ASHG **2019**
ANNUAL MEETING
HOUSTON • OCTOBER 15-19

American Society of Human Genetics 69th Annual Meeting
Plenary and Platform Abstracts

		Abstract #'s
Tuesday, October 15, 4:50 pm-6:00 pm:		
8. Featured Plenary Abstract Session I	Hall B	#1-#3
Wednesday, October 16, 9:00 am-10:00 am, Concurrent Platform Session A:		
9. Genetics of Addiction and Behavior	Hall B	#4-#7
10. Novel Discoveries in Large-Scale Genome-Wide Association Studies	Grand Ballroom A	#8-#11
11. Better Genome References and Representations	Grand Ballroom B	#12-#15
12. Large-scale Proteome and Metabolome Studies	Grand Ballroom C	#16-#19
13. Research Participants' Experiences and Preferences	Room 310A	#20-#23
14. New Discoveries in Long-studied Genes: Cancer Syndromes	Room 360D	#24-#27
15. Gene Expression Variation Across Diverse Global Populations	Room 361D	#28-#31
16. Telomeres: What You Need to Know at the End	Room 370A	#32-#35
17. Causes and Mechanisms Underlying Mendelian Neurogenetic Conditions	Room 371A	#36-#39
Wednesday, October 16, 1:00 pm-2:00pm, Platform Session		
112. Exome and RNA-based Sequencing Methods for Variant Interpretation to Improve Clinical Utility	Room 310A	#2609, #197, #2536, #225
Wednesday, October 16, 4:15 pm-5:45 pm, Concurrent Platform Session B:		
27. Precision Medicine: Models and Complex Disease	Hall B	#40-#45
28. Spectrum of Genomic Alterations in Cancer	Grand Ballroom A	#46-#51
29. Single Cell Transcriptomics of the Brain to Inform the Genetics of Neurological Disorders	Grand Ballroom B	#52-#57
30. Analyses Utilizing Biobanks	Grand Ballroom C	#58-#63
31. Genetics and Functional Insights into Cardiovascular Disease	Room 310A	#64-#69
32. Taking a Closer Look: New Discoveries in Mendelian Eye Diseases	Room 360D	#70-#75
33. Improved Structural Variation Detection Leads to New Insights into Disease and Development	Room 361D	#76-#81
34. Fine-scale Population Structure in Asia and America	Room 370A	#82-#87
35. Statistical Methods for GWAS Interpretation with Gene Expression Data	Room 371A	#88-#93
Wednesday, October 16, 6:35pm-7:15 pm		
39. Featured Plenary Abstract Session II	Hall B	#94-#95

Thursday, October 17, 9:00 am-10:30 am, Concurrent Platform Session C:

40. Methods and Resources in Large-scale Population Data	Hall B	#96-#101
41. Somatic Mosaicism in Affected and Unaffected Individuals	Grand Ballroom A	#102-#107
42. Genetics in Therapeutic Target Discovery	Grand Ballroom B	#108-#113
43. Genetic Risk Factors for Cardiovascular Diseases	Grand Ballroom C	#114-#119
44. Genetic Regulatory Variants and Complex Trait Associations	Room 310A	#120-#125
45. Strategies to Improve Genetic Counseling Practice & Education	Room 360D	#126-#131
46. Genetics of Prostate Cancer	Room 361D	#132-#137
47. Genetic Mechanisms of Autism and Related Disorders	Room 370A	#138-#143
48. Causal Genes in Skeletal Development	Room 371A	#144-#149

Thursday, October 17, 11:00 am-12:30 pm, Concurrent Platform Session D:

49. Variants Associated with Cancer in Large Cohorts	Hall B	#150-#155
50. Dominant and Recessive: Not that Simple? Lessons from Clinics and Cohorts	Grand Ballroom A	#156-#161
51. Chromatin Accessibility and Spatial Genome Organization in Disease	Grand Ballroom B	#162-#167
52. Considerations With Using Polygenic Risk Scores	Grand Ballroom C	#168-#173
53. Genetics of Cardiac and Vascular Disorders	Room 310A	#174-#179
54. Evolutionary Mechanisms Underlying Phenotypic Change	Room 360D	#180-#185
55. Genetic Effects on Transcriptome and Genome Traits	Room 361D	#186-#191
56. Functional Assays for Clinically Relevant Variant Interpretation	Room 370A	#192-#196, #2418
57. Solving the Unsolved: Strategies for Increasing Diagnostic Yield	Room 371A	#198-#203

Thursday, October 17, 4:15 pm–5:15 pm, Concurrent Platform Session E:

58. CRISPR-Based Approaches to Study Genome Function	Hall B	#204-#207
59. Mechanisms of Rare Neurogenetic Disorders	Grand Ballroom A	#208-#211
60. A Deep Dive into Deep Learning	Grand Ballroom B	#212-#215
61. Dissecting Molecular Pathways in Schizophrenia	Grand Ballroom C	#216-#219
62. Haplotype-level Interrogation of the Genome	Room 310A	#220-#223
63. RNAseq to Augment Variant Interpretation and Disease Diagnosis	Room 360D	#224-#227
64. Genetic Basis of Diabetes	Room 361D	#228-#231
65. Grim Inheritance: Germline Predisposition to Pediatric and Adult Cancers	Room 370A	#232-#235
66. Heritability and Dominance in Complex Traits	Room 371A	#236-#239

Friday, October 18, 9:00 am-10:00 am, Concurrent Platform Session F:

68. Transferability of Polygenic Risk Scores Across Populations	Hall B	#240-#243
69. Tangled Webs: Deconstructing Complex Regulatory Networks in Cancer	Grand Ballroom A	#244-#247
70. Fast Methods for Genome Analysis	Grand Ballroom B	#248-#251

71. Use of Single Cell RNA-seq to Dissect Fundamental Cellular Processes	Grand Ballroom C	#252-#255
72. Integrated Genomics and Transcriptomics in Parkinson's Disease	Room 310A	#256-#259
73. Alternative Methods for Evaluating Variant Pathogenicity	Room 360D	#260-#263
74. Detection and Evaluation of Actionable Findings: ACMG 59 and Beyond	Room 361D	#264-#267
75. Reproductive Fitness: Genetic Insights into Fertility	Room 370A	#268-#271
76. Sex Differences in Genetic Disorders	Room 371A	#272-#275

Friday, October 18, 5:15 pm-6:15 pm:

90. Featured Plenary Abstract Session III	Hall B	#276-#278

Saturday, October 20, 8:30 am-9:30 am, Concurrent Platform Session G:

92. Genomics and Therapeutics of Cancer and Prevention	Hall B	#279-#282
93. Bioinformatics and Machine Learning Methods	Grand Ballroom A	#283-#286
94. New Approaches for Novel Insights into Genetic Associations in Large-scale EHR and Biobank Studies	Grand Ballroom B	#287-#290
95. Large-scale Phenotype Association Studies	Grand Ballroom C	#291-#294
96. Computational Methods for Genetic Data	Room 310A	#295-#298
97. Chromosomes to Cell-free DNA: Balancing Genetic Contributions	Room 360D	#299-#302
98. Gene Regulation and Neurological Phenotypes	Room 361D	#303-#306
99. Precision Medicine: Rare Variants	Room 370A	#307-#310
100. Uncovering Genome Complexity and Function with Long-read Sequencing	Room 371A	#311-#314

Saturday, October 19, 9:45 am-11:15 am, Concurrent Platform Session H:

101. Pharmacogenomics and Gene Therapy	Hall B	#315-#320
102. Darwin's Tumor: Mutation, Selection, and Evolution in Cancer Genomes	Grand Ballroom A	#321-#326
103. DNA Methylation	Grand Ballroom B	#327-#332
104. Enhanced Analysis and Correction of Single-cell Data	Grand Ballroom C	#333-#338
105. Mechanisms of Immune Cell Phenotypes and Clonal Hematopoiesis	Room 310A	#339-#344
106. New Insights into Rare Skeletal, Growth, and Vascular Disorders	Room 360D	#349-#350
107. Methods and Resources for Improved Genomic Variant Interpretation	Room 361D	#351-#356
108. Prenatal Diagnosis and Pregnancy Loss	Room 370A	#357-#362
109. Novel Methods in Variant Association Studies	Room 371A	#363-#368

Saturday, October 19, 11:30 am-12:30 pm:

110. Featured Plenary Abstract Session IV	Hall B	#369-#371

A- A+

PgmNr 1: High-depth genome sequencing in diverse African populations reveals the impact of ancestral migration, cultural demography, and infectious disease on the human genome.

Authors:
N.A. Hanchard [1]; A. Choudhury [2]; S. Aron [2]; L. Botigue [3]; D. Sengupta [2]; G. Botha [4]; T. Bensellak [5]; G. Wells [6,7]; J. Kumuthini [7]; D. Shriner [8]; Y. Jaufeerally Fakim [9]; A. Wahed Ghoorah [10]; E. Dareng [11,12]; T. Odia [13]; D. Falola [13]; E. Adebiyi [13,14]; S. Hazelhurst [2,15]; G. Mazandu [4]; O.A. Nyangiri [16]; M. Mbiyavanga [4]; S. Kassim [17]; N. Mulder [4]; S.N. Adebamowo [18]; E.R. Chimusa [19]; C. Rotimi [8]; M. Ramsay [2,20]; A. Adeyomo [8]; Z. Lombard [20]; as members of the H3Africa Consortium

View Session Add to Schedule

Affiliations:
1) Department of Molecular & Human Genetics, Baylor College Medicine, Houston, Texas.; 2) Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; 3) Center for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Plant and Animal Genomics Program, Campus UAB, 08193 Bellaterra, Barcelona, Spain; 4) Computational Biology Group and H3ABioNet, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa; 5) System and Data Engineering Team, Abdelmalek Essaadi University, ENSA, Tangier, Morocco; 6) African Health Research Institute, Nelson R. Mandela School of Medicine, Durban, South Africa; 7) Centre for Proteomic and Genomic Research (CPGR), Cape Town, South Africa; 8) Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America; 9) Department of Agriculture and Food Science, Faculty of Agriculture, University of Mauritius, Reduit, Mauritius; 10) Department of Digital Technologies, Faculty of Information, Communication & Digital Technologies, University of Mauritius, Reduit, Mauritius; 11) Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; 12) Institute of Human Virology Nigeria, Abuja, Nigeria; 13) Covenant University Bioinformatics Research (CUBRe), Covenant University, Ota, Nigeria; 14) Department of Computer and Information Sciences, Covenant University, Ota, Nigeria; 15) School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa; 16) College of Veterinary Medicine, Animal Resources and Biosecurity, Makerere University, Kampala, Uganda; 17) Medical Biochemistry & Molecular Biology Department, Faculty of Medicine, Ain Shams University, Abbaseya, Cairo, Egypt; 18) Department of Epidemiology and Public Health, and Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, Maryland, United States of America; 19) Division of Human Genetics, Department of Pathology, Faculty of Health Sciences, Institute for Infectious, Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa; 20) Division of Human Genetics, National Health Laboratory Service, and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, South Africa

---

Africa is the cradle of human genetic variation; yet, to date, genomic studies in African populations have mostly focused on common variation in small, geographically-limited groups. The Human Health and Heredity in Africa (H3Africa) Consortium was convened to support genomic disease studies on the continent. To provide context for these studies, we undertook whole-genome sequencing of 426 individuals, including 323 at high depth, from 50 ethnolinguistic groups recruited from 13 countries. The resulting data were interrogated for patterns of admixture and selection as well as the distributions of rare, novel, and medically important variation.

Of the >3 million novel single nucleotide variants (SNVs) identified, most occurred in newly sampled groups from Botswana and Mali, with each surveyed population contributing at least 6,000 new common SNVs. Participants from Mali showed evidence of Northern patrilineal admixture and extended runs of homozygosity, reflecting regional cultural practices. ADMIXTURE analyses revealed a novel, putative East African ancestry component dating back 50-70 generations that constituted ~14% of a major Nigerian indigenous group, but was not seen in other West Africans. We also observed a lack of Khoesan ancestry among Zambians that distinguished them from their Southern neighbors, implicating present-day Zambia as an intermediate site in Southern and Eastern Bantu migrations.

Using a composite model that leveraged high depth data, we identified 63 loci with strong signatures of selection, including 33 novel loci converging upon genes involved in viral infection (*CAMK2B*, *MVB12B*, and *DKK2*) and metabolism (*ADRB3* and *GLIS3*). This was congruent with an enrichment of shared, putative loss-of-function variants and highly population-divergent allele frequencies at immune loci. ACMG medically-actionable variants were uncommon (<2%), but each person had a median of 7 reportedly pathogenic ClinVar alleles, some with frequencies 10 times higher than in current databases. Classical African disease alleles, including the sickle allele and *APOL1* G1 and G2 variants, were common, but frequencies varied widely by geography and ancestry.

Our results 1) highlight unexpected patterns of admixture and ancestry among ethnolinguistic groups that augment current theories of early migration, 2) illustrate the role of infections in shaping human genomes, and 3) demonstrate the importance of African genomic data to defining medically relevant variation.

A- A+

PgmNr 2: Pathogenic, loss-of-function mutations in *MRAP2* cause metabolic syndrome.

Authors:
A. Bonnefond [1]; M. Baron [1]; J. Maillet [1]; M. Huyvaert [1]; R. Boutry [1]; G. Charpentier [2]; M. Tauber [3]; R. Roussel [4]; B. Balkau [5]; M. Marre [4]; M. Canouil [1]; P. Froguel [1]

View Session Add to Schedule

Affiliations:
1) CNRS UMR8199, Lille, France; 2) CERITD (Centre d'Étude et de Recherche pour l'Intensification du Traitement du Diabète), Evry, France; 3) Inserm UMR1043, Toulouse, France; 4) Inserm U1138, Centre de Recherche des Cordeliers, Paris, France; 5) Inserm U1018, Center for Research in Epidemiology and Population Health, Villejuif, France

---

The G-protein-coupled receptor (GPCR) accessory protein MRAP2 was reported to be involved in rodent energy control. Notably, it was shown that MRAP2 interacted directly with the obesity gene MC4R, and enhanced MC4R downstream signaling, suggesting that this mechanism linked Mrap2 deficiency and rodent obesity. Although a couple of mutations in *MRAP2* were described in few obese human subjects, their functional consequences and their putative impact on obesity have remained elusive. Here, we performed a large-scale resequencing study of *MRAP2* in 9,418 participants, in combination with functional assays of detected variants, to accurately decipher the functional link between MRAP2 and obesity (and possibly associated phenotypes) in humans. We identified 23 rare variants (with a minor allele frequency between 0.0053 and 0.17%) that were significantly associated with increased obesity risk, in both adults and children. After functional assessment of each variant, we found that 7 pathogenic, loss-of-function *MRAP2* variants were totally penetrant for overweight or obesity in both adults and children. Surprisingly, when we investigated the clinical data of the carriers, we found that they actually had metabolic syndrome. In addition to high adiposity, this genetic form of metabolic syndrome was mainly associated with hyperglycemia (88% of carriers) and hypertension (71% of carriers). This is remarkable as only 30-60% of obese people present with metabolic syndrome. Importantly, MRAP2-deficient obese subjects are different from MC4R-deficient obese subjects who do not present with any other metabolic clinical features, with particularly low blood pressure. Through expression analysis of *MRAP2* in a panel of human tissues, we surprisingly found that the high expression of *MRAP2* was similar in brain regions and in human pancreatic islets and beta cells. We also discovered that *MRAP2* knockdown in human beta cells significantly decreased insulin secretion. As we also found that MRAP2 was expressed in key metabolic tissues including gut, kidney, muscle and fat, we suggest that MRAP2 deficiency causes combined metabolic abnormalities in humans possibly due to the failure of different GPCRs signaling (co)regulated by MRAP2 (such as ghrelin receptor). To our knowledge, it is only the second description of a monogenic form of metabolic syndrome, after the identification of *DYRK1B* deficiency few years ago, showing that complex metabolic phenotypes can be genetically driven.

A- A+

PgmNr 3: Exome sequencing of 25,000 schizophrenia cases and 100,000 controls implicates 10 risk genes, and provides insight into shared and distinct genetic risk and biology with other neurodevelopmental disorders.

Authors:
T. Singh [1,2]; B.M. Neale [1,2,3]; M.J. Daly [1,2,3]; on behalf of the SCHEMA consortium

View Session Add to Schedule

Affiliations:
1) Center for Genomic Medicine, Massachussetts General Hospital, Boston, Massachusetts.; 2) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA.; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA.

Schizophrenia (SCZ) is a severe psychiatric disorder with common intergenic and ultra-rare coding variants (URVs) contributing to risk. Despite hundreds of common risk loci discovered by GWAS, only a handful have resulted in validated functional variants that pinpoint novel biology underlying disease pathogenesis. This central challenge, common to complex polygenic disorders, could be addressed by sequencing studies where URVs (1 in 10,000 or rarer) can complement GWAS by pinpointing likely causal genes independently. However, success to this end has been hampered by power limitations.

The Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA) Consortium is one of the largest efforts to analyze sequencing data to advance gene discovery. We have completed the analysis of 24,248 sequenced cases and 97,322 controls comprising of individuals from five continental populations. The scale of SCHEMA enables us, for the first time, to implicate URVs in ten genes as conferring substantial risk for SCZ at genome-wide significance (odds ratios 4 - 50, $P < 2e-6$), and 34 genes at a FDR < 5%. Two of these, the NMDA receptor subunit *GRIN2A* and transcription factor *SP4*, reside in two loci implicated by SCZ GWAS. A second glutamate receptor subunit, *GRIA3*, is also implicated, providing support for the hypofunction of the glutamatergic system in the pathogenesis of schizophrenia.

Exploring the published results from severe neurodevelopmental delay (NDD) and autism (ASD) consortia, we find that top ten SCZ genes have no protein-truncating variant signal in either ascertainment, though a significant overlap between ASD (n = 102, FDR < 10%) and SCZ risk genes (n = 34, FDR < 5%) was observed. Partitioning the 102 ASD genes into those disrupted more frequently 1) in ASD and 2) in intellectual disability (ID), we show that this signal was driven by the ASD-preferential, and not ID-preferential genes. Thus, SCZ genes from exome sequencing have relevance for later-onset psychiatric disorders rather than more severe NDDs.

Despite not finding a notable overlap between genes identified by GWAS and exome sequencing, we find convergence in biological processes and tissue types, specifically in synaptic transmission and components of the post-synaptic density. After excluding associated genes, SCZ cases still carry a substantial excess of rare URVs, suggesting that many more remain to be discovered. Finally, we present a browser that displays variant-level results for use by the community.

A- A+

PgmNr 4: Genome-wide meta-analysis of alcohol use disorder and problematic drinking identifies over 30 risk variants.

Authors:
H. Zhou [1,2]; S. Sanchez-Roige [3]; T.K. Clarke [4]; J.M. Sealock [5]; R. Polimanti [1,2]; R.L. Kember [6,7]; R.V. Smith [8]; A.C. Justice [1,2,9]; L.K. Davis [5,10]; A.A. Palmer [3,11]; H.R. Kranzler [6,7]; J. Gelernter [1,2]; on behalf of the VA Million Veteran Program

View Session Add to Schedule

Affiliations:
1) Yale School of Medicine, New Haven, CT; 2) Veterans Affairs Connecticut Healthcare System, West Haven, CT; 3) Department of Psychiatry, University of California San Diego, CA; 4) Division of Psychiatry, University of Edinburgh, Edinburgh, UK; 5) Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; 6) University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 7) Crescenz Veterans Affairs Medical Center, Philadelphia, PA; 8) University of Louisville School of Nursing, Louisville, KY; 9) Yale School of Public Health, New Haven, CT; 10) Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN; 11) Institute for Genomic Medicine, University of California San Diego, CA

Alcohol use disorder (AUD) is among the leading causes of death and disability worldwide. Genome-wide association studies (GWASs) have identified few risk genes. However, the genetic architecture of AUD and the biological mechanisms are still understudied. We conduct genome-wide meta-analysis of AUD in European cohorts including: 1), AUD from Million Veteran Program (MVP) phase1; 2), AUD of newly genotyped individuals from MVP phase2; 3), alcohol dependence from Psychiatric Genomics Consortium (PGC), brings the total sample size to 313,959 (N_{case}=57,564, $N_{control}$=256,395). Given the high genetic correlation (r_g=0.7) between AUD and problematic drinking, measured by AUDIT-P (Alcohol Use Disorders Identification Test - Problems), we conduct a proxy-phenotype meta-analysis combining the above AUD datasets and AUDIT-P from UK Biobank, totaling in 435,563 subjects. All downstream analyses are based on the cross-trait meta-analysis. More than 30 risk loci have been identified, include 20 novel findings. Gene-based association analysis identified 70 genes associated with the trait. Genetic correlations with 133 traits have been detected, include smoking traits, alcohol traits, major depression, schizophrenia, and many other neuropsychological traits. Phenome-wide polygenic risk score (PRS) analysis in an independent biobank (BioVU) confirmed the genetic associations between AUD and multiple substance use disorders, anxiety, mood disorder, and lung-related diseases. In-depth functional analyses consistently show the genetic heritability of AUD is enriched in brain tissues, and in conserved region, enhancers and other regulatory regions. Gene expression were predicted using reference transcriptome data. 187 gene-tissue associations are significant, include *ADH1B*, *ADH4*, *ADH5*, *CADM2*, *DRD2*, *SLC39A13*, *C1QTNF4*, *MTCH2*, *SLC39A13*, *SNX17*, *NRBP1* and so on. Those associations in universal tissues might indicate pervasive functional consequences of genetic variation at the expression level. In summary, we present here the largest meta-analyses of AUD and problematic drinking, discover more genetic risks, genetic correlations with other traits, and potential functional mechanisms. More studies are warranted including genome-editing in near future.

A- A+

PgmNr 5: Trans-ancestry GWAS meta-analysis of tobacco and alcohol use.

Authors:
M. Liu; Trans-Omics for Precision Medicine; GWAS & Sequencing Consortium of Alcohol and Nicotine use

View Session Add to Schedule

Affiliation: Psychology, University of Minnesota, Minneapolis, MN

With ~1.1 billion smokers and ~2.3 billion drinkers world-wide, nicotine and alcohol are two of the most commonly used addictive substances. Together, they comprise the largest preventable cause of morbidity and mortality in economically developed countries. Despite the global nature of these behaviors, existing genetically informative research has focused predominantly on individuals of European ancestry. To better understand genetic influences on nicotine and alcohol, and by extension to addictive behavior more generally, we conducted a GWAS meta-analysis of nicotine and alcohol use in up to 3.4 million individuals of diverse ancestries as part of GSCAN and the TOPMed program. Our approach maximizes power for variant detection, allows evaluation of ancestry-specific variant effects, and provides greater fine-mapping resolution for five phenotypes representing different stages of cigarette use and alcohol use. Fixed effects meta-analysis across all ancestries were well-behaved, and loci defined conservatively as non-overlapping 1MB windows. For the phenotype indexing whether an individual had ever been a regular smoker (N=3,377,408 from 75 studies) we identified 780 loci (404 novel); for age of initiation of regular smoking (N=731,870; 64 studies) we identified 39 loci (30 novel); for cigarettes per day, a measure of heaviness of use (N=782,790; 81 studies) we identified 157 loci (103 novel); for a measure of smoking cessation (N=1,400,906; 74 studies) we identified 112 loci (91 novel); and finally for a measure of alcohol use, drinks per week (N=2,896,131; 62 studies), we identified 376 loci; (286 novel). The maximum sample sizes for non-European ancestry included 121,858 individuals with African ancestry, 285,155 Latinx individuals, and 298,624 individuals of East Asian ancestry, predominantly from Japan and China. We will report ancestry-based moderation and fine-mapping of variant associations, as well as heritabilities and utility of polygenic scores from this data set.

A- A+

PgmNr 6: The functional landscape and essential genes discovery in genetic etiology of tobacco use.

Authors:
F. Chen [1,2]; Y. Jiang [1,2]; D.J. Liu [1,2]

View Session Add to Schedule

Affiliations:
1) Department of Public Health Sciences, Penn State College of Medicine, Hershey,PA; 2) Institute of Personalized Medicine, College of Medicine, Pennsylvania State University, Hershey, PA

Tobacco use is a heritable and modifiable risk factor for a myriad of human diseases, including cardiovascular disease, respiratory disorder, and cancer. Developing novel therapeutics to aid in smoking cessation can bring substantial public health benefits. Genome-wide association study results have been established as a valuable resource for developing and repurposing drugs. In a recent GWAS with 1.2 million participants, we discovered more than 400 loci that show robust associations with phenotypes representing different stages of smoking, including smoking initiation, smoking cessation, and cigarettes per day. Despite a large number of novel discoveries, identifying the essential genes and pathways for drug development remains challenging.

Here we performed biological pathway, disease pathway, and drug targets enrichment analysis, in order to identify essential genes, and druggable targets. First, we noted that besides the typical smoking cessation drugs such as Cytisine and Varenicline, drugs that have already been used by other treatments are significantly enriched with tobacco-use associated SNPs (i.e., Fluspirilene, Lumateperone, and Acetophenazine). As many drugs for psychopharmacological and other treatments have been shown to be effective for smoking cessation, this observed enrichment showcased the effectiveness of our approach.

Next, to further prioritize the genes that are enriched with tobacco-use associated SNPs for drug development, we integrate information from gene-set analysis, PPI (protein-protein interaction), and eQTL database. The approaches utilized here include Magma, DEPICT, DAPPLE, and Locuscompare. Genes pass the Bonferroni correction p-value or FDR threshold in all methods are prioritized. To this end, we identified 19 genes, including GRID2, RUNX1T1, TOP2B, GRID1, NRXN3, RBFOX1, and BTRC, many of which are indeed deemed `druggable` according to the Illuminating the Druggable Genome (IDG) database. Finally, the GO analysis showed these genes are highly enriched in glutamate-related pathways, suggesting it as suitable therapeutic targets. Our prioritized gene list nicely complements the existing therapies for smoking cessation which are focused on nicotinic receptor subunits. They provide a set of much more flexible and promising targets in future clinical researches. We expect that these results are precious for smoking cessation pharmacotherapies development, repurpose, and possible side effects of the medications.

A- A+

PgmNr 7: Obesity-associated variants in the *FTO* locus: Dissecting the complex landscape of GWAS.

Authors:
D. Sobreira; I. Aneas; A. Joslin; M. Nobrega

View Session Add to Schedule

Affiliation: Department of Human Genetics, University of Chicago, Chicago, IL

The strongest obesity GWAS association lies in the first intron of the *FTO* gene. We have previously shown that these variants are functionally connected to *IRX3* in brain and adipose tissue. However, establishing the mechanistic basis for this association remains unresolved. Specifically, there is compelling and conflicting evidence implicating alterations in food preference and feeding behavior, regulated in the brain, and altered metabolic rate through disruptions in mitochondrial function regulating thermogenesis, autonomous to adipose tissue. In order to address this paradox, we applied an integrated platform to dissected the *FTO* locus regulatory circuit, and our findings reveal several nuances connecting genetic differences to phenotypic variation.Importantly, we found that several SNPs segregating on a common haplotype, each within distinct enhancers with neuronal and/or adipose tissue specificities, are capable of modulating *IRX3* and *IRX5* expression, thus implicating multiple causal variants with the association to obesity. Transcriptomic data in hypothalamus of *Irx3* knockout mice supports a role of the central nervous system in obesity susceptibility, suggesting alterations in feeding behavior as an important driver underlying obesity risk. Also, we show that mis-expression of *IRX3* and *IRX5* in murine and human adipocytes and hypothalamic neurons trigger a similar downstream cascade of mitochondrial dysfunction, illustrating an example of phenotypic convergence, where genetic variants impart their phenotypic effects through shared cellular processes in distinct tissues. Our results challenge the original model of GWAS reflecting an association of a SNP in a locus disrupting a regulatory element. Rather, our data support a model that includes pleiotropy, genetic heterogeneity, and molecular convergence of phenotypes ultimately regulating systems-wide organismal phenotypes. Our findingsunravel a new level of regulatory complexity at the *FTO* locus and illustrate that from a mechanistic perspective, the etiology of the genetic associations often is much more complex than previously thought.

A- A+

PgmNr 8: Genome-wide association study on vitamin D levels in 482,619 Europeans reveals 47 novel vitamin D-related loci.

Authors:
D. Manousaki [1,2]; R. Mitchel [3]; T. Dudding [3]; S. Haworth [3]; V. Forgetta [2]; N.J. Timpson [3]; J.B. Richards [1,2,4,5]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics,McGill Univ, Montreal, Quebec, Canada; 2) Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, QC H3T 1E2, Canada; 3) Medical Research Council Integrative Epidemiology Unit (IEU) at the University of Bristol, Bristol, BS8 2BN, UK; 4) Department of Epidemiology, Biostatistics and Occupational Health and Department of Medicine, McGill University, Montreal, QC H3A 1A2, Canada; 5) Department of Twin Research and Genetic Epidemiology, King's College London, London, WC2R 2LS, United Kingdom

Objective: We sought to increase our understanding of genetics of vitamin D levels by undertaking the largest to date genome-wide association study (GWAS) of plasma 25 hydroxyvitamin D (25OHD) levels, the most common biomarker of vitamin D status in humans.

Methods: Using data from 440,345 White British individuals from UK Biobank with available 25OHD levels and imputed genotypes we conducted a linear mixed model GWAS using the BOLT-LMM software, to account for population stratification and cryptic relatedness. We retained single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) > 0.1%, and imputation quality score > 0.3 from the autosomes and the X chromosome. We used standardized log-transformed 25OHD levels, adjusting for age, sex, season of measurement, and vitamin D supplementation. We next meta-analyzed this GWAS with our previous GWAS on 42,274 individuals of European ancestry using GWAMA. To identify conditionally independent SNPs from this meta-analysis, we performed a conditional analysis using GCTA-COJO.

Results: After quality control, a total of 20,370,875 SNPs were tested for association with 25OHD levels. The genomic control lambda was 1.23, and the LD score regression intercept was 1.06, implying little evidence of population stratification. The SNP heritability of vitamin D levels was estimated by BOLT-LMM to be 16.1%. After meta-analysis with our previous GWAS and using GCTA–COJO, we observed 105 independent 25OHD-associated SNPs (pre and post conditioning p-value< 6.6 x10^{-9}) among which 40 had MAF<5%. These SNPs map in 53 distinct loci (defined as 1Mb regions), among which 47 are novel, while all 6 previously described 25OHD loci replicated in our study. The 40 SNPs with MAF <5% conferred an average absolute effect of 0.22 standard deviations on standardized log transformed 25OHD levels per effect allele, compared to 0.03 of the 65 SNPs with MAF>5%.

Conclusions: Through the largest to date GWAS on 25OHD levels, we identified 47 novel 25OHD associated loci. Our findings support the polygenicity of vitamin D levels, increase substantially their assigned heritability and contribute to our knowledge on the genetic control of 25OHD levels. Moreover, by identifying more genocopies of vitamin D levels, this study will enable the development of a genomic predictor for vitamin D insufficiency, and will provide new instruments to test associations with diseases through Mendelian randomization.

PgmNr 9: A large cross-ancestry meta-analysis of genome-wide association studies identifies novel risk loci for primary open-angle glaucoma, and shows a genetic link to Alzheimer’s disease.

Authors:
P. Gharahkhani [1]; E. Jorgenson [2]; P. Hysi [3]; A. Khawaja [4]; S.A. Pendergrass [5]; X. Han [1]; A. Hewitt [6,7]; R. Igo [8]; H. Choquet [2]; N. Josyula [5]; D. Mackey [9]; C.P. Pang [10]; F. Pasutto [11]; P. Mitchell [12]; P. Bonnemaijer [13,14]; A. Lotery [15]; N. Pfeiffer [16]; A. Palotie [17,18]; C. van Duijn [14]; J. Haines [8]; C. Hammond [19]; M. Hauser [20]; L. Pasquale [21,22]; C.C.W. Klaver [13,14,23]; M. Kubo [24]; T. Aung [25,26,27]; J.E. Craig [28]; S. MacGregor [1]; J. Wiggs [29]; International Glaucoma Genetics Consortium, NEIGHBORHOOD consortium, ANZRAG study, FinnGen study

View Session Add to Schedule

Affiliations:
1) Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia; 2) Division of Research, Kaiser Permanente Northern California (KPNC), Oakland, CA, USA.; 3) Department of Twin Research and Genetic Epidemiology, King's College London, UK.; 4) Department of Public Health and Primary Care, Institute of Public Health, University of Cambridge, School of Clinical Medicine, Cambridge, UK.; 5) Geisinger Research, Biomedical and Translational Informatics Institute, Danville, PA, USA.; 6) Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia.; 7) Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, University of Melbourne, Melbourne, Australia.; 8) Institute for Computational Biology, Case Western Reserve University School of Medicine, Cleveland, Ohio, United States.; 9) Lions Eye Institute, Centre for Ophthalmology and Visual Science, University of Western Australia, Perth, Western Australia, Australia.; 10) Department of Ophthalmology and Visual Sciences, the Chinese University of Hong Kong, Hong Kong.; 11) Institute of Human Genetics, Universita¨tsklinikum Erlangen, Friedrich-Alexander-Universita¨t Erlangen-Nu¨rnberg, Erlangen, Germany.; 12) Centre for Vision Research, Department of Ophthalmology and Westmead Institute for Medical Research, University of Sydney, Sydney, NSW, Australia.; 13) Department of Ophthalmology, Erasmus MC, Rotterdam, The Netherlands.; 14) Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands.; 15) Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, England.; 16) Department of Ophthalmology, University Medical Center Mainz, Mainz, Germany.; 17) Institute for Molecular Medicine Finland, Helsinki Institute of Life Sciences, University of Helsinki, Helsinki 00014, Finland.; 18) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.; 19) Department of Ophthalmology, King's College London, St. Thomas' Hospital, London, UK.; 20) Department of Ophthalmology, Duke University Medical Center, Durham, North Carolina, United States.; 21) Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.; 22) Icahn School of Medicine at Mount Sinai, Department of Ophthalmology, New York, NY, USA.; 23) Department of Ophthalmology, Radboud University Medical Center, Nijmegen, The Netherlands.; 24) Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.; 25) Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.; 26) Ophthalmology & Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore.; 27) Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.; 28) Department of Ophthalmology, Flinders University, Flinders Medical Centre, Adelaide, Australia.; 29) Department of

Ophthalmology, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, MA, USA.

Glaucoma is the leading cause of irreversible blindness worldwide. Primary open-angle glaucoma (POAG) is one of the most common subtypes. Despite a high heritability, known risk loci only explain a small proportion of POAG risk. We conducted the largest meta-analysis of genome-wide association studies for POAG to date, using more than 32K cases and 338K controls across European, Asian, and African ancestries.

We identified 123 independent genome-wide significant loci for POAG, of which 98 were not previously reported. For the first time, we identified POAG risk loci at key genes known to be involved in dementia/cognitive function, e.g. *MAPT* (rs242559[C], OR=1.08, P=8.8e-10), TRIOBP (rs5750494[T], OR=1.08, P=2.4e-16), and *APP* (rs13049669[T], OR=1.1 , P=2.2e-09). We also found a genome-wide genetic correlation of 15% (P<0.05) between glaucoma and Alzheimer’s disease (based on 72K Alzheimer’s cases, 383K controls). Moreover, for the first time, we identified an association of an HLA gene (HLA-G/HLA-H) with POAG. Several of the risk loci (e.g. *GLIS1*, *DDR2*, and *THRB*) appeared to influence POAG independent of increased intraocular pressure and changes in optic nerve morphology (vertical cup-disc ratio), the major mechanisms known to be involved in development of POAG.

Prevalence of POAG varies with ethnicity in epidemiological studies, raising the question of ancestry-specific genetic effects. We found relatively consistent genetic effects across ancestries; the correlation of the effect estimates for the genome-wide significant loci was 0.8 between Europeans and Asians and 0.7 between Europeans and Africans.

The new risk loci have functional relevance supported by eQTL and chromatin interaction data. We also identified >30 additional new risk loci using gene-based analysis. The significant genes showed an enriched expression in the eye, artery, nerve, or nerve-enriched tissues. Pathway-based analyses identified >20 significant pathways including those involved in blood vessel morphogenesis, vasculature development, and collagen formation.

At least 15 of the risk genes are targeted by several drugs, some of which are already in use/clinical trials for retinal vein occlusion, age-related macular degeneration, diabetic retinopathy, Alzheimer’s, and cardiovascular diseases.

In summary, our study identified important new risk loci for POAG, supporting a biological link between POAG and Alzheimer’s disease, and suggesting target candidate genes for drug repurposing.

A- A+

PgmNr 10: Near-optimal trans-ethnic association and fine mapping of smoking associated genes integrating GWAS and TOPMed sequence data of 1.3 million individuals.

Authors:
Y. Jiang; TOPMed smoking working group and GSCAN consortium

View Session Add to Schedule

Affiliation: Department of Public Health Sciences, Penn State College of Medicine, Hershey, Pennsylvania.

Tobacco use is a heritable risk factor for numerous diseases, for which 353 associated genes were identified in European samples. Yet, its genetic architecture in non-European populations remains elusive. To address this, we assembled TOPMed whole genome sequences of ~150,000 individuals from diverse US populations as well as GWAS data of up to 1.2 million individuals. Four smoking phenotypes were studied, including smoking initiation, cigarettes per day, smoking cessation and the age of smoking initiation.

To analyze these amazingly rich datasets, we developed a novel mixed effect meta-regression method for near-optimal trans-ethnic meta-analysis (MEMO). MEMO summarizes ancestry for each study using principal components of genome-wide allele frequencies. It models the between-study genetic effect heterogeneities due to genetic ancestry differences as a fixed effect and that due to non-ancestry exposure differences as random effects. For each SNP, MEMO adaptively selects fixed effects and random effects to be included that best models the genetic effect heterogeneity. It thus combines the strength of fixed effect, random effect meta-analysis, and meta-regression. MEMO is consistently the most powerful (or close to the most powerful) across a wide variety of scenarios in simulations, even when the simulated disease model is in favor of alternative methods. We further extend MEMO for fine mapping, which can distinguish causal variants with homogeneous effects and that show ancestry-specific effects. Due to the improved model of multi-ethnic genetic effects, MEMO considerably improves fine mapping resolution. Simulation shows the method is well calibrated and on average, the posterior probability of association for causal variants estimated by our method is 50% higher, and our 95% credible interval for causal variants is ~33% shorter than alternative trans-ethnic fine-mapping methods.

Applying MEMO, we identified 265 loci with p<5e-9 among which 27 are novel, and >400 independent secondary associations. Our fine-mapping narrowed down the 95% credible interval for causal variants to less than 10 variants for 76 loci, and 17 of them contain a single SNP. We estimated that 56% of the causal variants show homogeneous effects across ancestries, while another 26% and 12% show African specific and Hispanic specific effects. In conclusion, our results elucidate the genetic architecture for smoking traits, and our developed methods will be valuable for other studies.

A- A+

PgmNr 11: Genetics of 38 blood and urine biomarkers in the UK Biobank.

Authors:
N. Sinnott-Armstrong [1]; Y. Tanigawa [2]; S. Naqvi [1,3]; N.J. Mars [4]; D. Amar [2]; H.M. Ollila [4,5,6]; M. Aguirre [2]; G.R. Venkataraman [2]; M. Wainberg [7]; J.P. Pirruccello [8,9]; J. Qian [10]; A. Shcherbina [2,11]; F. Rodriguez [11]; T.L. Assimes [11,12]; V. Agarwala [11]; R. Tibshirani [10]; T. Hastie [10]; S. Ripatti [3,9,13]; M.J. Daly [3,9,15]; J.K. Pritchard [1,4,14]; M.A. Rivas [2]; FinnGen

View Session Add to Schedule

Affiliations:
1) Genetics, Stanford Univ, Stanford, California.; 2) Biomedical Data Science, Stanford University, Stanford, California; 3) HHMI, Stanford, California; 4) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; 5) Stanford University, Department of Psychiatry and Behavioral Sciences, Palo Alto, CA, USA; 6) Center for Genomic Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA; 7) Department of Computer Science, Stanford University, Stanford, CA, USA; 8) Massachusetts General Hospital Division of Cardiology, Boston, MA, USA; 9) Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 10) Department of Statistics, Stanford University, Stanford, CA, USA; 11) Department of Medicine, School of Medicine, Stanford University, Stanford, CA, USA; 12) VA Palo Alto Health Care System, Palo Alto, CA, USA; 13) Department of Public Health, Clinicum, University of Helsinki, Helsinki, Finland; 14) Department of Biology, Stanford University, Stanford, CA, USA; 15) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA., USA

Biomarkers are well suited to testing how and when variation in the genome changes phenotype, as they are often understood on a molecular level. Here, we systematically evaluated the genetic basis of 38 blood and urine laboratory tests measured in 358,072 participants in the UK Biobank.

We identified 1,857 independent loci associated with at least one laboratory test, including 488 large-effect protein truncating, missense, and copy-number variants. These loci included membrane transporter SLC2A9 for urate; the chaperone IGFBP3 for IGF-1; and the activating enzyme SRD5A2 for testosterone, which were all key members of the corresponding gene pathways. More generally, up to 80% of genes in the relevant pathways contained common variation within 50 Kb that significantly altered biomarker levels. Moreover, rare variants also revealed novel coding associations with a number of genes with therapeutic potential.

Our findings suggest that biomarkers are driven by tissue-specific polygenic backgrounds and a few core genes with large effect. To this end, we found tissue- and cell-type specific polygenic signal in kidney tissue for urate (~35-fold); UACR and SHBG in podocytes and LDL in hepatocytes; and creatinine, alkaline phosphatase, and eGFR in proximal tubules. The polygenic architecture of biomarkers echos that of common diseases with the major exception that we have an a priori, molecularly-driven sense of core genes.

Finally, we built combined polygenic risk score (PRS) models using all 38 biomarker PRSs

simultaneously. We found substantially improved prediction of incidence in FinnGen (n = 135,500) with the multi-PRS for renal failure and alcoholic cirrhosis (hazard ratio = 1.1 vs no association with trait PRS alone).

Our results reveal that biomarkers are an ideal model system to understand the genetic architecture of complex phenotypes. By combining disease associations with measurements from a number of relevant biomarkers, we can improve the utility, interpretability, and portability of genetic associations.

A- A+

PgmNr 12: Advancements in the human genome reference assembly (GRCh38).

Authors:
T. Rezaie [1]; K. Howe [2]; T. Graves-Lindsay [3]; P. Flicek [4]; V.A. Schneider [1]; the Genome Reference Consortium

View Session Add to Schedule

Affiliations:
1) National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland; 2) The Wellcome Sanger Institute, Hinxton, Cambridge, UK; 3) The McDonnell Genome Institute at Washington University, St. Louis, MO; 4) European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK

The Genome Reference Consortium (GRC) provides updates to the human reference genome assembly, a resource critical to the basic and clinical research communities that serves as the basis of the coordinate system used for gene and other annotations, provides a representation of population diversity in the form of alternate loci, and supports identification of disease-associated variants. Although the reference has enabled numerous discoveries since its initial release in 2001, more recent studies have revealed the limitations of its linear pseudo-haploid chromosome assemblies and highlighted the need for a reference that represents population diversity. We will present features of the GRC assembly model that support inclusion of such diversity in the current reference, GRCh38, and discuss how this reference may also contribute to future graph and non-graph pangenome representations. In addition, we will highlight updates made to GRCh38 since its 2013 release. The GRC has generated a total of 13 publicly available non-coordinate-changing patch releases. As of the latest, (GRCh38.p13, GCA_000001405.28), these cumulatively include 113 fix and 72 novel patch scaffolds, which respectively represent corrections to the GRCh38 chromosomes and alternate sequence representations of variant genomic regions. All patches are available from INSDC as accessioned sequences and have had their chromosome context defined by alignment to GRCh38. The current collection of patches covers 61 Mb (~1.97% of GRCh38), of which 4.99 Mb is unique sequence not found in GRCh38 and its alternate loci. Notably, the novel patch scaffolds include coding regions and provide variant representations of genes found on chromosomes, as well as the only assembly representation for other genes. In addition, we will present data on the most recent set of fix patches and the 0.5 Mb of previously missing sequence they add to the GRCh38 chromosomes, resulting in the closure of 28 GRCh38 assembly gaps and improved gene representations in clinically important regions. We will also present the results of ongoing analyses of potentially rare alleles (MAF<5%) in GRCh38 coding sequences, as well as efforts to correct erroneous bases and provide representation for biologically valid haplotypes. The GRC remains committed to transparency in its curation efforts and to the production of a reference assembly that supports the widest range of analyses. See updates at the GRC website, https://www.genomereference.org

A- A+

PgmNr 13: Constructing a reference genome that captures global genetic diversity for improved interpretation of whole genome sequencing data.

Authors:
K.H.Y. Wong [1]; W. Ma [1]; N. Wei [2]; E.C. Yeh [2]; W.J. Lin [2]; E.H.F. Wang [2]; J.P. Su [2]; F.J. Hsieh [2]; Y. Mostovoy [1]; M. Levy-Sakin [1]; S. Chow [1]; E. Young [3]; C. Chu [4]; A. Poon [4]; M. Xiao [3,5]; P.Y. Kwok [1,2,4,6]

View Session Add to Schedule

Affiliations:
1) Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA.; 2) Institute of Biomedical Sciences, Academia Sinica, Taiwan.; 3) School of Biomedical Engineering, Drexel University, Philadelphia, PA.; 4) Institute for Human Genetics, University of California, San Francisco, San Francisco, CA.; 5) Institute of Molecular Medicine and Infectious Disease in the school of Medicine, Drexel University, Philadephia, PA.; 6) Department of Dermatology, University of California, San Francisco, San Francisco, CA.

The flagship product of the human genome project is a collection of high-quality DNA sequences that provides a reference that serves as the basis for understanding health and disease. Integral as it has been to the scientific community, the current reference genome does not represent the genetic diversity found in different human populations. In fact, 70% of the reference sequences originated from a single DNA donor. We and others have identified numerous “non-reference unique insertions” (NUIs) found in multiple individuals from around the world but missing in the reference. Furthermore, our group previously demonstrated that 1/3 of the NUIs are found in the human transcriptome, and thus are likely to be of functional significance. Alignment of whole genome sequencing (WGS) reads is less accurate when sequencing reads contain alleles that are different from the reference. While some biases can be mitigated through careful analysis, the problem of missing sequences is particularly difficult to resolve. This problem arises when the study genome contains stretches of DNA that are missing from the reference. Reads that fail to align are discarded altogether. As WGS is being performed widely, it is important that we include as much NUIs as possible in the human reference genome, both for better alignment of WGS reads and for more comprehensive interpretation of the WGS data.

To construct a more representative genome reference, we generated 220 whole genomes *de novo* assemblies from diverse populations using 10x Genomics Linked-Reads technology. These assemblies were aligned to the reference genome to determine the insertion sites of all NUIs. NUI breakpoints and sequence content were analyzed for consistency among the different samples. Recurrent NUIs with consistent breakpoints totaling 7Mb were integrated into the Hg38 primary chromosomal assemblies so that these sequences can be annotated based on the local genomic context. To demonstrate the utility of the NUI-integrated reference, we showed that many of the unmapped reads in WGS datasets from the Simon Genome Diversity Project could be salvaged when aligning to the new reference. The NUI-integrated reference is the first step towards creating a comprehensive human reference genome with inclusion and annotation of sequences found across the global populations, a genome reference that reduces the number of unaligned WGS reads while enhancing the value of existing and future WGS datasets.

A- A+

PgmNr 14: Development of sequence variation graphs and graph-based software for genomics studies.

Authors:
S. Tetikol; V. Semenyuk; A. Dolgoborodov; A. Jain; J. Browning; I. Johnson; D. Turgut; O. Kalay; D. Kabakci; Graph Development Team

View Session Add to Schedule

Affiliation: Seven Bridges, Boston, MA.

As the pan-genome paradigm gains momentum, there are numerous challenges and design decisions concerning how to represent genomic variants in graphs, as well as how to implement tools and workflows in order to process sequencing data efficiently and effectively. We reported our initial efforts to develop an approach based on sequence variation graphs, which utilize a graph-based representation of genomic variation, along with an aligner and variant caller that take advantage of graph information to improve alignment and variant calling with next-generation sequencing data (Nature Genetics, v51, pages 354–362 (2019)). Here we present advancements to our methods and performance benchmarks that show significant improvements in indel detection performance compared to non-graph or other graph-based approaches while maintaining competitive performance in single-nucleotide variant detection, based on head-to-head comparisons using publicly available whole genome sequencing data such as Genome-in-a-Bottle. In terms of computational performance (runtime & cost), our graph-based tools are competitive with current best practice tools for whole genome sequence analysis, such as BWA & GATK4, without requiring specialized hardware due to our memory-optimized implementation.

A- A+

PgmNr 15: Candidate variant discovery using graph genomes: Leveraging familial genetic structures to improve detection of causal variation to rare diseases.

Authors:
C. Markello [1]; J. Eizenga [1]; A. Novak [1]; E. Garrison [1]; G. Hickey [1]; J. Siren [1]; X. Chang [1]; J. Sibbesen [1]; J. Monlong [1]; R. Rounthwaite [1]; B. Pusey [2]; T. Markello [2]; C. Lau [2]; D. Adams [2]; W. Gahl [2]; B. Paten [1]

View Session Add to Schedule

Affiliations:
1) University of California, Santa Cruz, Santa Cruz, CA.; 2) NIH Undiagnosed Diseases Network, NHGRI, Bethesda, MD

Traditional methods that use a linear reference for analyses of whole genome sequencing data have been found to be inadequate for detection of structural variants, rare variation and variants that originate in high-complexity and repetitive regions of the human genome. Over the last few years our lab has developed methods for leveraging common human variation for the purpose of improving read mapping and calling of variants in the difficult-to-analyse regions of the genome.

The Genomics Institute of the University of California at Santa Cruz (UCSC) in collaboration with the Undiagnosed Diseases Program (UDP) of the NHGRI have developed a workflow for detecting candidate variants that are causal to rare genetic disorders. The software developed at UCSC leverages new techniques provided by the Variation Graph (VG) toolkit to encode human genetic variation of pedigrees for improved mapping and variant calling capabilities. Techniques developed by the UDP and published for use in exome data have been adapted for genomic data produced by the VG graph toolkit. The entire pathway is intended to detect causal variants in cases with previously negative clinical exome testing. For ease-of-use, portability and scalability purposes, the software was built using the Broad Institute’s Cromwell and WDL software framework and is hosted on Github and Dockstore. Code for candidate analysis was written in Java for software portability and is also on Github. We have applied this software to detect candidate variants in 55 whole-genome quartet nuclear families. Each family contains both unaffected parents, at least one unaffected sibling and only a single individual that expresses an undiagnosed genetic disease. There are no known cases with the same phenotype. A previous test using exomes demonstrated that a significant excess of deleterious candidates were found in the proband group versus an equally sized unaffected sibling group. The results of the present analysis are consistent with the exome findings, even with a much larger potential space to produce false positive candidates. Results show improvement in the search space for candidate variants when using parental genomes over traditional linear-reference based methods. This provides new opportunities to search regions that have previously been difficult to study using whole genome sequencing data.

A- A+

PgmNr 16: Genetic control of the human brain proteome.

Authors:
C. Robins [1]; W. Fan [1]; D. Duong [2]; J. Meigs [1]; E. Gerasimov [1]; D. Cutler [3]; E. Dammer [2]; P. De Jager [4,5]; D. Bennett [6]; J. Lah [1]; A. Levey [1]; N. Seyfried [2]; A. Wingo [7,8]; T. Wingo [1,3]

View Session Add to Schedule

Affiliations:
1) Department of Neurology, Emory Univ School of Medicine, Atlanta, GA; 2) Department of Biochemistry, Emory Univ School of Medicine, Atlanta, GA; 3) Department of Human Genetics, Emory Univ School of Medicine, Atlanta, GA; 4) Cell Circuits Program, Broad Institute, Cambridge, MA; 5) Center for Translational and Computational Neuroimmunology, Department of Neurology, Columbia University Medical Center, New York, NY; 6) Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL; 7) Division of Mental Health, Atlanta VA Medical Center, Decatur, GA; 8) Department of Psychiatry, Emory University School of Medicine, Atlanta, GA

Alteration of the brain proteome is thought to be important in neurodegenerative diseases, but little is known about the genetic variation that controls protein abundance in the brain. To identify SNPs that underlie variation in protein abundance in the human brain, we performed protein quantitative trait loci (pQTL) analyses using tandem mass tag (TMT) protein data from the dorsolateral prefrontal cortex (DLPFC) and whole genome sequencing of 144 cognitively unimpaired older participants of the Religious Order Study (ROS) and Memory and Aging Project (MAP). The cis-genetic control of 8,002 proteins was tested using linear regression to model protein abundance as a function of genotype for each SNP in within a 10-kb window around the corresponding protein-coding sequence. Each regression assumed additive genetic effects and included age at death, sex, post-mortem interval, study, genetic principal components, and estimated proportions of brain cell types as covariates. We identified 100 SNPs significantly associated with the abundance of 78 proteins (Bonferroni threshold: 3.8×10^{-7}). These results were compared to expression (RNA) quantitative trait loci (eQTL) analyses performed using RNA-sequencing data from the DLPFC of 169 cognitively unimpaired ROS and MAP participants. Using similar statistical procedures to our pQTL analysis, we find 1,460 SNPs significantly associated with the expression of 790 genes (Bonferroni threshold: 2.0×10^{-7}). Only 10 sites were both eQTL and pQTL associated. While this level of overlap is itself statistically significant (Fisher's exact test: $p = 1.3 \times 10^{-14}$), and 9 out of 10 sites had an effect direction consistent between eQTL and pQTL, the vast majority of eQTLs are not pQTLs (97%; 347 of the 357 eQTLs also tested in the pQTL analysis 97%), and vice versa (79%; 37 of the 47 pQTLs also tested in the eQTL analysis). Our results suggest that if one believes that protein dysregulation is important in neurodegenerative diseases, the eQTL analysis of protein coding loci in the brain should be treated with caution, as eQTL significance does not often translate to pQTL significance.

A- A+

PgmNr 17: Genomic architecture of 184 plasma proteins in 20,000 individuals: The SCALLOP Consortium.

Authors:
J. Wilson [1,2]; E. Macdonald-Dunlop [1]; P.K. Joshi [1]; J.E. Peters [3]; L. Folkersen [5]; I. Ingelsson [6,7,8]; K. Michaelsson [9]; S. Gustafsson [10]; S. Enroth [11]; A. Johansson [11]; G. Smith [12]; D. Zhernakova [13]; A. Siegbahn [14]; A. Kalnapenkis [3,15]; N. Eriksson [16]; J. Fu [13]; L. Franke [13]; C. Hayward [2]; L. Wallentin [14]; T. Esko [15,17]; E. Zeggini [18]; C. Teunissen [19]; O. Hansson [20,21]; P. Eriksson [22]; U. Gyllensten [10]; A.S. Butterworth [3]; A. Mälarstig [22,23]; on behalf of the SCALLOP Consortium

View Session Add to Schedule

Affiliations:
1) Usher Institute for Population Health Sciences and Informatics, Univ Edinburgh, Edinburgh, United Kingdom; 2) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom; 3) Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Worts Causeway, Cambridge, United Kingdom.; 4) Health Data Research UK, United Kingdom.; 5) Institute of Biological Psychiatry Copenhagen, 2000 Denmark; 6) Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA 94305; 7) Stanford Cardiovascular Institute, Stanford University, Stanford, CA 94305; 8) Stanford Diabetes Research Center, Stanford University, Stanford, CA 94305; 9) Department of Surgical Sciences, Uppsala University, Uppsala, Sweden.; 10) Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden.; 11) Department of Immunology, Genetics, and Pathology, Biomedical Center, Science for Life Laboratory (SciLifeLab) Uppsala, Uppsala University, Uppsala, Sweden.; 12) Department of Cardiology, Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden.; 13) Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.; 14) Department of Medical Sciences and Uppsala Clinical Research Center, Uppsala University, Uppsala Sweden.; 15) Estonian Genome Center, University of Tartu, Estonia.; 16) Uppsala Clinical Research Center, Uppsala University, Uppsala Sweden.; 17) Broad Institute of MIT and Harvard; 18) Helmholtz Zentrum München, Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH),Ingolstädter Landstr. 1, 85764 Neuherberg.; 19) Department of Clinical Chemistry, Amsterdam Neuroscience, Amsterdam UMC, Vrije Universiteit Amsterdam, 1081HZ Amsterdam, The Netherlands.; 20) Clinical Memory Research Unit, Department of Clinical Sciences, Lund University, Malmö, Sweden.; 21) Memory Clinic, Skåne University Hospital, Malmö, Sweden.; 22) Department of Medicine, Karolinska Institutet, Stockholm, Karolinska University Hospital, Solna, Sweden.; 23) Pfizer Inc, USA.

Proteins are the fundamental building blocks of life and participate in all biological processes. Plasma proteins show great promise as novel disease biomarkers, and their genetic determinants also help unravel underlying networks and causal pathways, pointing to new drug targets. However, to date, discoveries have been limited due to small sample sizes. The SCALLOP consortium of 20 cohorts was established to discover protein quantitative trait loci (pQTLs) in a large combined sample. We used Olink proximity extension assays to quantitate the abundances of 184 plasma proteins (from the CVD II & III panels) in up to 19,578 subjects.

Genome-wide association meta-analysis revealed 22,518 genome-wide significant SNPs ($P<2.72x10^{-10}$) associated with at least one protein, including both cis- and trans-pQTLs. Contrary to findings from previous smaller individual studies, our increased sample size reveals the majority of proteins to have at least one pQTL.

Utilising a variety of state-of-the-art methodologies, we show how a subset of these pQTL co-localise with eQTLs and disease-related traits. We also elucidate novel regulatory mechanisms and protein-protein interaction networks from our greatly expanded catalogue of trans-pQTL. Finally, we apply two-sample Mendelian randomisation to explore causal mechanisms in a broad set of complex diseases and risk factors.

A- A+

PgmNr 18: A genome-wide association study reveals 51 novel loci of human metabolome in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).

Authors:
E.V. Feofanova [1]; H. Chen [1]; Q. Qi [2]; R.C. Kaplan [2,3]; M.L. Grove [1]; K.E. North [4,5]; Y. Dai [6]; P. Jia [6]; C.C. Laurie [7]; M. Daviglus [8]; J. Cai [9]; E. Boerwinkle [1]; B. Yu [1]

View Session Add to Schedule

Affiliations:
1) Human Genetics Center, University of Texas, Health Science Center, Houston, TX, USA; 2) Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA; 3) Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; 4) Department of Epidemiology, University of North Carolina Gilling School of Global Public Health, Chapel Hill, NC, USA; 5) Carolina Center of Genome Sciences, University of North Carolina, Chapel Hill, NC, USA; 6) Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA; 7) Department of Biostatistics, University of Washington, Seattle, WA, USA; 8) National Heart, Lung and Blood Institute (NHLBI), National Institutes of Health, Bethesda, MD, USA; 9) Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA

Variation in levels of the human metabolome reflect changes in homeostasis, providing a window into health and disease. Previous studies have identified multiple genetic loci associated with levels of metabolites; the genetic impact on circulating metabolites in US Hispanics, a population with disproportionately high cardiometabolic disease burden, is largely unknown.

We conducted genome-wide association analyses using 1000G imputation genotypes on 640 circulating metabolites (quantified by liquid chromatograph-mass spectrometry) in 3926 participants from the HCHS/SOL. Replication was performed using the Atherosclerosis Risk in Communities (ARIC) Study and/or publicly available databases. We identified and replicated 51 novel variant-metabolite pairs (P-value<1.2e-10, MAF≥1%), and reproduced 281 previously reported loci-metabolite associations. The estimated heritability for 640 metabolites ranged between 0-54%. The identified variants explain 1-22% of variance of the corresponding metabolites, with half of the variants located in genes, including 5 nonsynonymous variants. Some of the identified genes are directly involved in metabolite conversion (*PCMT1*), catabolism (*FAAH*), and inactivation (*FOLH1*). For example, a nonsynonymous variants rs324420, belonging to a degrading enzyme of endocannabinoids, FAAH, was associated with higher levels of an endocannabinoid N-oleoyltaurine.

For all detected loci-metabolite pairs, we performed co-localization analyses using eQTLs from available tissues (GTEx V7). We identified co-localization at 59 novel and 28 known genetic regions. A novel variant rs5855544, upstream of *SLC51A* (intestinal transporter for steroid-derived molecules), was associated with higher levels of four steroid sulfates (androsterone sulfate, epiandrosterone sulfate, 16a-hydroxy DHEA3-sulfate and androsteroid monosulfate), and co-localized with expression levels of SLC51A in transverse colon and terminal ileum. An intergenic variant rs2014127, associated with lower N-acetyl-aspartyl-glutamate (NAAG) levels, co-localized with expression levels of PSMC3, involved in protein catabolism. *PSMC3* was also identified by PPI network analysis. DAVID enrichment analysis suggests macromolecule catabolic process pathway for NAAG.

The data document the genetic architecture of circulating metabolomics analytes in an underrepresented Hispanic/Latino community. The identified loci are involved in various metabolic processes, thus shedding new light on disease etiology.

A- A+

PgmNr 19: Quantitative proteomics as a complementary diagnostic tool for Mendelian disorders.

Authors:
R. Kopajtich [1,2]; C. Ludwig [3]; C. Mertes [4]; C. Meng [3]; V.A. Yépez [4]; L.S. Kremer [1,2]; M. Gusic [1,2]; A. Nadel [1,2]; D. Smirnov [1,2]; J. Behr [3]; K. Murayama [5]; T.M. Strom [1,2]; B. Küster [6]; J.A. Mayr [7]; D. Rokicki [8]; S. Wortmann [2,7]; J. Gagneur [4]; H. Prokisch [1,2]

View Session Add to Schedule

Affiliations:
1) Institute of Human Genetics, Helmholtz Center Munich, Munich, Germany; 2) Institute of Human Genetics, Technical University Munich, Munich, Germany; 3) Bavarian Center for Biomolecular Mass Spectrometry, BayBioMS, Technical University Munich, Freising-Weihenstephan, Germany; 4) Department of Informatics, Technical University Munich, Garching, Germany; 5) Department of Metabolism, Chiba Children's Hospital, Chiba, Japan; Chiba Cancer Center Research Institute, Chiba, Japan; 6) Chair of Proteomics and Bioanalytics, Technical University Munich, Freising-Weihenstephan, Germany; 7) Department of Paediatrics, Paracelsus Medical University, SALK, Salzburg, Austria; 8) Department of Pediatrics Nutrition and Metabolic Diseases, The Children's Memorial Health Institute CMHI, Warsaw, Poland

The care of rare Mendelian diseases has been revolutionized by genome sequencing. However, across a large variety of Mendelian diseases, analysis of the coding sequence does not lead to a diagnosis for 50-75% of patients. This indicates that in many cases pathogenic variants evade detection or were detected but remained of uncertain signi?cance (VUS).

We and others recently demonstrated that combining DNA and RNA sequencing can increase the diagnostic yield by 10-30% through detection of expression outliers, monoallelic expression or aberrant splicing events. Still, many pathogenic alterations cannot be seen at the RNA level, but may affect protein folding and stability. To address this, we investigated protein levels in fibroblast cell lines using antibodies specific for the proteins affected by pathogenic mutations and found reduced levels in 95 out of 103 cases. To develop a diagnostic tool, we established a protocol for quantitative deep proteome analysis using TMT-10plex labeling and trimodal mixed phase fractionation combined with the MultiNotch MS3 method for peptide quantification that allows us to quantify about 7.800 proteins per sample.

Systematic analysis of protein outliers from 120 samples confirmed 40 RNA outlier and detected >100 protein only outliers, providing new candidates for pathogenic VUS. These outliers facilitate interpretation of functional consequences of missense variants or small in-frame insertions and deletions. Clinical interpretation of those candidates delivered a diagnosis in 10 cases (8.3%).

Moreover, we identified several cases where reduced levels of an affected protein resulted in reduced levels of interaction partners within known protein complexes. Compound heterozygous missense and 5’-UTR mutations in *MRPL38*, encoding a mitochondrial ribosomal protein resulted in a decrease of many large ribosomal subunits and as a consequence reduced levels of translated proteins. Thus proteomic data also immediately provides functional evidence for the underlying pathomechanisms.

In summary, quantitative proteomics is a powerful complementary tool to genome and transcriptome sequencing. It delivers functional data for interpretation of VUS at the level of the affected protein, it provides insights into disease mechanisms and increases the diagnostic rate for Mendelian disorders.

A- A+

PgmNr 20: Patient and public preferences on being recontacted with updated genomics results: A mixed methods study.

Authors:
Y. Bombard [1,2]; A. Sebastian [1,2]; C. Mighton [1,2]; M. Clausen [2]; S. Muir [2]; S. Shickh [1,2]; N. Baxter [1,2]; A. Scheer [1,2]; T.H. Kim [4]; D. Regier [6]; E. Glogowski [7]; K. Schrader [6]; R.H. Kim [3,4,5]; J. Lerner-Ellis [2,5]; A. Bayoumi [1,2]

View Session Add to Schedule

Affiliations:
1) University of Toronto, Toronto, ON, Canada; 2) St. Michael's Hospital, Toronto, ON, Canada; 3) University Health Network, Toronto, ON, Canada; 4) The Hospital for Sick Children, Toronto, ON, Canada; 5) Mount Sinai Hospital, Sinai Health System, Toronto, ON, Canada; 6) BC Cancer Center, Vancouver, BC, Canada; 7) GeneDX, Gaithersburg, MD, USA

Background: ASHG guidelines strongly recommend recontacting a research participant if reinterpretation of a genomic variant is related to a condition under study and is expected to change their medical management. However, research participants’ preferences for recontact are not well known.

Aim: Characterize participants’ preferences for recontact and the factors driving their preferences.

Methods: We developed a survey with a discrete choice experiment (DCE) to evaluate participants’ preferences for incidental sequencing results. Recontact was one of several attributes related to preferences for receiving results and was categorized into 4 levels: doctor updates you, login to online database for updates, contact your doctor for updates, or no updates provided. Semi-structured interviews in the pre-test of the DCE explored participants’ preferences for being recontacted. Interviews were analyzed using qualitative description. The DCE survey results (n=1000) will be presented at the ASHG conference.

Results: We conducted interviews with 31 participants 11 cancer and 20 public. Preferences were consistent between cancer patients and members of the public. Participants responded favorably to being recontacted; many assumed that they would be recontacted with updates. While most participants considered updates to have personal and clinical utility, they would still be willing to receive initial results without future updates because they valued the genomic information. The few participants who did not want to be recontacted anticipated that updates would cause them anxiety. Many preferred updates delivered through a database. Participants’ prior negative healthcare experiences, such as their doctor not following up with their test results, led to a desire for “control” and access to updates via database. Participants who had more trust in their physician preferred clinician-involved delivery of updates. A few were indifferent to how updates are delivered. Participants also recognized feasibility challenges related to recontact, such as added burden to providers.

Conclusion: Many, but not all, of our study participants assumed they would be recontacted with

updated results and prefered to have these updates delivered by accessing a database. Past healthcare experiences are important determinants of preferences for recontact. If confirmed, these findings could inform the development of strategies to optimize delivery of updated genomic results to patients.

A- A+

PgmNr 21: Participant perceptions on return of secondary findings in a clinical research setting: Low decisional conflict and potential need for targeted education and counseling.

Authors:
M. Similuk [1]; J. Yan [2]; L. Jamal [2, 3]; M. Walkiewicz [2]; M. Lenardo [1]; H. Su [1]

View Session Add to Schedule

Affiliations:
1) NIAID, NIH, Bethesda, MD.; 2) Medical Science and Computing, Bethesda, MD.; 3) NIH Bioethics Department, Bethesda, MD.

Background: The most appropriate strategies for managing genomic secondary findings (SF) in a research setting is an area of controversy, prompting a need to understand patient perceptions and preferences. We sought to assess the degree of decisional conflict regarding receipt of SF reported by research participants enrolled in genomic research at the National Institute of Allergy and Infectious Diseases when electing to learn SFs as well as what attributes are correlated with decisional conflict. All participants received genetic counseling.
Methods: Cross-sectional survey done after consent and before return of results.
Results: Seventy-six of 116 eligible participants returned the survey, for a response rate of 66%. None of the participants approached for this study explicitly declined due to or asked to opt-out of SF receipt. When asked generally about receiving SFs in genomic research projects, 74/76 (97%) survey participants reported thinking it was appropriate to return SFs; two participants were unsure. Twenty-seven (35.5%) of participants reported zero decisional conflict regarding receiving SFs. Most participants (75-88%) reported agreeing or strongly agreeing with statements related to perceived decision quality. Lower genetic literacy was weakly associated with higher total decisional conflict (r=-0.218,p=0.049) and multiple sub-variables of decisional conflict, as well as a lower reported capacity to 'deal with' primary findings (r=0.328, p=0.004) and SFs (r=0.287, p=0.012).Additionally, a notable minority of participants reported confusion about basic aspects of the decision, such as being unsure when asked if they chose to receive SFs or not (n=4, 5.3%). This confusion was correlated with overall decisional conflict (t statistic=-2.526, 7.2 df, p=0.038). Participants reported a high perceived likelihood of receiving a positive result, with 46.1% and 42.1% of participants reporting it was likely or very likely that they "receive a genetic cause for their/their child's immune system disorder" and a SF, respectively. Risk perception for primary and SFs were correlated (r=0.428,p<0.0001).
Conclusions: These data support the acceptability of genomics SFs return for participants in this study and suggests some participants may need further support and education in understand genomics concepts, as well as forming accurate risk perceptions and expectations of sequencing.

This research was supported by the intramural research program at NIAID, NIH.

A- A+

PgmNr 22: Variation in intention to participate in genetic research among Hispanic/Latinx populations by Latin America birth-residency concurrence: A global study.

Authors:
J.G. Perez-Ramos [1]; T.D. Dye [1]; I.D. Fernandez [3]; C.M. Velez Vega [5]; D. Vega Ocasio [6]; E. Avendaño [2]; N.R. Cardona Cordero [1]; C. Di Mare Herring [2]; Z. Quiñones Tavarez [6]; A. Dozier [3]; S. Groth [4]

View Session Add to Schedule

Affiliations:
1) School of Medicine and Dentistry, Department of Obstetrics and Gynecology, University of Rochester, NY; 2) Universidad de Ciencias Medicas, San Jose, Costa Rica; 3) School of Medicine and Dentistry, Department of Public Health Sciences, University of Rochester, NY; 4) School of Nursing, University of Rochester, NY; 5) Escuela Graduada de Salud Publica, Universidad de Puerto Rico, PR; 6) School of Medicine and Dentistry, Translational Science Program, University of Rochester, NY

Background: Hispanic/Latinx (H/L) populations are underrepresented in clinical and genetic research (GR). This lack of inclusion in GR violates the Belmont Report’s construct of justice (i.e., distributional justice) in that by not participating in GR, H/L populations risk not benefiting from the outputs of the research process.

Methods: We conducted a global study using Amazon Mechanical Turk(mTurk). Race and ethnicity was provided by 1,718 respondents from 69 countries, 251 (14.6%) of whom identified as Hispanic or Latin American and Caribbean (LAC). H/L respondents were further classified as: 1) Born and live outside of LAC (56.2%), 2) Born within but live outside LAC (23.9%), and 3) Born and live within LAC (19.9%). We ascertained a likelihood of participating in genetic research with a range of genetic attitudes and psychometric scales.

Results: More than half of those respondents self-identifying as H/L indicated they would participate in research that used their DNA, a similar level as non-H/L (52.8% v. 56.2%). Analysis of birth-residency subtypes indicates that respondents born in LAC and living outside of it and those born and living within LAC more commonly reported they would participate in GR when compared with H/L respondents from outside LAC (53.3% and 70.0%, v. 46.4%). The US-born and US-living H/L diaspora were significantly (p=.037) less likely to participate in research (46.5%) than the US-born/living non-H/L population. Respondents indicating they would participate in GR were significantly more likely to trust researchers (<.05), believe that GR could lead to better understanding of disease (<.05), and felt that GR could lead to new treatments (p<.05) when compared with respondents not interested in participating. Adjusting for Genetic Test Experience Score and Genetic Research Beliefs Score, the adjusted OR for GR participation among H/L born and living in LAC compared with others remained significant (OR: 2.38; 95% CI: 1.04, 5.42, p<.05).

Discussion: H/L, overall, were equally likely to indicate they would participate in GR as non-H/L. Respondents born in and living in LAC were significantly more likely to participate in GR when compared with other H/L Diaspora subgroups. In particular, the H/L populations born/living in the USA were least likely of all to indicate they would participate in GR, had the least positive attitudes toward

GR, were most distrustful of researchers, and expected the least benefit to society from GR.

A- A+

PgmNr 23: Utilization of a post-result follow-up chatbot and family sharing tool among patients receiving clinically actionable exome sequencing results.

Authors:
T. Schmidlen; C. Jones; C. McCormick; E. Vanenkevort; A. Sturm

View Session Add to Schedule

Affiliation: Genomic Medicine Institute, Geisinger, Danville, Pennsylvania.

Reducing morbidity and mortality via population genomic screening may be possible if patients take risk-reducing actions before disease onset and share risk knowledge with family. Geisinger’s MyCode$^{®}$ Community Health Initiative is a large research biobank returning actionable exome sequencing results to participants at risk for cancer, cardiac and other heritable diseases. To assist post-disclosure follow-up and cascade testing, Geisinger and Clear Genetics, Inc. developed chatbots deployable by link via electronic health record patient portal (MyGeisinger), email, text, or messenger. Chatbots are a technology-based simulated conversation used to scale communications. The follow-up chatbot reminds probands of actions to take after result receipt (see doctor, share with family). The Family Sharing Tool (FST) is a launching tool for probands to send a cascade chatbot to their family that describes the result, disease risks, management, and how to get cascade testing. During result disclosure, consent to receive the chatbot and communication preference (text, email, MyGeisinger) is collected. Probands receive the FST 2 weeks post-disclosure, a follow-up chatbot 1 month post-disclosure, and are offered the FST again in the follow-up chatbot. Chatbot uptake was tracked from August 2018-March 2019: 106/195 (54%) consented and 89/195(46%) declined. Receipt preferences were: MyGeisinger (54%), email (28%), text (18%). 101 patients received a follow-up chatbot, 54% opened it and 64% of those who opened, completed. Younger patients were more likely (p=.001) to consent. There were no differences within sex (p=.30) and no age differences in likelihood to complete a follow-up chatbot (p=.49). MyGeisinger users were more likely (p< .001) to consent, but MyGeisinger receipt preference did not affect likelihood (p=.40) to complete. Most (30/36, 83%) said in the chat that they already shared with family by talking (83%), letter (57%), FST (10%), or other (7%). The chatbot encourages use of the FST to supplement information already shared. Most reported not having other family to share with (27/34, 79%) but 63% (17/27) accepted the FST and 70% (12/17) used it after the chat. The MyCode$^{®}$ follow-up chatbot may be an acceptable, scalable tool to follow-up with patients receiving genomic risk information and to increase proband sharing with at-risk family members. Future work will determine if risk-reducing behaviors and cascade testing uptake is greater among chatbot users.

A- A+

PgmNr 24: Rare germline functional variant in *ARHGAP30* gene predisposes to Li-Fraumeni syndrome-like cancers.

Authors:
R. Krahe [1,2,3]; J.W. Wong [1,2]; Y. Deng [1]; L.L. Bachinski [1]; S.E. Olufemi [1]; Q. Chen [1]; J. Hsu [1]; M. Sirito [1]; Y. Wang [4]; K.A. Baggerly [4]; S.T. Arnold [5]; J.E. Ladbury [5]; P. Gang [6]; W. Wang [6]; B.R. Gracia [1]; G.I. Karras [1]; H. Hampel [6]; A. de la Chapelle [6]; P.L. Mai [7]; S.A. Savage [7]; C.L. Snyder [8]; H.T. Lynch [8]; J. Bojadzieva [1]; G. Lozano [1,2,3]; L.C. Strong [1,2,3]

View Session Add to Schedule

Affiliations:
1) Dept. of Genetics, University of Texas MD Anderson Cancer Center, Houston, TX; 2) Program of Genetics & Epigenetics, MD Anderson Cancer Center/UTHealth Graduate School of Biomedical Sciences, Houston, TX; 3) Center of Cancer Genetics & Genomics, University of Texas MD Anderson Cancer Center, Houston, TX; 4) Dept. of Bioinformatics & Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX; 5) Center for Biomolecular Structure & Function, University of Texas MD Anderson Cancer Center, Houston, TX; 6) Div. of Human Genetics, Dept. of Internal Medicine & Comprehensive Cancer Center, Ohio State University, Columbus, OH; 7) Clinical Genetics Branch, Div. of Cancer Epidemiology & Genetics, NCI, NIH, Bethesda, MD; 8) Hereditary Cancer Center, Dept of Preventive Medicine, Creighton University, Omaha, NE

Li-Fraumeni Syndrome (LFS) is a rare, clinically and genetically heterogeneous cancer predisposition syndrome characterized by a diverse tumor spectrum, including a high prevalence of sarcomas, breast, brain and adrenal gland cancers. Most cases characterized to date that meet the classic criteria are caused by autosomal dominant germline mutations in the tumor suppressor gene *TP53* (p53) on chromosome 17p13.1. However, a subset of patients and families that phenotypically meet the classic or variously relaxed LFS criteria lack pathogenic *TP53* mutations. Using a 5-generation pedigree of European ethnicity with 13 samples from 3 generations (including 7 affecteds, 4 obligate carriers, estimated penetrance 55%), we had mapped a novel LFS-like (LFL) predisposition locus to a 3.8-Mb region in chromosome 1q23.2-q23.3. Here, we used whole genome/exome sequencing and segregation analysis to identify a co-segregating rare germline functional missense variant in *ARHGAP30*, encoding a Rho GTPase-activating protein, in the same region, while excluding the remainder of the genome. Genetic testing of 47 additional probands from independent LFL families identified 3 more of European ethnicity (including 6 affecteds) with apparent co-segregation of the same candidate pathogenic variant (for a total of 4 of 48 families, 8.3%) and significant enrichment over the general population (gnomAD Global MAF=1.843%; European MAF=2.614%, *p*-value=3.87e-2). *In silico* structural analysis suggested the missense variant potentially disrupts a consensus protein binding site and sites for post-translational modifications. *In vitro* functional testing of the missense variant in the short and long ARHGAP30 protein isoforms generated by site-directed mutagenesis indicated that the variant proteins increase cell migration and proliferation, consistent with the previously reported main molecular and cellular functions. *ARHGAP30* was recently identified as a novel tumor suppressor gene in colorectal and lung cancer; it functions as a scaffolding protein for p300-mediated p53 acetylation and activation to genotoxic stress. Analysis of existing TCGA sporadic tumor datasets for the same pathogenic variant showed an enrichment in multiple sporadic cancers that are part of the LFS tumor spectrum, further underscoring the identification of *ARHGAP30* not only as a novel LFL cancer predisposition gene, but also as a gene with possible relevance to

multiple sporadic cancers, similar to mutant *TP53*.

A- A+

PgmNr 25: Li-Fraumeni syndrome? A multi-tissue NGS strategy to define constitutional TP53 status.

Authors:
J.N. Weitzel [1]; J. Garber [1]; D. Castillo [1]; S. Sand [1]; R. Mejia [1]; A. Cervantes [1]; K.W.K. Tsang [1]; J. Mokhnatkin [1]; J. Wang [2]; X. Wu [2]; J. Herzog [1]; B. Nehoray [1]; T.P. Slavin [1]

View Session Add to Schedule

Affiliations:
1) Division of Clinical Cancer Genomics, City of Hope, Duarte, CA; 2) Integrative Genomics Core, City of Hope, Duarte, CA

Purpose: Pathogenic germline *TP53* variants (PV) underlie the rare Li Fraumeni Syndrome (LFS), associated with predisposition to sarcoma, brain, breast, adrenocortical and other malignancies at unusually early ages, but about which many questions remain unanswered and new questions are emerging. NGS-based multi-gene panel testing (MGPT) including *TP53* is now being performed on many people who do not meet LFS criteria and has raised concerns about: 1) a broader phenotypic spectrum for germline PV; and 2) the clinical relevance of PV identified in blood or saliva with variant allele frequencies (VAF) below the 50% expected for a germline carrier. We demonstrated frequent occurrence of aberrant clonal expansion (ACE), most often from hematopoietic clones (CH) with an acquired PV, spontaneously with age or after exposure to chemotherapy; this must be distinguished from germline or constitutional PVs, true post-zygotic mosaicism (PZM) and from contaminating malignant cells, given the very different clinical implications. We tested a multi-tissue NGS strategy as a tool to interrogate constitutional *TP53* status.
Methods: Among 105 cases, consented and enrolled in the Clinical Cancer Genomics Community Research Network registry, with MGPT-detected *TP53* PVs, some of which had skewed VAF results, we obtained additional tissues on 31 cases. DNA extracted from formalin fixed paraffin embedded (FFPE) tumor/non-tumor tissues, blood, saliva, eyebrow plucks, was analyzed using a custom myeloid and CH gene (n=79) amplicon-based QIAseq panel. PVs with VAF> 2% were included in analyses.
Results: With an average read depth of 275X, VAF ranged from 10.55%- 45.56%. Within-subject standard error was 3%, with no significant excess variability (P = 0.17). Of 31 cases, 25 cases had comparable samples: 12 had results supporting ACE/CH, with additional CH-associated PV(s) identified in 5/12 (41%); n=2 of each *TET2*, *ATM*, *TP53*; and increasing VAF over time for the driver *TP53* PV was noted in 2. Germline status was confirmed for 6 cases (one with a CH PV), post-zygotic mosaicism was supported for 5 cases and 2 were indeterminant.
Conclusions: Our multi-tissue NGS strategy appears valid, with reproducible VAF and ability to discern constitutional *TP53* status. This work has direct translational impact, refining risk estimation and improving the clinical care of patients with *TP53* PVs, while avoiding unnecessary LFS-related care and enabling appropriate care for those with ACE.

A- A+

PgmNr 26: Comprehensive functional classification of Lynch syndrome missense variants.

Authors:
Jia [1]; B. Burugula [1]; V. Chen [1]; M. Maksutova [1]; S. Jayakody [1]; J. Kitzman [1,2]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics,; 2) Department of Computational Medicine and Bioinformatics, Univ MICHIGAN, Ann Arbor, Michigan.

The rapid expansion of clinical genetic testing has shifted the bottleneck in human genetics from data acquisition to variant interpretation. For many clinically actionable genes, there is a large burden of individually rare variants of uncertain significance (VUS), particularly missense variants for which the impact upon the encoded protein remains undetermined and can vary from loss of function to entirely benign. We selected one such gene, *MSH2*, which is frequently mutated in Lynch Syndrome, the highest prevalence inherited cancer risk syndrome (1:279 individuals). We systematically classified variant function by synthesizing and introducing into human cells a library containing every possible single-codon variant of *MSH2* full-length cDNA (N=58,842). Targeted deep sequencing of *MSH2* from the resulting cellular population indicated that >95% of all possible variants were stably integrated, representing the largest human gene subjected to full saturation mutagenesis and deep mutational scanning to date in a mammalian cellular model. We performed mismatch repair activity-dependent screening to measure the molecular function of each *MSH2* mutation in cells, using as a read-out ultra-deep sequencing of the mutant *MSH2* library before and after drug selection. After stringent data quality control and filtering, we arrived at loss-of-function scores for 16,588 (~93.5%) single amino-acid substitution missense variants of *MSH2*. These scores were bimodally distributed, and indicated that most missense alleles of *MSH2* (88.7%) retain mismatch repair function, while a small minority (11.3%) are functionally impaired. These functional scores showed near-perfect concordance with published biochemical characterization of individual variants. Additionally, these high-throughput measurements show strong agreement with the expert-panel reviewed classification of *MSH2* missense variants in ClinVar. This dataset will enable greatly improved interpretation of clinically observed variants of *MSH2*, towards prospective, genotype-guided early detection and intervention for inherited colorectal cancer.

A- A+

PgmNr 27: Disparate genes with germline pathogenic and likely pathogenic variants converging on p53 network can be etiologic for pediatric/adolescents and young adults solid tumor cancer predisposition.

Authors:
S. Akhavanfard [1,2]; R. Padmanabhan [1]; L. Yehia [1]; F. Chen [1]; C. Eng [1,2,3,4]

View Session Add to Schedule

Affiliations:
1) Genomic Medicine Institute, Cleveland Clinic Lerner Research Institute, Cleveland, OH; 2) Cleveland Clinic Lerner College of Medicine, Case Western Reserve University School of Medicine, Cleveland; 3) Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH; 4) CASE Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH

Cancer is believed to be caused by an accumulation of DNA damage beyond repair mechanisms, resulting in inevitable “wear and tear” that increases with age. In contrast to adult cancers, environmental risk factors are less relevant in Children, Adolescents and Young Adults (C-AYA) who develop cancer. Although there has been substantial advancement in understanding somatic mutations in cancers, our knowledge about the spectrum and implications of germline mutations in C-AYA especially with solid tumors (STs) are limited. We hypothesized that C-AYA patients with STs are likely to have germline mutations in a specific spectrum of genes making them more susceptible to the development of very early-onset cancers. We performed variant-prioritization analysis and ACMG classification on germline exome sequencing data of 1507 early-onset STs in patients initially diagnosed under 29 years old. We identified 17.4% of these patients carrying germline pathogenic and/or likely pathogenic (P/LP) mutations in 54 of 206 known cancer predisposition genes (KCPG), including *RB1*, *NF1*, *ERCC2*, *TP53*, while adding unexpected KCPG for adrenocortical carcinoma, astrocytoma, CNS tumors, Ewing sarcoma, neuroblastoma, osteosarcoma, retinoblastoma, soft tissue sarcoma, and Wilms tumor. An additional 56.4% had germline pathogenic mutations in other critical genes, including *PRKN*, *MCPH1*, *SMARKAL1*, *SMAD7*, which we call them “candidate” genes. Despite mutations in a broad gene spectrum, pathway analysis led to distinct top canonical pathways and networks, centering around p53. Our drug-target network analysis showed that more than one-third of patients with germline P/LP mutations have at least one druggable gene, with 31.5% of those genes having existing FDA-approved antineoplastic and immunomodulating-related compounds. More than half of these gene-target candidates are from our “candidate” gene group which would go unidentified in routine clinical care. We show here a broad spectrum of susceptibility genes that may account for C-AYA ST predisposition, but with pathways converging on the p53 network. Our drug-target network data uphold the importance of precision oncology practices for all C-AYA STs, which currently have a paucity of targeted treatments, and suggests the immediate need for C-AYA-specific “basket” (ie, by gene) clinical trials.

A- A+

PgmNr 28: Optimizing gene expression predictive performance across global populations.

Authors:
P. Okoro [1,2]; R. Schubert [1,2,5]; A. Luke [4]; L.R. Dugas [4]; H.E. Wheeler [1,2,3]

View Session Add to Schedule

Affiliations:
1) Bioinformatics Program, Loyola University Chicago, Chicago, Illinois.; 2) Biology Department, Loyola University Chicago, Chicago, Illinois.; 3) Computer Science Department, Loyola University Chicago, Chicago, Illinois.; 4) Division of Public Health Sciences, Loyola University Chicago, Maywood, Illinois.; 5) Mathematics and Statistics Department, Loyola University Chicago, Chicago, Illinois.

The tremendous progress made in understanding the genetics of complex traits through GWAS has been largely achieved in populations of European ancestry. Leveraging transcriptome-based methods like PrediXcan that identify potential gene regulatory function underlying GWAS loci, we have shown that the genetic architecture of gene expression is sparse and that genetic predictors of gene expression are more accurate in populations of similar ancestry. Here, our goal is to use machine learning to optimize gene expression prediction within and across diverse populations.

We used genotype and monocyte transcriptome data from the Multi-Ethnic Study of Atherosclerosis (MESA) in self-identified African Americans (AFA, n=233), Hispanics (HIS, n=352), and Caucasians (CAU, n=578) with nested cross-validation of elastic net to build gene expression prediction models for 9623 protein coding genes using SNPs with 1Mb of each gene. We found 1814 (AFA), 1532 (HIS), and 1753 (CAU) genes with significant cross-validated performance ($R^2 > 0.1$). *NUDT2*, *CHURC1*, and *HLA-DRB5* were consistently among the top 10 performing gene models across populations. Using the models fit in MESA ($R^2 > 0.1$), we predicted gene expression in 77 Ghanaians and African Americans from the Modeling the Epidemiologic Transition Study (METS), and found the mean Spearman correlations (r) between predicted and observed whole blood expression were 0.102, 0.047, and 0.034 with AFA, HIS, and CAU models, respectively. *ERAP2* and *JUP* had the highest r in METS using the models from each MESA population. To determine if predictive performance can be improved with other models, we focused on top performing AFA genes (elastic net mean R^2=0.786 for *NUDT2*, *HLA-DRB5*, *CHURC1*, *ERAP2*, *JUP*). For these five genes, we trained other machine learning algorithms with 5-fold cross-validation in AFA and found the mean R^2 for random forest=0.803, *k*-nearest neighbor R^2=0.402, support vector machine linear kernel R^2=0.506, support vector machine radial basis function kernel R^2=0.581.

Thus, random forest may improve gene expression predictive performance compared to elastic net. We show predictive performance increases with shared ancestry. Using varying machine learning approaches and diverse population data, we will continue to optimize models for genetically regulated gene expression.

A- A+

PgmNr 29: Local-ancestry based models for improved gene-expression prediction in African Americans.

Authors:
L. Nahlawi; Y. Zhong; T. De; C. Alacron; M.A. Perera

View Session Add to Schedule

Affiliation: Feinberg School of Medicine - Department of Pharmacology, Northwestern University, CHICAGO, Illinois.

Motivation Over the past decade, more than 10,000 robust genome-wide associations (GWA)s, between genetic- variants and many complex-traits, revealed potential drivers of disease and drug response. However, determining the underlying mechanism of action is costly and time consuming. An intermediary phenotype, namely gene expression (GE), has been adopted to facilitate the discovery of gene-based associations and identify plausible biological mechanisms for GWAS.

Background A plethora of genome-wide genotyping data is publicly available. Yet, matched GE/genotype data is harder to obtain, specifically in relevant tissues. More importantly, such data in non-European populations are only available for limited tissue types, like lymphobastoid cell lines (LCL). Most drug metabolizing enzymes are not expressed in LCLs, making them suboptimal for key precision medicine phenotypes like drug response. Many computational models have been proposed to predict the genetically-regulated component of GE such as *PrediXcan*, which models GE using GTEx genotype data through linear regression analysis. Such models are limited by training mostly on European-ancestry data and do not generalize well to other populations where ancestral differences have a key role in tissue specific GE.

Methods Our previous work has shown that incorporating local-ancestry (LA) improves eQTL mapping in ad- mixed populations. Thus we propose a LA-based model to predict GE in African Americans (AA), an admixed minority population. We extend *PrediXcan*'s framework to incorporate loci-specific inferred LA into the prediction model. For training, we use genotype and GE data from cultured primary hepatocytes isolated from 60 AA donor livers. We analyzed 7 million SNPs in 14000 genes. We train a linear model per gene to map genotype to GE levels. For each model, we generate 3 sets of predictors: dosage data for *cis*-SNPs, LA data for the respective loci, and interaction terms consisting of the product of dosage and LA data for each locus.

Results Our models led to 1323 well predicted genes in comparison to 1027 genes predicted using GTEx models. We were able to identify 11 genes, related to xenobiotic metabolism, a key determinant of drug response, including *CYP1A1*, unlike previously published models.

Conclusion Our LA-based models pave the way for a population-specific GE prediction, facilitating the elucidation of the genetic impact on complex traits in admixed populations.

A- A+

PgmNr 30: Impact of discordant ancestral structure between coding and regulatory regions on gene expression.

Authors:
W.J. Alvarez [1]; S.W. Kong [1,2]; I.H. Lee [1]

View Session Add to Schedule

Affiliations:
1) Computational Health Informatics Program, Boston Children's Hospital, Boston, MA; 2) Department of Pediatrics, Harvard Medical School, Boston, MA

Expression quantitative trait locus (eQTL) analysis showed the impact of genetic variants on gene expression regulation, which could contribute to the pathogenesis of some diseases as shown in fine-mapping analysis for common diseases. The transferability of findings from the studies in European populations into non-European populations is an outstanding question. Allele frequency spectrum for coding and regulatory regions vary between populations, which could cause variation in the gene expression regulation. Moreover, admixed populations have increased risks for some diseases such as hypertension compared to ancestral populations. We investigated the impact of genetic variants on tissue-specific gene expression levels by focusing on local ancestral structure of coding and regulatory regions using the paired whole-genome sequence (WGS) and tissue-wide transcriptome datasets from the Genotype-Tissue Expression (GTEx) project. For each WGS, ancestral origins of coding region, immediate upstream of coding region (i.e., 50Kbps upstream from the transcription start sites), distant regulatory region (i.e., 50-500Kbps), and annotated regulatory regions discovered by the ENCODE project were predicted using RFMix with reference panels from the 1000 Genomes Project. Then we checked gene expression variation explained by discordant local ancestry between coding and regulatory regions in each tissue. Among the individuals enrolled in the GTEx project, the most frequent change of predicted local ancestry was from European to African in chromosome 6p22.1, which includes the major histocompatibility complex that is highly polymorphic for immune response genes. In the left ventricle samples from 27 individuals, a total of 43 genes including 17 putative disease-associated genes had predicted admixed events in distant regulatory regions including CTCF binding sites. In these samples, expression levels of 43 genes showed significant deviation (i.e. 2 or more standard deviations) from the mean expression levels of 254 individuals without admixed events. Gene expression variation explained by discordant ancestral structure between coding and regulatory regions also showed tissue-specific pattern. Further validation studies including additional populations are required for understanding the impact of local ancestral structure on gene expression regulation for both rare and common diseases.

A- A+

PgmNr 31: Cross-population portability and statistical power of predictive models for gene expression.

Authors:
K. Keys [1]; A.C.Y. Mak [1]; M.J. White [1]; W.L. Eckalbar [1]; A.W. Dahl [1]; J. Mefford [1]; A.V. Mikhaylova [2]; M.G. Contreras [3]; J.R. Elhawary [1]; C. Eng [1]; D. Hu [1]; S. Huntsman [1]; S.S. Oh [1]; S. Salazar [1]; M.A. Lenoir [4]; J.C. Ye [5,6]; T.A. Thornton [2]; N. Zaitlen [7]; E.G. Burchard [1,6]; C.R. Gignoux [8,9]

View Session Add to Schedule

Affiliations:
1) Medicine, UCSF, San Francisco, California.; 2) Department of Biostatistics, University of Washington, Seattle, WA, USA; 3) San Francisco State University, San Francisco, CA, USA; 4) Bay Area Pediatrics, Oakland, CA, USA; 5) Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA; 6) Department of Bioengineering and Therapeutic Biosciences, University of California, San Francisco, CA, USA; 7) Department of Neurology, University of California, Los Angeles, CA, USA; 8) Colorado Center for Personalized Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA; 9) Department of Biostatistics and Informatics, School of Public Health, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA

Transcriptome-wide association studies (TWAS) analyze the contribution of genetic variants to complex traits mediated through gene expression. Public genotype-expression repositories such as the Genotype-Tissue Expression (GTEx) project allow researchers to perform TWAS by predicting gene expression levels from reference genotypes using linear models from PrediXcan. Predictive models with GTEx expression data are built largely in individuals of European (EUR) descent. This fact, combined with the availability of genotype data from EUR populations, foreshadows extensive use of transcriptome prediction models trained in EUR and applied to non-EUR populations.

We highlighted challenges with cross-population portability of eQTL-based models: while gene expression prediction depends on the genetic distance between populations and the amount of shared eQTL architecture, prediction accuracy drops sharply with any deviation from homogeneity.

We extend these results by quantifying the loss of statistical power in the absence of cross-population portability. We simulate two ancestral populations from 1000 Genomes CEU and YRI and one admixed African-American population (AA) of equal sample sizes with admixture matching observed proportions. Under a simulation framework of 1000 subjects per population with 100 genes determined by 10 cis-eQTL per gene, we perform TWAS using within- and cross-population predicted expression across a range of heritabilities.

We note three major observations. Firstly, under a highly genetically-controlled gene model of $h^2 \sim$ 95%, power for TWAS (Bonferroni *p*-value < 5 x 10^{-4}) is highest when training and testing in the same population (approaching 1), intermediate in populations with smaller genetic distance (YRI and AA, power 0.49) and smallest when genetic distance is large (CEU and YRI, power 0.23). Secondly, highly heritable genes do not always retain high power, with 9 of 100 genes showing at most 0.50 power across populations. Lastly, highly heritable genes retain some statistical power across populations even without any shared eQTL structure, likely due to local LD, which affects downstream interpretation. We extend this to discuss models across a range of heritabilities and the implications for TWAS discovery for gene expression predicted across populations. Our work underscores the pressing need for continued transcriptome generation in populations from across the world to ensure that biomedical research benefits all individuals.

PgmNr 32: Determinants of telomere length across human tissues.

Authors:
K. Demanelis [1]; F. Jasmine [1]; L.S. Chen [1]; M. Chernoff [1]; L. Tong [1]; J. Shinkle [1]; M. Sabarinathan [1]; H. Lin [1]; E. Ramirez [1]; M. Oliva [2]; S. Kim-Hellmuth [3]; B.E. Stranger [2]; K. Ardlie [4]; F. Aguet [4]; J. Doherty [5]; H. Ahsan [1,6,7,8]; M.G. Kibriya [1]; B.L. Pierce [1,6,7]; Genotype-Tissue Expression (GTEx) Consortium

View Session Add to Schedule

Affiliations:
1) Department of Public Health Sciences, University of Chicago, Chicago, IL, USA; 2) Section of Genetic Medicine, Department of Medicine, Institute for Genomics and Systems Biology, Center for Data Intensive Science, University of Chicago, Chicago, USA; 3) New York Genome Center, New York, NY, USA; 4) The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA; 5) Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA; 6) Department of Human Genetics, University of Chicago, Chicago, IL, USA; 7) University of Chicago Comprehensive Cancer Center, Chicago, IL, USA; 8) Department of Medicine, University of Chicago, Chicago, IL, USA

Leukocyte telomere length (TL) has been studied extensively as a biomarker of aging and risk for age-related diseases. However, little is known about variability in TL across human tissues. To better understand this variation and its determinants, we measured relative TL (RTL, telomere repeat abundance in a DNA sample compared to a standard) in 40 tissue types from 962 donors (age 20-70) in the Genotype-Tissue Expression (GTEx) Project.

RTL was measured for 6,391 tissue samples using a Luminex assay. We analyzed relationships between RTL and covariates using linear mixed models (across all tissues) and linear models (within tissues).

Variation in RTL was attributable to tissue (24%), donor (9%), and age (3%). RTLs were generally positively correlated among tissues. Whole blood (WB) RTL was a proxy for RTL in most tissues, with correlations ranging from 0.15 to 0.37. RTL was shortest in WB and longest in testis. RTL was inversely associated with age in most tissues, and this association was strongest for tissues with shorter average RTL. African ancestry (estimated from SNPs) was associated with longer RTL across all tissues ($p=0.007$) and within WB, cerebellum, lung, thyroid, and prostate ($p<0.05$), suggesting ancestry-based differences in TL exist in germ cells and are passed to the zygote. A polygenic TL score, based on 9 leukocyte TL-associated SNPs from prior GWAS, was positively associated with RTL in WB ($p=0.005$) and across all tissues ($p=0.006$), but there was heterogeneity in these SNP effects among tissues. Six of these TL loci showed evidence of co-localization with *cis*-eQTLs in one or more tissues, suggesting these loci impact TL via regulation of gene expression. Components of telomerase (*TERT, TERC, DKC1*), a TL maintenance enzyme, were higher expressed in testis than any other tissue, and RTL was associated with *TERT* ($p=0.01$) and *DKC1* ($p=0.004$) expression across tissues. RTL mediated the effect of age on gene expression in several tissues (between 12% and 21% of age-associated genes had $p_{mediation}<0.05$). Within specific tissue types, cell type enrichment scores (from XCell) were associated with RTL, indicating that cell composition impacts TL measurement.

These findings enhance our understanding of tissue-specific effects of aging and SNPs on TL, mitotic inheritance of TL during development, and inherited ancestry-based differences in TL. Our results have important implications for interpreting epidemiologic studies of leukocyte TL and disease.

A- A+

PgmNr 33: Thirteen novel genetic loci identified for telomere length leveraging 75K whole genome sequences in the Trans-Omics for Precision Medicine (TOPMed) Program.

Authors:
M.A. Taub [1]; J. Weinstock [2]; K. Iyer [3]; L.R. Yanek [4]; M.P. Conomos [5]; M. Arvanitis [6,7]; A.R. Keramati [4]; J. Lane [8]; T. Blackwell [2]; C. Laurie [5]; T. Thornton [5]; A. Battle [7]; J.A. Perry [9]; N. Pankratz [8]; A. Reiner [10]; R.A. Mathias [4]; on behalf of the NHLBI TOPMed Consortium

View Session Add to Schedule

Affiliations:
1) Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD; 2) Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI; 3) Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD; 4) GeneSTAR Research Program, Johns Hopkins University School of Medicine, Baltimore, MD; 5) Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA; 6) Department of Medicine, Division of Cardiology, Johns Hopkins University, Baltimore, MD; 7) Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD; 8) Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN; 9) Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD; 10) Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA

Telomere length (TL) is considered a molecular/cellular hallmark of aging. Fifteen recent genome-wide association studies (GWAS) have found 16 TL loci. These prior GWAS have two limitations: (i) almost all have been in European ancestry individuals; and (ii) all have relied on array genotype data. Therefore, very little is known about the specific causal variants, and even less about the genetic architecture of these loci in individuals with other ancestral backgrounds.

We leverage TOPMed whole-genome sequencing (WGS) data to estimate TL bioinformatically using TelSeq software in the largest multi-ethnic dataset for TL GWAS to date. Genomewide tests for association in a meta-analysis of n=46,458 discovery and n=28,718 replication samples were performed using GENESIS on 82M variants with minor allele count >= 5, adjusting for age, sex, study, sequencing center, population structure and relatedness. We identified 22 loci ($p < 5\text{x}10^{-8}$), including 9 prior and 13 novel loci. Several of the novel loci map to genes that play a role in telomere biology: *RFWD3, TERF1, TINF2, POT1, ATM, SAMHD1*, and *TERF2*. Of the top 25 pathways identified in gene set enrichment analysis for these loci ($\text{FDR} < 5.6\text{x}10^{-5}$), 24 are related to telomere length/maintenance, DNA regulation, telomere capping/loop disassembly, and telomere organization.

We estimate TL heritability to be 47%, consistent with previous reports. Stratified analysis was performed by race/ethnicity: African (n=21179), Asian (n=4754), Hispanic/Latino (n=9808), European (n=38193), and Samoan (n=1242), and several loci show population differences. In particular, *TINF2* has a strong association in the Samoan (alternate allele frequency (AAF)=0.23; $p=1.3\text{x}10^{-7}$), Asian (AAF=0.09; $p=1.3\text{x}10^{-5}$) and African (AAF=0.01; $p=2.6\text{x}10^{-4}$) groups, and no association in the European group (AAF <0.005). PheWAS of sentinel variants at *TERT* and *TERC* had associations with myeloproliferative neoplasms, cancers of skin/brain, and leiomyoma/benign neoplasms of the uterus (all $p<10^{-8}$) in the UK Biobank. Sentinel variants at *NAF1, TERF1, ZNF729, POT1*, and *CHKB-AS1* had suggestive associations with uterine fibroids (p=0.008 to 0.07).

We showcase the promise of leveraging WGS in TOPMed for TL genetics in the context of race and ancestry. Future efforts include fine-mapping and co-localization analysis using GTEx and eQTLGen whole blood eQTLs to identify functional variants, with an emphasis on loci showing population differences in signal.

A- A+

PgmNr 34: Leukocyte telomere length and occurrence of chronic health conditions among survivors of childhood cancer: A report from the St. Jude Lifetime Cohort.

Authors:
N. Song [1]; Z. Li [1]; N. Qin [1]; C.R. Howell [1]; C.L. Wilson [1]; J. Easton [2]; H. Mulder [2]; M.N. Edmonson [2]; M.C. Rusch [2]; J. Zhang [2]; M.M. Hudson [1]; Y. Yasui [1]; L.L. Robison [1]; K.K. Ness [1]; Z. Wang [1]

View Session Add to Schedule

Affiliations:
1) Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, TN; 2) Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN

Background: Leukocyte telomere length (LTL) has not been extensively evaluated among childhood cancer survivors, even though it is widely believed that survivors are vulnerable to telomere attrition following exposures to cytotoxic cancer treatments. We aimed to compare LTL and age-dependent LTL attrition rates between survivors and non-cancer controls, and to evaluate associations between LTL and childhood cancer treatment exposures, chronic health conditions (CHCs), and health behaviors among survivors of childhood cancer.
Methods: We included 2,430 survivors and 293 non-cancer community controls of European ancestry, drawn from the participants in St. Jude Lifetime Cohort Study (SJLIFE), a retrospective hospital-based study with prospective clinical follow-up (2007-2016). Common non-neoplastic CHCs (59 types) and subsequent malignant neoplasms (5 types) were clinically assessed and based on follow-up through December 2016. LTL was measured with whole-genome sequencing data. Cumulative doses of chemotherapy and region-specific radiation exposures were abstracted from medical records. Health behaviors including physical activity, diet, smoking, and alcohol intake were assessed from questionnaires.
Results: After adjusting for age at DNA sampling, sex, and genetic risk score based on 9 SNPs known to be associated with telomere length, LTL among survivors was significantly shorter both overall (adjusted mean [AM]=6.12kb; 95% CI=6.06-6.19kb) and across diagnoses than among non-cancer controls (AM=6.72kb; 95% CI=6.54-6.90kb). Among survivors, specific treatment exposures associated with shorter LTL included chest or abdominal irradiation, glucocorticoids, and vincristine chemotherapy. Significant negative associations between LTL and 14 different CHCs (e.g. obesity), and a positive association between LTL and subsequent thyroid cancer occurring out of irradiation field were identified. Health behaviors were associated with LTL among survivors aged 18-35 years (favorable health behaviors [AM=6.55kb; 95%CI=6.39kb-6.71kb] vs. unfavorable health behaviors [AM=6.17kb; 95% CI=5.87kb-6.46kb], p=0.05).
Conclusion: LTL is significantly shorter among survivors of childhood cancer than non-cancer controls, and is associated with CHCs and health behaviors, suggesting this aging biomarker as a potential mechanistic target for future intervention studies designed to prevent or delay onset of CHCs in survivors of childhood cancer.

A- A+

PgmNr 35: The nuclear organization of telomeres as a genomic exploration tool in triple negative breast cancer.

Authors:
O. Samassekou [1]; S. Mai [2]

View Session Add to Schedule

Affiliations:
1) University of Science, Technic and Technology of Bamako, Bamako, Mali; 2) Research Institute in Oncology and Hematology, Winnipeg, Manitoba, Canada

---

In Mali, triple negative breast cancers (TBNC) represent 46% of all breast cancers and more than 90% of TBNC patients have grade III tumors underscoring the high malignancy of this molecular form in the Malian population. Half of TNBC patients respond well to chemotherapy while the other half does not. It is crucial to find biomarkers that can predict aggressive forms at diagnosis, so treatment can be better tailored to patients' needs. Some evidence suggests that high malignancy of TNBC is due to a high level of genomic instability, and one of the earliest predictors of genomic instability in cancer is the alteration of the telomeric nuclear organization (TNO). We ask whether the alteration of nuclear organization of telomeres of TNBC, at diagnosis, can predict genomic instability and aggressive forms of TNBC.
We studied the telomeric nuclear organization in 100 TNBC patients. Telomeres were labeled on a 5 microns-tissue section by using the three-dimensional (3D) fluorescent *in situ* hybridization (FISH), the cell and telomere images and were captured by a 3D fluorescent microscopy, and parameters defining the TNO were assessed by the TeloView™ software. We did statistical analysis by correlating parameters of TNO (number of telomere aggregates, telomere number, the position of telomeres within nuclei, telomere intensities, nuclear volume, and telomere distance respective to the center of nuclei) to clinical outcomes of patients (progression-free survival and overall survival). We found that the number of telomere aggregates, the position of telomere within nuclei, and telomere intensities were the most statistically significant parameters of TNO which predict progression-free survival and overall survival of TNBC patients.
This study has enabled us to find a potential predictive biomarker for TBNC patients. If this finding is confirmed by a large study and validated in subsequent studies, treatment can be tailored for TNBC patients, at diagnosis, for an improvement of their survival. Most importantly, the approaches we have used in this study can be rapidly translated into clinical practice for the benefit of breast cancer patients.

A- A+

PgmNr 36: Noncoding CGG repeat expansions as common causative mutations for three diseases, neuronal intranuclear inclusion disease, oculophryngodistal myopathy, and an overlapping disease.

Authors:
H. Ishiura [1]; S. Shibata [1]; J. Yoshimura [2]; Y. Suzuki [2]; W. Qu [2]; K. Doi [2,3]; M.A. Almansour [1]; J.K. Kikuchi [1]; M. Taira [1,4]; J. Mitsui [1,5]; Y. Takahashi [1,6]; Y. Ichikawa [1,7]; T. Mano [1]; A. Iwata [1]; Y. Harigaya [8]; M.K. Matsukawa [1]; T. Matsukawa [1,5]; M. Tanaka [1,9]; H. Kowa [1,10]; S. Murayama [11]; Y. Shiio [12]; Y. Saito [13]; S.Y. Lim [14]; A.H. Tan [14]; J. Shimizu [1]; J. Goto [1,15]; I. Nishino [16]; T. Toda [1]; S. Morishita [2]; S. Tsuji [1,5,9]

View Session Add to Schedule

Affiliations:
1) Department of Neurology, The University of Tokyo Hospital, Tokyo, Japan; 2) Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan; 3) School of Bioscience and Biotechnology, Tokyo University of Technology, Tokyo, Japan; 4) Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Miyagi, Japan; 5) Department of Molecular Neurology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; 6) Department of Neurology, National Center of Neurology and Psychiatry, Tokyo, Japan; 7) Department of Neurology, Kyorin University, Tokyo, Japan; 8) Department of Neurology, Maebashi Red Cross Hospital, Gunma, Japan; 9) Institute of Medical Genomics, International University of Health and Welfare, Chiba, Japan; 10) Department of Rehabilitation Science, Kobe University Graduate School of Health Sciences, Hyogo, Japan; 11) Department of Neurology and Neuropathology (the Brain Bank for Aging Research), Tokyo Metropolitan Geriatric Hospital and Institution of Gerontology, Tokyo, Japan; 12) Department of Neurology, Tokyo Teishin Hospital, Tokyo, Japan; 13) Department of Pathology and Laboratory Medicine, National Center Hospital, National Center of Neurology and Psychiatry, Tokyo, Japan; 14) Division of Neurology, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia; 15) Department of Neurology, International University of Health and Welfare Mita Hospital, Tokyo, Japan; 16) Department of Neuromuscular Research, National Institute of Neuroscience, National Center of Neurology and Psychiatry, Tokyo, Japan

---

[Background] Neuronal intranuclear inclusion disease (NIID) is a neurodegenerative disease characterized by leukoencephalopathy associated with abnormal signals in diffusion-weighted images and ubiquitin-positive intranuclear inclusions in the central and peripheral nervous systems as well as other organs. The clinical, radiological, and pathological characteristics of NIID are strikingly similar to those of fragile X tremor/ataxia syndrome (FXTAS) caused by expanded CGG repeats in the 5’ untranslated region (UTR) in *FMR1*. We, furthermore, have noticed two other diseases, oculopharyngeal myopathy with leukoencephalopathy (OPML) and oculopharyngodistal myopathy (OPDM), that exhibit overlapping clinical presentations with each other. Inspired by the similarity of the clinical presentations of these diseases and FXTAS, we aimed to identify causative mutations in these diseases focusing on expanded CGG repeats.
[Methods] Repeat expansions were directly investigated from whole genome sequence data using TRhist, which efficiently extracts short reads filled with repeat sequences. The repeat expansions were confirmed by repeat-primed PCR, Southern blot, and long-read sequence analyses.

[Results] TRhist detected short reads filled with CGG repeats in the 5’ UTR in *NBPF19* (*NOTCH2NLC*) from 4 patients with NIID. The expanded CGG repeats in *NBPF19* were further confirmed in 28 families with NIID and absent in 1,000 controls. We similarly identified noncoding CGG repeat expansions in *LOC642361/NUTM2B-AS1* in a family with OPML, which were found in 4 affected patients and absent in 7 unaffected family members or 1,000 controls. Finally, we identified expanded CGG repeats the 5’ UTR of *LRP12* in 13/34 (38.2%) of OPDM patients with rimmed vacuoles (RVs), while only two of the 1,000 controls (0.2%) had the repeat expansions.
[Discussion] On the basis of clinical similarities, we directly searched for CGG repeats and identified noncoding CGG repeat expansions in the three independent genes as the causes of NIID, OPML, and OPDM. Together with our previous finding that noncoding TTTCA and TTTTA repeat expansions in the three genes (*SAMD12*, *TNRC6A*, and *RAPGEF2*) cause the similar disease, benign adult familial myoclonic epilepsy, the results expand our knowledge on the molecular genetics of repeat expansion diseases as highlighted by the concept that expanded noncoding repeats with the same repeat motifs lead to similar clinical presentations, repeat motif-phenotype correlation.

A- A+

PgmNr 37: *De novo EIF2AK1* and *EIF2AK2* variants are associated with developmental delay, movement disorders, cerebellar ataxia, leukoencephalopathy, and neurologic decompensation.

Authors:
D. Mao [1,2,25]; C. Reuter [3,4,25]; M. Ruzhnikov [5,6]; A. Beck [7]; E. Farrow [8,9,10]; L. Emrick [11,12]; J. Rosenfeld [12]; K. Mackenzie [5]; L. Robak [2,12]; M. Wheeler [3,13]; L. Burrage [12]; M. Jain [14]; D. Calame [11]; M. Graf [15]; S. Masters [16,17]; B. Lee [12,18]; I. Thiffault [8,9,19]; P. Agarwal [20, 26]; J. Bernstein [3,21,26]; H. Bellen [2,12,22,23,24,26]; H. Chao [1,2,11,12,26]; Undiagnosed Diseases Network

View Session Add to Schedule

Affiliations:
1) Department of Pediatrics, Baylor College of Medicine, Houston, Tx.; 2) Jan and Dan Duncan Neurological Research Institute, Texas Children’s Hospital, Houston, TX; 3) Stanford Center for Undiagnosed Diseases, Stanford University, Stanford, CA; 4) Stanford Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine, Stanford School of Medicine, Stanford, CA; 5) Department of Neurology and Neurological Sciences, Stanford, CA; 6) Division of Medical Genetics, Department of Pediatrics, Stanford Medicine, Stanford, CA; 7) Department of Pediatrics, Division of Genetic Medicine, University of Washington, Seattle, WA; 8) Center for Pediatric Genomic Medicine, Children's Mercy Hospital, Kansas City, MO; 9) University of Missouri-Kansas City School of Medicine, Kansas City, MO; 10) Department of Pediatrics, Children's Mercy Hospitals, Kansas City, MO; 11) Division of Neurology and Developmental Neuroscience, Department of Pediatrics, BCM, Houston, TX; 12) Department of Molecular and Human Genetics, BCM, Houston, TX; 13) Department of Medicine, Stanford University School of Medicine, Stanford, CA; 14) Department of Bone and Osteogenesis Imperfecta, Kennedy Krieger Institute, Baltimore, MD; 15) CardioVascular Thoracic Institute, Keck Hospital of University of Southern California, Los Angeles, CA; 16) Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia; 17) Department of Medical Biology, University of Melbourne, Parkville, Victoria 3052, Australia; 18) Texas Children’s Hospital, Houston, TX 77030, USA; 19) Department of Pathology and Laboratory Medicine, Children's Mercy Hospitals, Kansas City, MO; 20) Division of Genetics and Genomics, Boston Children's Hospital and Harvard Medical School, Boston, MA; Division of Newborn Medicine, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA; Manton Center for Orphan Disease Research, Bos; 21) Department of Pediatrics, Stanford University School of Medicine, Stanford, CA; 22) 22Program in Developmental Biology, BCM, Houston, TX; 23) Department of Neuroscience, BCM, Houston, TX; 24) Howard Hughes Medical Institute, BCM, Houston, TX; 25) Equal contribution; 26) Corresponding authors

EIF2AK1 and *EIF2AK2* encode members of the Eukaryotic Translation Initiation Factor 2 Alpha Kinase (EIF2AK) family that inhibits protein synthesis in response to different stress conditions, such as ER stress, oxidative stress, heme deficiency, amino acid starvation, and viral infection. EIF2AK2 is also involved in innate immune response and the regulation of signal transduction, apoptosis, cell proliferation, and differentiation. In the gnomAD database, loss of function variants in both *EIF2AK1* (o/e = 0.47) and *EIF2AK2*(o/e = 0.30) are less tolerated than expected by statistical predictions. Despite their key cellular function and restricted variation in population databases, human Mendelian

disorders associated with *EIF2AK1* and *EIF2AK2* have not been reported. Here, we describe the identification of seven unrelated individuals through the Undiagnosed Diseases Network and GeneMatcher with heterozygous *de novo* missense variants in *EIF2AK1* or *EIF2AK2* that map to known functional domains of the respective proteins. Recurrent features seen in these individuals include developmental delays (7/7), hypotonia (6/7) and hypertonia (5/7), involuntary movements (2/7), cerebellar ataxia (4/7), and variable cognitive impairments (5/7). Notably, individuals with *EIF2AK2* variants exhibit sensitivity to febrile inflammatory events that appear to provoke neurological deterioration and white matter alterations. EIF2AKs mediated the phosphorylation of EIF2S1(eukaryotic translation initiation factor 2 subunit alpha, eIF2α). Intriguingly phosphorylation of EIF2S1 inhibits EIF2Bs, which are associated with vanishing white matter (VWM) disease. Our findings in mammalian cell lines and patient-derived fibroblasts indicate that loss-of-function variants in *EIF2AK1* and *EIF2AK2* cause a genetic neurodevelopmental syndrome that may share common pathogenic mechanisms with VWM disease.

A- A+

PgmNr 38: *De novo* truncating *TRIM8* mutations cause a novel pediatric neuro-renal syndrome and abrogate protein localization to nuclear bodies.

Authors:
K. Khan [1]; P.L. Weng [2]; A.J. Majmundar [3]; T.Y. Lim [4]; S. Shril [3]; J.A. Martinez-Agosto [5]; N. Mann [3]; Y. Jin [4]; V. Aggarwal [6]; A. Onuchic-Whitford [3]; F. Buerger [3]; J. Musgrove [1]; B.B. Beck [7,8]; K.M. Riedhammer [9]; M.R. Benz [10]; J. Hoefele [9]; H.L. Rehm [11]; D.G. MacArthur [11]; S. Mane [12]; V. D'Agati [13]; F. Hildebrandt [3]; S. Sanna-Cherchi [4]; E.E. Davis [1]

View Session Add to Schedule

Affiliations:
1) Center for Human Disease Modeling, Duke University, Durham, North Carolina.; 2) Div. of Pediatric Nephrology, UCLA, Los Angeles.; 3) Dept. of Medicine, Boston Children's Hospital, Boston, Massachusetts.; 4) Div. of Nephrology, Columbia University, New York, New York.; 5) Department of Pediatrics, Division of Medical Genetics, University of CA, Los Angeles.; 6) Institute of Genomic Medicine, Columbia University, New York.; 7) Institute of Human Genetics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany.; 8) Center for Molecular Medicine, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany.; 9) Institute of Human Genetics, Klinikum rechts der Isar, Technical University Munich, Germany.; 10) Kindernephrologie in Dachau, Germany.; 11) Broad Center for Mendelian Genetics, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts.; 12) Dept. of Genetics, Yale University School of Medicine, New Haven, Connecticut.; 13) Dept. of Pathology, Columbia University, New York, New York.

Genetic forms of pediatric focal segmental glomerulosclerosis (FSGS) are predominantly caused by recessive mutations, but the contribution of dominant *de novo* mutations (DNMs) to this trait is unknown. Here, we performed genetic studies using whole exome sequencing (WES) in 43,146 individuals. First, we used trio-based WES in a 4-year-old individual with epilepsy and FSGS and the unaffected parents and identified a novel DNM in *TRIM8* (p.Q459*). To support the potential link between *TRIM8* and FSGS, we conducted WES in unrelated cases at two independent research centers. We identified nonsense mutations in *TRIM8* in 6 additional individuals out of 2,051 children with FSGS, a significant enrichment compared to 0/35,885 controls ($P=2.5x10^{-8}$). Parental data were available for 4/6 cases and in all instances the mutations were *de novo* ($P=2.6x10^{-10}$), establishing *TRIM8* as a novel FSGS gene. Review of clinical data showed that all 6 FSGS cases also had epilepsy, indicating that *TRIM8* DNMs define a neuro-renal syndrome. Accordingly, we identified additional *TRIM8* truncating mutations in 2 of 5,209 epilepsy patients. *TRIM8* encodes a 551 amino acid E3 ubiquitin ligase, and notably, all the mutations clustered between amino acid positions 410-460 in the last exon, and are predicted to escape nonsense mediated decay. Further, *TRIM8* is intolerant to loss of function (pLI=0.99) as evidenced by few gnomAD subjects with heterozygous truncating variants, all of which reside outside of the constrained region impacted in affected individuals. We performed *in vitro* studies using immunohistochemistry and observed TRIM8 localization in podocytes. In immortalized podocytes transfected with wild type TRIM8, we noted targeting to nuclear bodies, while constructs harboring patient-specific mutations (p.Q411* and p.Q459*) localized diffusely to the nucleoplasm. Zebrafish *trim8* models support further the dominant toxic variant effect, arguing

against a haploinsufficiency mechanism in affected humans. Together, our data show that dominant *TRIM8* DNMs cause a novel pediatric neuro-renal syndrome likely due to aberrant cellular localization, implicating nuclear bodies in FSGS pathogenesis.

A- A+

PgmNr 39: Hypomorphic variants in the deubiquitylase OTUD5 cause multiple congenital defects through altered chromatin dynamics.

Authors:
D.B. Beck [1]; A.B. Basar [2]; H. Oda [1]; D.T. Uehara [6]; J. Inazawa [6]; E. Macnamara [1]; P. D'Souza [1]; J. Bodurtha [3]; W. Mu [3]; K. Baranano [3]; T. Kosho [8]; M. Kempers [7]; M. Walkiewicz [4]; R. Wang [5]; C.J. Tifft [1]; I. Aksentijevich [1]; A. Werner [2]; D. Kastner [1]

View Session Add to Schedule

Affiliations:
1) National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland.; 2) National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda; 3) Institute of Genomic Medicine, Johns Hopkins Hospital, Baltimore, MD; 4) National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD; 5) Children’s Hospital of Orange County, University of California Irvine School of Medicine, Orange, CA; 6) Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan; 7) Radboud University, Nijmegen, Netherlands; 8) Department of Medical Genetics, Shinshu University School of Medicine

Embryonic development occurs through commitment of pluripotent stem cells to differentiation programs that require highly coordinated changes in chromatin architecture. While many factors controlling chromatin dynamics are known, mechanisms how different chromatin regulators are orchestrated during development are not well understood. Here, we describe a novel multiple congenital anomaly disorder caused by hypomorphic, hemizygous variants in *OTUD5,* which encodes for a Lys48/Lys63-chain-specific deubiquitylation enzyme (DUBA). We identified seven males from five families each with distinct, novel variants in *OTUD5,* all with overlapping characteristics*.* Affected individuals have a spectrum of clinical manifestations including structural brain malformations, congenital heart disease, ambiguous genitalia, post-axial polydactyly, and craniofacial findings with phenotypic overlap with Coffin-Siris Syndrome (CSS), and Cornelia de Lange Syndrome (CdLS). Studying these mutations *in vitro* and *in vivo* using mouse and human pluripotent stem cell models, we uncover a novel regulatory circuit that coordinates chromatin remodeling pathways during early differentiation. We show that OTUD5’s Lys48-linkage specific deubiquitylation activity is essential for murine and human development and, if reduced, leads to aberrant differentiation, in particular affecting neuroectodermal lineages. Intriguingly, OTUD5 interacts with and regulates the stability of several key chromatin regulators during differentiation, many of which are known to control neural fates and are linked to CSS and CdLS. Consistent with these findings, loss of OTUD5 causes aberrant chromatin architecture during specific stages of lineage commitment. Thus, we conclude that OTUD5 orchestrates changes in chromatin dynamics during early differentiation events to allow for robust changes in gene expression networks required for proper human development.

A- A+

PgmNr 2609: Outcomes for reanalysis of exome sequencing data in a large cohort.

Authors:
M. Guillen Sacoto; A. Begtrup; R. Willaert; A. Crunk; E. Heise; L. Rhodes; C. Kucera; L. Havens; J. Juusola

View Session Add to Schedule

Affiliation: GeneDx, Gaithersburg, MD

The American College of Medical Genetics and Genomics (ACMG) encourages clinical laboratories to reevaluate Exome Sequencing (ES) cases and variants (Deignan et al., 2019). This recommendation is supported by studies showing that reanalysis of previously generated ES data increases the positive yield by 10-16% in small cohorts (Hiatt et al., 2018; Shashi et al. 2019).

Between September 2013 and April 2019 we reanalyzed 4255 ES cases. Our cases include a broad range of phenotypes with neurological disorders (48.7%) and multiple congenital anomalies (MCA) (24.9%) being the most common. Of those 4255 cases, 377 (8.9%) were initially reported as positive, 1366 (32.1%) negative, and 2512 (59.0%) as of uncertain clinical significance. 2819 (66.3%) cases were trios and the average time between analyses was 21 months (SD 11.1 months).

We were able to make a new definitive molecular diagnosis in 419/4255 (9.9%) cases. We reported a candidate gene in 610 (14.3%) and no additional variants in 2586 (60.8%) cases. The majority of new positive cases (304/419, 72.6%) resulted from increased knowledge about the gene or variant from publications or internal observations; 162 (38.7%) cases involved genes that had not been associated with human disease at the time of the initial analysis. Fifty-six (13.4%) new diagnosis were identified by pipeline improvements.

Among the 377 cases initially reported as positive, we downgraded a pathogenic or likely pathogenic variant in 43 (11.4%), and had no additional findings to report in 264 (70.0%) cases. Of the 91 cases that had an additional variant reported, 46.2% had new phenotypic information. In 13 (3.4%) cases the reanalyzed results were consistent with a dual diagnosis.

The diagnostic yield of ES increased by 10% with reanalysis of previously generated data in our diverse cohort. The likelihood of identifying a definitive molecular diagnosis, is lower for already positive cases, but since ES reporting is phenotype-driven, reanalysis should be considered if there is new clinical information. Since 2012, 207/3052 (6.8%) novel candidate genes reported by our laboratory have been published as disease-causing genes, impacting 2,696/34,000 (7.9%) total ES cases.

A- A+

PgmNr 197: High-throughput RNA splicing profile increases detection of clinically-actionable variants while reducing inconclusive results in patients with hereditary cancer predisposition.

Authors:
T. Landrith; B. Li; A. Cass; B.R. Conner; S. Wu; H. Vuong; S. Charpentier; J. Burdette; H. LaDuca; T. Pesaran; J. Rae-Radecki Crandall; H. Lu; B. Tippin-Davis; A. Elliot; R. Karam

View Session Add to Schedule

Affiliation: Ambry Genetics, Aliso Viejo, CA

The diagnostic yield of cancer genetic testing has improved over time as new technologies are established. Despite these improvements, a substantial proportion of patients with suspected hereditary cancer syndromes receive either inconclusive results or remain without a molecular diagnosis. Inconclusive results, or variants of uncertain significance (VUS), are a barrier to diagnosis and can often lead to challenges in the management of patients with hereditary cancer predisposition. Furthermore, genetic testing may not detect clinically actionable (pathogenic/likely pathogenic) variants due to intrinsic limitations of DNA-based assays. RNA data can aid in the resolution of these cases. However, this determination is complicated by multiple isoforms and alternative splicing events that exist within the normal range of biological variation. To address this, we obtained blood samples from 345 healthy donors and generated a control reference dataset using an RNA massively parallel sequencing assay and analysis pipeline. This allowed us to characterize normal alternative splicing across clinically significant hereditary cancer predisposition genes expressed in blood, many of which have not been previously characterized with this approach. Using our control reference dataset, we were able to determine whether the Percent Splicing Index (PSI) for a splicing event associated with a given DNA variant deviates significantly from normal splicing. Here, we present cases studies in which this approach yielded RNA splicing data that was used as evidence to detect clinically actionable variants. Our results demonstrate the utility of this approach to increase the positive yield of genetic testing for hereditary cancer predisposition in the clinical diagnostic setting.

A- A+

PgmNr 2536: Improving diagnostic yield in clinical and acute care genomics: Rapid PCR studies of splicing variants in tissue-specific genes using blood or fibroblast mRNA.

Authors:
A. Bournazos [1,2]; L.G. Riley [2,3]; L.S. Akesson [4,5,6]; N.L. Baker [4,5]; K. Boggs [7,8]; N. Brown [4,5]; M. Buckley [9]; M.R. Davis [10]; M. Edwards [11]; L.J. Ewans [12,13,14]; A. Fennell [5,6]; M. Field [15]; M.L. Freckmann [16]; H. Goel [15]; S. Goh [2]; L. Goodwin [2,17]; K. Kumar [2,13,18]; S. Lunke [4,7,19]; A. Ma [2,17]; H. McCarthy [20,21]; M.P. Menezes [2,22]; D. Mowat [23,24]; S.A. Sandaradura [1,2,25]; Z. Stark [4,5,7]; C. Sue [26]; T.Y. Tan [4,5]; E. Tantsis [22]; M. Tchan [2,27]; S.M. White [4,5]; M. Wilson [2,25]; D.C. Wright [2,28]; K. Jones [1,2]; B. Bennetts [2,29]; S.T. Cooper [1,2,30]

View Session Add to Schedule

Affiliations:
1) Kids Neuroscience Centre, Kids Research, The Children's Hospital at Westmead, Sydney, Australia; 2) Sydney Medical School, University of Sydney, Sydney, Australia; 3) Rare Diseases Functional Genomics, Kids Research, The Children’s Hospital at Westmead and The Children’s Medical Research Institute, Sydney, Australia; 4) Victorian Clinical Genetics Services, Murdoch Children’s Research Institute, Melbourne, Australia; 5) Department of Paediatrics, University of Melbourne, Melbourne, Australia; 6) Monash Genetics, Monash Health, Melbourne, Australia; 7) Australian Genomics Health Alliance, Melbourne, Australia; 8) Sydney Children’s Hospital Network, The Children's Hospital at Westmead and Sydney Children's Hospital, Randwick, Sydney, Australia; 9) NSW Health Pathology East Laboratory, Prince of Wales Private Hospital, Randwick, New South Wales, Australia; 10) Neurogenetics Unit, Department of Diagnostic Genomics, PathWest Laboratory Medicine, QEII Medical Centre, Nedlands, Australia; 11) School of Medicine, Western Sydney University, Australia; 12) St Vincent’s Clinical School, Faculty of Medicine, University of New South Wales, Sydney, Australia; 13) Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, Australia; 14) Department of Medical Genomics, Royal Prince Alfred Hospital, Sydney, Australia; 15) Genetics of Learning Disability Service, Hunter Genetics, Waratah, Australia; 16) Department of Clinical Genetics, Royal North Shore Hospital, Sydney, Australia; 17) Department of Clinical Genetics, Nepean Hospital and Western Sydney Genetics Program, The Children's Hospital at Westmead, Sydney, Australia; 18) Molecular Medicine Laboratory, Concord Hospital, Sydney, Australia; 19) Department of Clinical Pathology, University of Melbourne, Melbourne, Australia; 20) Centre for Kidney Research, The Children’s Hospital at Westmead, Sydney, Australia; 21) Department of Renal Medicine, Sydney Children’s Hospital, Randwick, Sydney, Australia; 22) TY Nelson Department of Neurology and Neurosurgery, The Children's Hospital at Westmead, Sydney, Australia; 23) Centre for Clinical Genetics, Sydney Children’s Hospital, Randwick, Sydney, Australia; 24) School of Women’s and Children’s Health, The University of New South Wales, Sydney, Australia; 25) Department of Clinical Genetics, Children’s Hospital at Westmead, Sydney, Australia; 26) Department of Neurogenetics, Kolling Institute, Royal North Shore Hospital, Sydney, Australia; 27) Department of Genetic Medicine, Westmead Hospital, Sydney, Australia; 28) Sydney Genome Diagnostics, The Children’s Hospital at Westmead, Sydney, Australia; 29) Department of Molecular Genetics, The Children’s Hospital at Westmead, Sydney, Australia; 30) The Children’s Medical Research Institute, Sydney, Australia

---

Background

Variants that may affect pre-mRNA splicing are often classified as variants of uncertain significance

(VUS), which are uninterpretable in terms of provision of a genetic diagnosis. We have established a Rapid Splicing Diagnostics program to provide experimental evidence to enable accurate classification of putative splice variants, in accordance with American College of Medical Genetics (ACMG) guidelines. Due to the tissue specific expression of many genes, biopsies from relevant manifesting vital organs are limited. Here we present 20/21 cases where RNA derived from blood and skin fibroblasts was used to determine pathogenicity of splice variants in the absence of an available biopsy.

Methods

Reverse transcriptase polymerase chain reaction (RT-PCR) and Sanger sequencing was performed for 40 individuals from 21 families with a putative splicing VUS identified in gene relevant to their clinical presentation. RNA was derived from blood (18/21), skin fibroblasts (7/21), and/or an available biopsy (2/21) from 6 trios, 5 duos, 10 singleton cases.

Results

Families were affected with a range of genetic conditions (developmental, cardiac, metabolic, neurological) with 2/21 fetal death, 4/21 severe congenital, 10/21 early childhood and 5/21 teenage-adult onset. Diagnostically informative results were obtained for 20/21 cases using RNA from blood or skin fibroblasts. Splicing analyses were performed for at least 2 carriers of the variant in 11/21 of families, and two or more biospecimens (blood, skin fibroblasts, autopsy specimen) for 6/21 cases.

Conclusions

RT-PCR and Sanger sequencing provided experimental evidence to support re/classification of splice variants following ACMG guidelines. Furthermore, we successfully implemented rapid functional studies of mRNA splicing within clinically relevant timeframes (reported within 10–14 days) for two neonates in acute care, enabling variant re-classification and directly informing clinical management. In conclusion, RNA from blood and skin fibroblasts can provide experimental evidence for abnormal splicing of genes with highly restricted tissue-specific expression (7/21 cases < 1 transcript per million). Importantly, an identical pattern of mis-splicing was observed for each variant among biological replicates (different individuals and different specimens; including the manifesting tissue), showing blood and fibroblast RNA reproducibly informs abnormal splicing events elicited by the splicing variant across tissues.

A- A+

PgmNr 225: Reclassification of splicing VUS in neurological disease genes via RNA-seq.

Authors:
S. Ichikawa [1]; B.R. Conner [2]; S. Wu [3]; R. Karam [2]

View Session Add to Schedule

Affiliations:
1) Department of Clinical Diagnostics, Ambry Genetics, Aliso Viejo, CA; 2) Department of Research and Development, Ambry Genetics, Aliso Viejo, CA; 3) Department of Bioinformatics, Ambry Genetics, Aliso Viejo, CA

Clinical genetic testing for neurological disorders is becoming widely available. However, its clinical utility is diminished by a large number of variants of unknown significance (VUS) detected in patients. In particular, the interpretation of splicing VUS poses a significant challenge due to their unknown effects on splicing. In this study, we sought to determine whether RNA-seq analysis can effectively reduce the number of splicing VUS in genes associated with neurological disorders. VUS that might affect splicing were identified in patients who ordered neurology genetic testing (single gene, multi-gene panel, or exome). VUS detected in genes expressed in bone marrow were selected for RNA-seq analysis. Blood was collected from the patients and healthy controls. RNA extracted from blood was analyzed using massively-parallel RNA-seq of cloned RT-PCR products (CloneSeq). Based on splicing events detected in RNA-seq, 86% (19/22) of the variants changed classification from VUS: seven were reclassified to pathogenic/likely pathogenic variants, while twelve were likely benign variants. Those variants that became clinically actionable included: four alterations affecting guanine at the last nucleotide of an exon (*ANDP* c.201G>C, *ANKRD11* c.226G>A, *NF1* c.586G>A, and *TSC1* c.2041G>A), two small deletions in introns (*FMR1* c.104+3_104+6delAAGT and *WDR45* c.976+5_976+10delGTGGGA), and one single nucleotide substitution in an intron (*ATRX* c.4957-3A>G). Of the three variants that remained VUS after RNA-seq analysis, two were missense alterations that had no splicing impact, but may still affect protein function. The remaining VUS was the only variant that had ambiguous RNA evidence. In summary, RNA-seq analysis provided useful evidence for all but one VUS and resulted in reclassification of 86% of the variants that would have remained VUS otherwise. Although RNA-based analysis may be limited to genes expressed in the blood or other readily obtainable tissues, our data indicates that RNA-seq analysis can help clarify pathogenicity of many splicing VUS identified in genes associated with neurological disorders.

A- A+

PgmNr 40: Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders.

Authors:
M.T. Oetjens [1]; M.A. Kelly [1]; A.C. Strum [1]; R.G.C. Regeneron Genetics Center [2]; C.L. Martin [1]; D.H. Ledbetter [1]

View Session Add to Schedule

Affiliations:
1) Geisinger Health System, Danville, PA; 2) Regeneron Genetics Center, Tarrytown, NY

Rare genetic disorders (RGDs) often exhibit significant variation in clinical severity among affected individuals. However, the genetic and environmental factors that contribute to variable expressivity are not well understood. Recently, the aggregate effect of common variation, quantified as polygenic scores (PGSs), has emerged as an effective tool for predictions of disease risk and trait variation in the general population. To determine if common variants contribute to variable expressivity of RGDs, we measured the effect of PGSs in RGDs affecting three quantitative phenotypes: low-density lipoprotein (LDL-C) levels, body-mass index (BMI), and height. From 92,455 patients in the DiscovEHR cohort, an unselected health system-based population, we identified rare pathogenic variants underlying 11 RGDs that affect these three phenotypes, including familial hypercholesterolemia (FH; *LDLR* and *APOB*), familial hypobetalipoproteinemia (FHBL; *PCSK9* and *APOB*) for LDL-C, 16p11.2 deletions and duplications, melanocortin 4 receptor deficiency (*MC4R*) for BMI, and four sex-chromosome aneuploidies (47,XXX; 47,XXY; 47,XYY; 45,X) for height. We first characterized the strength of these rare variants and showed each has a large effect size (0.5 - 2.6 SD) on the affected trait. Next, we developed PGSs for the three quantitative phenotypes ($\text{PGS}_{\text{LDL-C}}$, $\text{PGS}_{\text{BMI,}}$ and $\text{PGS}_{\text{HEIGHT}}$). PGSs were strong predictors of phenotypic variance in the general population (R^2 = 8 - 21%), and in some cases we observed an equivalence between the effect size of an extreme PGS and the rare variant. The effect size of highly penetrant rare variants in *LDLR* ($p = 1.83 \times 10^{-28}$; n=146) and *APOB* ($p = 1.59 \times 10^{-7}$; n=87) are significantly greater than a $\text{PGS}_{\text{LDL-C}}$ in the 100th percentile. However, the BMI of individuals with rare pathogenic *MC4R* variants is approximately equal to variant-negative individuals in the ~99th percentile of PGS_{BMI}. Our examination of PGSs within RGDs revealed a polygenic component to variable expressivity in most RGD/trait pairs tested, including FH caused by pathogenic *LDLR* variants ($p = 7.69 \times 10^{-3}$), obesity caused by pathogenic *MC4R* variants ($p = 3.71 \times 10^{-3}$, n=58), and tall stature caused by 47,XXY ($p = 3.87 \times 10^{-3}$, n=44) and 47,XXX ($p = 1.33 \times 10^{-3}$, n=42). These results demonstrate that common, polygenic factors often contribute in an additive fashion to variable expressivity in RGDs and PGSs may a useful metric for stratifying affected individuals by clinical severity.

A- A+

PgmNr 41: Aberrant post-translational modifications (PTM) characterize the hepatic proteome of methylmalonic acidemia (MMA).

Authors:
P. Head [1]; I. Manoli [1]; Y. Chen [2]; M. Gucek [2]; C. Venditti [1]

View Session Add to Schedule

Affiliations:
1) National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland.; 2) National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland.

Organic acidemias (OAs), such as methylmalonic acidemia (MMA), are a group of inborn errors of metabolism that typically arise from defects in the catabolism of amino- and fatty acids. OAs are difficult to treat and have multisystemic manifestations, leading to increased morbidity and mortality. Accretion of acyl-CoA species is postulated to cause intracellular toxicity. Here, we explore an alternative pathophysiological consequence of impaired acyl-CoA metabolism: the accumulation of aberrant posttranslational modifications (PTMs) that modify enzymes in critical intracellular pathways, especially during periods of increased stress. Using a mouse model that recapitulates the hepatic mitochondriopathy of MMA ($Mut^{-/-};Tg^{INS\text{-}MCK\text{-}Mut}$), we surveyed PTMs in hepatic extracts with propionyl- and malonyl-lysine antibodies. We discovered widespread hyper-acylation in the MMA mice compared to controls, but not in animals with Acsf3 deficiency, a disorder of acyl-CoA synthesis. Next, we prepared anti-PTM antibody columns, purified hepatic extracts from MMA and control mice, and performed mass spectrometry to characterize the PTM proteome. Excessive acylation of enzymes involved in glutathione, urea, arginine, lysine, tryptophan, valine, isoleucine, methionine, threonine, and fatty acid metabolism were detected in the MMA mice but not controls, and further validated via immunoprecipitation analysis and Western blotting. In parallel, we generated, via nonenzymatic acylation reactions, PTM-modified BSA targets for *in vitro* analyses. We purified, then assayed, SIRT1-7 deacylase activity using BSA-PTM standards to identify the SIRT(s) that most efficiently remove MMA related PTMs. Because PTMs usually inhibit enzyme function, our observations suggest that hyperacylation of key enzymes in pathways known to be dysregulated in MMA likely contributes to altered metabolism and identifies a new set of targets for therapeutic intervention.

A- A+

PgmNr 42: Targeting transforming growth factor-β (TGFβ) signaling for osteogenesis imperfecta treatment.

Authors:
I. Song [1]; S. Nagamani [1]; D. Nguyen [1]; I. Grafe [1]; E. Munivez [1]; M. Jiang [1]; P. Esposito [2]; J. Goodwin [2]; E. Strudthoff [2]; S. McGuire [2]; V. Shenava [3]; S. Rosenfeld [3]; B. Lee [1]; Brittle Bone Disorders Consortium

View Session Add to Schedule

Affiliations:
1) Dept. of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA; 2) Orthopaedic Surgery, University of Nebraska Medical Center, Omaha, NE, USA; 3) Dept. of Orthopedic Surgery, Texas Children's Hospital, Baylor College of Medicine, Houston, TX, USA

Osteogenesis imperfecta (OI) is a common genetic disorder characterized by increased bone fragility. Currently, there is no FDA-approved OI specific treatment, though bisphosphonates are used as standard of care. In search of a mechanism-specific approach, we previously showed that excessive TGFβ signaling is a common pathogenic mechanism in both dominant (*Col1a2*$^{+/G610C}$) and recessive (*Crtap*$^{-/-}$) OI mouse models. Here, we tested whether this signaling defect is present in humans with OI and whether they could be safely treated with a pan-anti-TGFβ therapy (fresolimumab). We performed histology, Western blot (WB), immunohistochemistry, and RNASeq on human bone from 9 OI type III (OI-III) patients and 4 controls. OI-III bone showed increased osteocyte density, increased phospho-Smad2 staining, and increased phospho-Smad2/total-Smad2 on WB, together validating the preclinical finding of increased TGFβ activity in OI bone. Principle component analysis and hierarchical clustering of RNASeq data showed distinct OI-III vs. control gene expression patterns. Gene Ontology analysis showed molecular signatures of high bone turnover, reduced trabecula formation, and reduced bone maturation. Interestingly, regulation of SMAD phosphorylation was the most significantly up-regulated molecular event, while Ingenuity Pathway Analysis predicted TGFβ activation as the most significant upstream regulator. The RNASeq data was validated by NanoString gene expression analysis. Finally, in the initial dosing stage of a phase I trial, we tested the safety of fresolimumab (1 mg/kg) in 4 patients with OI. After a single infusion, patients were followed for 6 months. The treatment was well tolerated with no significant adverse events. In 2 subjects (OI type IV, *COL1A2* mutations), areal bone mineral density at the spine increased by 6 and 9%; one subject (OI type VIII, *LEPRE1* mutation) showed no change and one subject (OI type III, *COL1A1* mutation) showed a decrease from baseline. The magnitude of increase in responders is similar to what has been observed with 18 months of anabolic therapy (teriparatide) in studies of postmenopausal osteoporosis. Furthermore, the lack of response in more severe OI may reflect a dose requirement related to magnitude of TGFβ alteration; this will be tested in the next phase of dose escalation (4 mg/kg). Here, our data confirms that excessive TGFβ signaling is a mechanism of OI pathogenesis and anti-TGFβ therapy should be further investigated.

A- A+

PgmNr 43: Nonhuman primate models of rare diseases support the development of precision medicine, pharmacogenomics, and gene therapy approaches.

Authors:
S.M. Peterson [1]; B.N. Bimber [1]; L.M. Colgin [2]; A.L. Johnson [2]; A.D. Lewis [2]; B. Ferguson [1]

View Session Add to Schedule

Affiliations:
1) Genetics Division, Oregon National Primate Research Center, Oregon Health & Sciences University, Beaverton, OR 97006; 2) Division of Comparative Medicine, Oregon National Primate Research Center, Oregon Health & Sciences University, Beaverton, OR 97006

Nonhuman primates provide unparalleled opportunities to advance the development and testing of gene therapy approaches for the treatment of rare diseases. We have identified numerous natural genetic models of human disease through whole genome sequence analysis of pedigreed rhesus macaques housed at the Oregon National Primate Research Center. To date, we have identified more than 30 million unique macaque sequence variants, including over 24,000 alleles that are either identical to SNVs associated with human disease (ClinVar) or predicted to be damaging to protein structure and likely to be pathogenic (PolyPhen, SnpEff, CADD). The predicted functional variants map to over 2,600 OMIM-linked traits, implicating risk for vision loss, neurodegenerative disease, developmental disorders, metabolic disorders, and respiratory disease. Complementing the genomic data are extensive electronic health records (EHR) summarizing clinical and behavioral data including diagnostic imaging, histology, and pathology reports collected on each subject. Combining both genomic and phenotypic data sets, we are rapidly uncovering a broad array of spontaneous primate models of rare disease, including Bardet-Biedl syndrome (*BBS7*), neuronal ceroid lipofuscinosis (*CLN7*), leukodystrophy (*CLCN2*), Pelizaeus-Merzbacher disease (*PLP1*), dyshormonogenetic goiter (*TG*), and epidermolysis bullosa (*KRT5*). All genetic variants, extensively annotated with predicted functional impact, associated gene information, population frequencies, species conservation, and relevant OMIM links are publically-accessible through the macaque Genotype And Phenotype database (mGAP; https://mgap.ohsu.edu/). As appropriate, *in vitro* fertilization and zygotic genotyping are being used to selectively breed unique genetic models for collaborators pursuing critical disease research and treatment goals. These preclinical models provide powerful new opportunities for the study of rare disease biology, biomarker discovery, and the development of precision medicine, pharmacogenomic or gene therapy approaches. Funded by NIH R24OD021324 and P51OD011092.

A- A+

PgmNr 44: An interventional clinical trial to evaluate the role of elamipretide in individuals with Barth syndrome.

Authors:
H.J. Vernon [1]; W.R. Thompson [2]; R. Manuel [1]; A. Aiudi [3]; J.J. Jones [3]; J. Carr [3]; B. Hornby [4]

View Session Add to Schedule

Affiliations:
1) Department of Genetic Medicine, Johns Hopkins Hosp, Baltimore, Maryland.; 2) Department of Pediatric Cardiology, Taussig Heart Center, Johns Hopkins University School of Medicine, Baltimore, Maryland; 3) Stealth BioTherapeutics, Inc Newton MA; 4) Department of Physical Therapy, Kennedy Krieger, Baltimore MD

Barth Syndrome (BTHS) is an X-linked disorder caused by defects in TAZ, a mitochondrial transacylase involved in the final remodeling step of cardiolipin (CL). This transacylation defect leads to mitochondrial dysfunction due to an accumulation of the unremodeled, immature form of CL, monolysocardiolipin (MLCL) and a reduction in the remodeled, mature form of CL, tetralinoleoyl-cardiolipin (L4-CL) on the inner mitochondrial membrane. Clinical features of BTHS include cardiomyopathy, skeletal myopathy, neutropenia and growth abnormalities. There are no specific targeted therapies for BTHS, and high morbidity and mortality. Previously, we identified that impairment on functional exercise performance (measured by 6-Minute Walk Test [6MWT]) is significantly associated with the MLCL:L4-CL ratio in affected individuals, thus providing a quantitative clinical endpoint and an associated biomarker for evaluation of ongoing clinical status, and for testing novel therapeutic interventions. Here we describe a placebo-controlled, interventional clinical trial in patients with BTHS to investigate the role of elamipretide, a tetrapeptide known to stabilize CL and improve mitochondrial energy production. The study design consisted of a Part 1 placebo-controlled crossover trial, followed by a Part 2 open label extension (OLE). Endpoints included functional exercise performance, quality-of-life (QoL) questionnaires, and the MLCL:CL ratio. At the end of Part 1, statistical significance was not achieved on the primary endpoints. However, a pre-specified analysis of subjects with lower MLCL:L4-CL ratios showed improvements. Ten patients continued into the ongoing OLE. With longer elamipretide therapy (week 12 OLE) data demonstrate consistent improvements across functional and QoL assessments as well as a mean improvement in the MLCL:L4-CL ratio (-7.4 [-38.7%, P=0.03]), suggesting a physiologic basis to the observed clinical improvements. In summary, we determined that continued elamipretide therapy is associated with improvements in clinical status and the CL content in BTHS patients. Moreover, this work illustrates a pathway towards development of treatments for ultra-rare genetic diseases: 1.) thorough collection of longitudinal and cross-sectional clinical natural history data, 2.) identification of clinical endpoints and correlation to biomarkers, and 3.) selection of functionally relevant therapeutic molecules.

A- A+

PgmNr 45: Moving human genetics into the mouse: Full human gene-replacement models.

Authors:
M. Koob; K. Karanjeet; K. Benzow

View Session Add to Schedule

Affiliation: Institute for Translational Neuroscience, University of Minnesota, Minneapolis, MN.

Although geneticists have now identified many of the sequence variants that underlie a wide array of human diseases, most of the animal models we have made in light of these findings have failed to capture more than a fraction of the molecular impacts of these pathogenic variants. In the first public presentation of this work, we report that we have developed Gene Replacement (GR) technology that allows us to more fully model the genetics of human disease in mice by replacing mouse genes with their full human orthologs. We used this technology to replace the full mouse *Microtubule-associated protein tau* (*Mapt*) genomic coding and regulatory region (156,547bp) with the full human *MAPT* genomic sequence (190,081bp). We have confirmed that mice homozygous for this *MAPT*-GR allele express human tau at endogenous levels, and that all expected splice variants are found in the appropriate tissues and in ratios expected for the fully functional human *MAPT* gene. The genomes of all subsequent models generated in this *MAPT*-GR series of mouse lines will precisely match our first *MAPT*-GR line except for those *MAPT* nucleotide differences that we specifically introduce. These matched sets of animal models will allow the research community to evaluate the molecular impact of pathogenic mutations within the context of the human genomic sequence in which they occur in patients, and these mouse lines will contain all potential human therapeutic targets of the *MAPT* gene. Furthermore, by fully replacing the endogenous *Mapt* gene with its human counterpart, a) the *MAPT* gene in these lines is situated in the only location within the mouse genome that has specifically evolved to express the tau gene, and b) the *MAPT*-GR lines do not express any endogenous mouse tau gene products that could confound analyses of these mice. We are working to generate similar lines in which other human genes involved in the etiology of Alzheimer’s Disease (AD) replace their mouse homologs, and will discuss our progress in characterizing our first lines with a 356kb *Amyloid Precursor Protein* (APP)-GR allele. Finally, to demonstrate the utility of these approaches to diseases other than AD, we will describe matched sets of partial Gene Replacements (pGR) lines with either 31kb *ATXN1*-pGR alleles (wt + 5 matched SpinoCerebellar Ataxia type 1 variants) or 192kb *TCF4*-pGR alleles (wt + 2 matched Fuchs Endothelial Corneal Dystrophy CTG expansion mutations).

A- A+

PgmNr 46: Gene mapping in breast cancer extended pedigrees using PAM50 tumor dimensions.

Authors:
J. Gardner [1]; M. Madsen [1]; S. Knight [1]; M. Cessna [1]; R. Factor [1]; C. Sweeney [1]; B. Caan [2]; L. Kushi [2]; P. Bernard [1]; N. Camp [1]

View Session Add to Schedule

Affiliations:
1) University of Utah, Salt Lake City, UT; 2) Kaiser Permanente, Oakland, CA

Breast cancer is a heterogeneous disease whose etiology includes inherited risk. Large pedigrees contributed to landmark discoveries of high-risk breast cancer genes. However, subsequent progress with this design has been hindered by disease complexity. We hypothesized that heritable tumor molecular phenotypes would reduce germline heterogeneity and revitalize the large pedigree approach for gene mapping. We previously used principal component (PC) analysis to derive five orthogonal tumor dimensions from PAM50 gene expression (PC1—PC5). We further illustrated that extreme PC3 tumors were enriched in pedigrees and used shared genomic segment (SGS) analysis to identify a 0.5 Mb region at 12q15 containing *CNOT2* (subunit of CCR4-NOT complex, a global transcriptional regulator for cell growth and survival). The region was inherited through 32 meioses to 8 cases ($p=2.610^{-8}$, Madsen et al 2018). Here, we perform a gene mapping tumor dimension study in 6 breast cancer pedigrees, each containing 68—159 cases, with tumors and PAM50 dimensions available for 20—35 cases per pedigree. We first identified groups of women within pedigrees whose tumors were significantly more similar than expected under the premise that this would reduce germline heterogeneity. To achieve this we performed hierarchical clustering within pedigrees for each of PC1-PC5. Resulting groups were compared to 911 sporadic breast tumors to determine excessive clustering. This procedure resulted in 22 groups, including the previously published PC3-extreme group. Subsequent SGS analysis for selected groups identified 15 genomewide significant regions. Of particular interest, was a 1.04 Mb region at 4q21 shared by 6 breast cancer cases across 30 meioses ($p=8.710^{-9}$), containing *CNOT6L*, another subunit of CCR4-NOT. Another region harboring intriguing candidates is 8p22 (0.4 Mb, 6 cases, 30 meioses, $p=2.110^{-8}$), which contains *NAT1* and *NAT2*, drug metabolizing enzymes associated with cancer cell proliferation and survival. *NAT1* is on the PAM50 panel and its expression has been associated with the estrogen receptor (ER) and proposed as a prognostic marker for ER+ cancers. In summary, we have identified novel breast cancer loci segregating in large pedigrees using an innovative approach that incorporates tumor molecular phenotypes, illustrating its potential to address heterogeneity. These novel tumor traits may contribute not only to mapping risk for breast cancer but also advance tumor characterization.

A- A+

PgmNr 47: A 585bp structural variant in *CTRB2* underlies a pancreatic cancer GWAS risk signal at chr16q23.1.

Authors:
A. Jermusyk [1]; J. Zhong [1]; N. Gordon [1]; I. Collins [1]; E. Abdolalizadeh [1]; S. Parera [2]; T. Zhang [3]; J. Hoskins [1]; K. Connelly [1]; D. Eiser [1]; R. Stolzenberg-Solomon [3]; B. Wolpin [4]; S. Chanock [3]; G. Petersen [5]; J. Shi [3]; C. Westlake [2]; L. Amundadottir [1]; PanScan, PanC4

View Session Add to Schedule

Affiliations:
1) Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA; 2) Laboratory of Cell and Developmental Signaling, Center for Cancer Research, National Cancer Institute, NIH, Frederick, Maryland 21702, USA; 3) Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA; 4) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215 USA; 5) Department of Health Sciences Research, College of Medicine, Mayo Clinic, Rochester, Minnesota, USA

GWAS have successfully identified common variants that influence complex traits, generally with small effect sizes. Most of these loci are believed to mediate their effects through noncoding variants lying within regulatory elements that have subtle effects on gene expression. We performed functional follow-up of a pancreatic cancer risk locus on chr16q23.1 identified through GWAS in Europeans. Fine-mapping identified rs72802365 as the most significant variant at this locus (P = 2.51×10^{-17}, OR = 1.36). This SNP lies between *CTRB1* and *CTRB2*, two chymotrypsin precursor genes which encode serine proteases whose main roles are to digest food. Expression QTL analysis did not reveal allele-specific gene expression for genes within this locus, but we observed a strong splicing QTL for *CTRB2* (*CTRB2* exon5 → exon 7 splicing; rs72802365, P=7.0×10^{-74}, β=1.97 in GTEx), which colocalizes with the GWAS signal (*PP* = 99.6%). As this finding is consistent with a 585bp insertion-deletion variant that overlaps exon 6 of *CTRB2*, we genotyped this deletion in the 1000genomes European ancestry samples (n = 438, MAF=8.2%, r^2=0.70 to rs72802365) and imputed the deletion variant into the pancreatic cancer GWAS datasets (OR = 1.37, P = 2.83×10^{-16}). Conditioning the analysis on rs72802365 or the 585bp *CTRB2* deletion variant resulted in a complete loss of the association signal, indicating the deletion marks the same pancreatic cancer risk signal as rs72802365 and may be the functional variant underlying the association.

The deletion allele of this variant results in a truncated CTRB2 protein due to a premature stop codon early in exon 7. This truncated protein lacks one of the three amino acids in the catalytic triad of the full-length protein. When expressed in HEK293 cells, the truncated protein showed a 7.6-fold lower chymotrypsin activity (P = 0.017) and 5.1-fold lower secretion (P = 3.7×10^{-4}) as compared to full length CTRB2. Increased expression of ER stress markers was observed in cell lines, and in normal pancreas tissue samples correlated with deletion allele dosage (GTEx: GSEA and IPA, FDR < 1×10^{-4}). In summary, we identified a 585bp genomic deletion as a functional variant underlying risk at the chr16q23.1 pancreatic cancer locus. Individuals carrying this deletion express an inactive CTRB2 protein that accumulates intracellularly. The ensuing ER stress likely contributes to pancreatic cancer risk at this locus.

A- A+

PgmNr 48: Variable expression quantitative trait loci analysis of breast cancer risk variants.

Authors:
G. Wiggins [1]; J. Pearson [1,2]; M. Black [3]; A. Dunbier [3]; T. Merriman [3]; L. Walker [1]

View Session Add to Schedule

Affiliations:
1) Pathology and Biomedical Science, University of Otago, Christchurch, Christchurch, New Zealand; 2) Biostatistics and Computational Biology Unit, University of Otago Christchurch, NZ.; 3) Department of Biochemistry, University of Otago Dunedin, NZ.

Genome wide association studies in breast cancer have identified more than 170 risk variants. A major challenge has been to understand the functional consequences of these variants. Expression quantitative trait loci (eQTL) analysis has been an important approach to identify candidate susceptibility genes at risk loci on the basis that expression of many genes is regulated by genetic variants. Studies to date have focussed on the mean expression levels of nearby genes in relation to the genotype at each locus. Here, we demonstrate a new variable expression quantitative trait loci (veQTL) method for testing the association of variants with the variability (not mean) of gene expression and identify new candidate genes and pathways associated with risk variants. We predict that veQTLs interact in a genotype-specific manner with intrinsic and extrinsic exposures.
We performed veQTL analysis using genotype and tissue specific RNA-sequencing data from 635 subjects obtained through GTEx (v7). Four tissue-specific (breast, ovary, kidney and lung) veQTL datasets were generated for known breast cancer risk variants. This analysis found that genes identified by veQTL were distinctly different from those identified by eQTL (rank correlation=0.15). For breast tissue, a significantly greater number of genes were found at eQTL (n=448) compared to veQTL(n=39). Pathway enrichment analysis highlighted the differences in the biological functions of genes identified by each method. Genes at veQTL were enriched for pathways involved in androgen and glucocorticoid processes, while genes at eQTL were enriched in pathways involved in leukocyte activity. Only 10/39 breast veQTLs were observed in a second tissue, nine of which were found in lung. No veQTL was shared between more than two tissues, suggesting a tissue-specific effect of exposures.
Here we have described a novel method for identifying candidate susceptibility genes associated with breast cancer risk. Furthermore, these genes were shown to be involved in hormonal processes that are associated with breast tumourigenesis, thus highlighting a potential relationship between genetic regulation, expression variability and breast cancer development. This study has demonstrated the potential utility of veQTL analysis to provide key insights into genes and pathways underlying variants identified by genome wide association studies.

A- A+

PgmNr 49: Oncogenes co-opt endogenous and ectopic enhancers on extrachromosomal DNA amplifications.

Authors:
A.R. Morton [1]; N. Dogan-Artun [2]; G. MacLeod [3]; C.F. Bartels [1]; M.S. Piazza [4]; S.C. Mack [5]; X. Wang [6]; R.C. Gimple [6,7]; Q. Wu [6]; Z.J. Faber [1]; B.P. Rubin [8]; S. Shetty [4]; S. Angers [3,9]; P.B. Dirks [10,11]; M. Lupien [2,11,12]; J.N. Rich [6,13]; P.C. Scacheri [1]

View Session Add to Schedule

Affiliations:
1) Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA; 2) Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 1L7, Canada; 3) Leslie Dan Faculty of Pharmacy, University of Toronto, Ontario M5S 3M2, Canada; 4) Center for Human Genetics Laboratory, University Hospitals, Cleveland, OH 44106, USA; 5) Department of Pediatrics, Division of Hematology and Oncology, Baylor College of Medicine, Texas Children's Hospital, Houston,?TX 77030, USA; 6) Department of Medicine, Division of Regenerative Medicine, University of California, San Diego, La Jolla, CA 92037, USA; 7) Department of Pathology, Case Western Reserve University, Cleveland, OH 44120, USA; 8) Departments of Anatomic Pathology and Molecular Genetics, Cleveland Clinic, Lerner Research Institute and Taussig Cancer Center, Cleveland, Ohio 44195, USA; 9) Department of Biochemistry, Faculty of Medicine, University of Toronto, Ontario, M5G 0A4, Canada; 10) Developmental and Stem Cell Biology Program and Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada; 11) Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada; 12) Department of Medical Biophysics, University of Toronto, Toronto, ON M5S 1A8, Canada; 13) Department of Neurosciences, University of California, San Diego, School of Medicine, La Jolla, CA 92037, USA

In glioblastoma, focal amplification of *EGFR* occurs in 48% of tumors. Despite the prevalence of this copy number aberration, little is known about the cis-regulatory architecture of *EGFR* amplicons. To determine whether the amplicons contain features other than the oncogene that are important for tumorigenesis, we systematically evaluated the active regulatory regions of the locus through H3K27ac ChIP-seq of 41 patient-derived glioblastoma cell models. Despite the heterogeneous sizes of the *EGFR* amplicons, two enhancer elements located 130-kb and 86-kb upstream of the *EGFR* promoter were co-amplified with *EGFR* in every sample with focal amplification of this locus, indicating selection. These two enhancers showed consistent co-amplification with *EGFR* in 173 of 174 glioblastoma samples from TCGA. 4C-seq studies revealed that the two enhancers physically engage the *EGFR* promoter in both *EGFR*-amplified and unamplified tumors. CRISPRi-targeting of the two enhancers led to a dramatic reduction of *EGFR* expression and compromised viability, indicating that the enhancers are bona fide regulators of *EGFR*. Using paired-end sequence reads to reconstruct amplicon architecture and DNA FISH, we determined that the *EGFR* amplicons occur as double minutes: circularized extrachromosomal DNA common in solid tumors. In addition to contacts between the *EGFR* promoter and its two endogenous enhancers, the double minutes contain new, TAD-oblivious regulatory interactions with *EGFR* that do not normally occur on linear chromosomes. Preferential co-amplification of oncogenes and their endogenous enhancers was also observed for *MYC* in medulloblastoma, and *MYCN* in Wilms' tumor and neuroblastoma, indicating that enhancer selection is a general feature of focal amplifications. Moreover the selected enhancers are specifically active in the presumed cell type of origin for these cancers. Our results indicate that for focal oncogene amplifications, the preexisting regulatory circuitry of the driver locus is preserved from the cell of origin upon eviction from linear chromosomes and subsequent amplification, maintaining the oncogene’s expression. The evicted oncogene adopts ectopic regulatory interactions that further augment its activity and ensure robustness. Our studies indicate that oncogene amplifications are shaped by non-coding regulatory elements, and provide new insights into tumorigenesis and cancer evolution.

A- A+

PgmNr 50: Epigenomic translocation of H3K4me3 broad domain: A mechanism of super-enhancer hijacking following oncogenic translocations.

Authors:
A. Mikulasova [1]; K. Fung [1]; N. Karataraki [1]; B.A. Walker [2]; G.J. Morgan [2]; C. Ashby [2]; A. Corcoran [3]; S. Hambleton [4]; D. Rico [4]; L.J. Russell [1]

View Session Add to Schedule

Affiliations:
1) Northern Institute for Cancer Research, Newcastle upon Tyne, United Kingdom; 2) Myeloma Center, University of Arkansas for Medical Sciences, Little Rock, AR, USA; 3) Babraham Institute, Cambridge, United Kingdom; 4) Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom

---

Introduction: Chromosomal translocations are common events in hematological malignancies with oncogenic power generated via aberrant fusion proteins or juxtaposition of proto-oncogenes and strong regulatory regions promoting their over-expression. One of the most frequent translocations involves the *immunoglobulin heavy locus* (*IGH*) at 14q32.33, due to errors in the recombination processes; V(D)J and class-switch recombination, and somatic hypermutation. H3K4me3 is a histone mark characteristic of active promoters, but broad domains (H3K4me3-BDs) can cover entire genes, providing consistent expression. We present a new model of “epigenomic translocation”, where a wild-type H3K4me3-BD disappears in malignant cells and “re-locates” into the target oncogene of the genomic translocation.
Methods: We used chromatin states of 108 BLUEPRINT samples (PMID:28934481) including healthy B-cells/plasma cells (n=15), T-cells (n=15), myeloid-lineage cells (n=55) and samples from B-cell-derived hematological malignancies (n=23). Myeloma cell line U266 was sequenced using a custom targeted-capture covering 95% of the *IGH*-DJC locus (https://doi.org/10.1101/515106).
Results: Using chromatin states, we fine mapped three B-cell-lineage specific enhancers within the *IGH* constant region: chr14:106,025,200-106,056,800 (Eα2), chr14:106,144,200-106,179,400 (Eα1) and chr14:106,281,800-106,323,000 (Eμ). We also discovered B-cell specific H3K4me3-BD present in healthy donors at chr14:106,346,800-106,387,800, telomeric from the *IGH* promoter. Interestingly, this H3K4me3-BD is absent in samples with known translocation, t(11;14)(p13.3;q32.33)/*IGH-CCND1*, including mantle-cell lymphoma and myeloma samples. These cases showed abnormal appearance of H3K4me3-BD over the *CCND1*. We mapped the translocation in U266 as a cut-and-paste mechanism of the Eα2 from *IGH* next to *CCND1*. We hypothesize that the translocation of the enhancer is followed by interaction between this enhancer and the *CCND1* locus, generating the aberrant H3K4me3-BDs.
Conclusions: We describe how a genomic translocation of the *IGH* enhancer close to the *CCND1* gene results in the “epigenomic translocation” of an H3K4me3-BD, resulting in *CCND1* activation. This could be a wide-ranging mechanism of oncogenic activation in hematological malignancies.

A- A+

PgmNr 51: Somatic mutations in HLA-class I genes: Associations with tumor burden and HLA homozygosity.

Authors:
M. Dean [1]; M. Viard [2]; M. Yeager [3]; V. Naranbhai [4]; M. Carrington [2]

View Session Add to Schedule

Affiliations:
1) Laboratory of Translational Genomics, NCI, Gaithersburg, Maryland.; 2) Center for Cancer Research, National Cancer Institute, NIH, Frederick, MD USA; 3) Leidos Biomedical Research, Inc National Cancer Institute, NIH, Gaithersburg, MD USA; 4) Massachusetts General Hospital, Boston, MA USA

Cancer cells are recognized as foreign by multiple immune cell types and therefore tumors have evolved immune escape mechanisms including mutation of HLA class I (*HLA-A*, *HLA-B*, *HLA-C*), antigen processing (*TAP1*, *TAP2*, *TAPBP*, *TAPBL*) and beta-2-microglobulin (*B2M*) genes, involved in antigen presentation and *CASP8*, critical for T-cell directed apoptosis. To study the role of these mutations in cancer we developed allele-specific alignment and variant calling methods to identify *HLA-A, -B* and *-C* mutations and determine the phase of the mutated allele in public (TCGA) data of 9562 tumors from 13 cancer types. Of tumor types with at least 100 samples, the mutation rate for any *HLA-A, B* or *C* gene is highest in cervical cancer (10.3%), for *B2M* in colon and stomach cancers (4.4 and 4.8%), and *CASP8* in uterine and head and neck cancers (10.2 and 10.5%). Mutations were 2-3 times more frequent in *HLA-A* and *HLA-B* than *HLA-C*, and *HLA-A* and *HLA-B* have a significantly higher proportion of null alleles (51-55%) than *HLA-C* (28%) X^2=18, P<10-4), distinctions that may have to do with the central importance of HLA-C as a ligand for inhibitory killer cell immunoglobulin-like receptors. The number of mutations in the tumor (mutation burden) is known to be variable by cancer type and high mutation burden (MB) has been associated with a positive response to immune checkpoint inhibitors (ICI). One or more mutations in *HLA-I*, *TAP* or *CASP8* genes are associated with higher MB, and tumors with missense mutations in *HLA-A* or *HLA-B* have higher MB than tumors with null alleles in those genes. *HLA-A, B and C* mutations significantly co-occur with each other and with mutations in *TAP1* and *TAP2* genes, suggesting that these lesions are cooperative. Homozygosity in the germline for HLA-I genes is known to associate with poorer outcome in infectious disease, and in ICI therapy. Mutations in *HLA-A* are less common in individuals homozygous for *HLA-A* and *HLA-B*.

In conclusion, several cancers, especially those with high mutation burden show frequent alterations in the genes involved in antigen processing and presentation, and response to immune recognition. It is likely that these mutations will reduce the response to immune therapies and therefore will have prognostic utility.

A- A+

PgmNr 52: Single-cell dissection of brain circuitry across 800+ individuals including AD, schizophrenia, bipolar, ALS, HD, and controls.

Authors:
M. Kellis [1,2]; J. Davila-Velderrain [1]; S. Mohammadi [1]; Y.P. Park [1]; L. He [1]; C. Boix [1]; M. Kousi [1]; J. Mantero [1]; K. Galani [1]; L.-L. Ho [1]; H. Mathys [3]; J. Young [3]; D.A. Bennett [4]; L.-H. Tsai [2,3]

View Session Add to Schedule

Affiliations:
1) Computer Science, MIT & Harvard, Cambridge, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Picower Institute, MIT, Cambridge, MA; 4) Rush University, Chicago, IL

Alzheimer’s disease (AD), schizophrenia (Sz), bipolar disorder (BD), amyotrophic lateral sclerosis (ALS), Huntington’s disease (HD) and related neuropsychiatric and neurodegenerative diseases are pervasive disorders characterized by changes in brain activity, whose complex etiologies involves multiple brain cell types, but remain poorly characterized. To address this challenge, we carry out systematic profiling of single-cell transcriptomes (scRNA-seq) and epigenomes (scATAC-seq) from 800+ post-mortem brain samples across prefrontal cortex and other of disease vs. control individuals, and integrate these data with providing the single-cell cortical views of brain pathologies. We identify transcriptionally-distinct cell subpopulations capturing six major brain cell-types, and analyzed their association with pathology. In the case of Alzheimer’s disease (AD), we uncover coherent neuronal and glial subpopulations associated with AD, and identified regulators of myelination, inflammation, and neuron survival as top markers characterizing them. We find that the strongest AD-associated changes appear early in disease progression, and are highly cell-type specific. By contrast, genes up-regulated in late-stage disease progression are common across cell types, and primarily involved in global stress response. Surprisingly, we found that cells isolated from female individuals are overrepresented in the AD-associated cell subpopulations, revealing substantially different transcriptional responses between sexes. We also combined single-cell profiles, tissue-level variation, and genetic variation across healthy and diseased individuals to deconvolve bulk profiles into single-cell profiles, to recognize changes in cell type proportion associated with disease and aging, and to partition genetic effects into the individual cell types where they act. These revealed a change in cell type proportion associated with aging and with AD, with decreased excitatory and inhibitory neurons, and increased astrocytes, which can become cytotoxic. They also revealed genetic variants underlying cell-type-proportion, including in the TMEM106B locus, which has been previously implicated in Fronto-temporal lobal degeneration (FTLD), and was associated with decreased fraction of inhibitory neurons, but not associated with AD directly. These results provide a roadmap for translating genetic findings into mechanistic insights and ultimately new therapeutic avenues for complex brain disorders

A- A+

PgmNr 53: Integrated analysis of rare variants and single-cell expression data provides insights into schizophrenia risk.

Authors:
T. Nguyen [1]; A. Charney [2]; X. He [3]; K. Kendler [1]; P. Sullivan [4]; S. Bacanu [1]; B. Riley [1]; E. Stahl [2]

View Session Add to Schedule

Affiliations:
1) Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, USA.; 2) Psychiatry Department, Icahn School of Medicine at Mount Sinai, New York, USA.; 3) Department of Human Genetics, University of Chicago, Chicago, 60637, IL, USA.; 4) Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, 27599-7264, North Carolina, USA.

Schizophrenia is a severe, complex neuropsychiatric disorder affecting >2 million people in the US. Even though the disorder has a high genetic heritability (60-80%) and common-variant studies have identified multiple risk loci, only a handful of risk genes have been discovered by rare-variant studies. However, rare disruptive and damaging variants of schizophrenia are enriched in some specific biological pathways and are concentrated in brain-expressed genes, particularly in neuronally expressed genes. Therefore, integrating brain gene-expression information into the analysis of rare variants can increase the power of risk gene identification for schizophrenia. Here, we jointly modeled single-cell gene-expression data and rare variants to identify additional risk genes; and to better understand the genomic architecture for schizophrenia. We analyzed different types of rare variants from 2,772 parent-offspring trios, 5,601 cases and 10,634 controls; and 149 gene sets from mouse single-cell RNA sequencing expression data using gTADA, an integrative method developed recently by our group. We identified 50/149 enriched gene sets. Using these 50 gene sets, we prioritized 9 genes with maximum posterior probability (maxPP) > 0.95; including *TRIO, SETD1A, TAF13* which were reported in previous studies. To better understand the top genes, we further analyzed genes with maxPP>0.8 by using gene-expression data from the BrainSpan consortium. We observed that these genes were significantly up-regulated in early pre-natal stages and down-regulated in post-natal stages. When analyzed on both spatial and temporal scales, significant results were observed for the cortical regions in both early fetal and early mid-fetal stages. Interestingly, those genes were also expressed in the early fetal stage of the cerebellum and thalamus.
Our results present a new integrated framework for the analysis of rare variants and functional genomic data as well as new biological insights into schizophrenia.

A- A+

PgmNr 54: Human single-cell transcriptomes identify cell-types and states relevant to brain disorders.

Authors:
S. Gerges [1,2,3,5]; T. Singh [1,2,3]; M. Goldman [1]; S. Berretta [1,4]; S. McCarroll [1,2,3]; M. Daly [1,2,3]

View Session Add to Schedule

Affiliations:
1) Department of Genetics, Harvard Medical School, Boston, MA.; 2) Stanley Center, Broad Institute, Cambridge, MA; 3) Analytic and Translational Genetics Unit, MGH, Boston, MA; 4) Harvard Brain Tissue Resource Center, McLean Hospital, Belmont, MA.; 5) John A. Paulson School Of Engineering And Applied Sciences, Harvard University, Cambridge, MA

Genetic studies have identified hundreds of genetic loci associated with brain-related traits. However, the fundamental neurobiology that explains how genetic etiology translates into disease manifestation remains elusive. Single-cell RNA-sequencing (RNA-seq) allows for unprecedented resolution into the cell-types most relevant to psychiatric and neurological disease etiology as well as complex traits more broadly.

Here, we combine single-cell RNA-seq of over 40,000 cells from caudate and prefrontal cortex from multiple individuals with genome-wide association study (GWAS) summary statistics and rare-variants from over 25 brain-related traits. We show that brain traits have both shared and distinctive expression signatures which map onto diverse cell-types and states in the human brain. Specifically, we use partitioned LD score regression (LDSC) to show that schizophrenia, bipolar disorder, educational attainment and IQ are most enriched in a new cell-type population known as “eccentric” spiny projection neurons (eSPN) (p=1.29e-15, p=9.83e-11, p=9.22e-10 and p=3.72e-10 respectively), as well as both “direct” (dSPN) and “indirect” (iSPN) populations in the patch and matrix compartments of the striatum (p<1e-10) and cortical pyramidal neurons (p<1e-8). Importantly, while the cell-type associations are similar, we demonstrate they are driven by different sets of genes.

We contrast this with other traits, such as neuroticism and major depressive disorder which are exclusively enriched in excitatory pyramidal neurons (p < 1.0e-05) and interneurons defined during development (p=1.23e-06), respectively. In contrast to psychiatric and cognitive traits, we show that neurological disorders such as Alzheimer’s disease and multiple sclerosis are significantly enriched in specific types of “disease-associated” microglial cell-states.

To our knowledge, this is the most in depth single-cell RNA-seq survey of the human caudate nucleus, which is associated with roles in procedural learning, and one of the most comprehensive surveys of the frontal cortex, which is associated with higher-order functions. Importantly, both regions are heavily implicated in brain disorders, and provide an atlas for further investigation of brain phenotypes. Altogether, our approach to utilizing human single-cell data allows for unprecedented resolution into the most relevant cells that contribute to disease onset using the both rare and common genetic studies of the brain.

A- A+

PgmNr 55: Regional heterogeneity in gene expression, regulation, and coherence in the frontal cortex and hippocampus across development and schizophrenia.

Authors:
L. Collado Torres; E.E. Burke; A. Peterson; J.H. Shin; R.E. Straub; A. Rajpurohit; S.A. Semick; W.S. Ulrich; BrainSeq Consortium

View Session Add to Schedule

Affiliation: Lieber Institute for Brain Development, Baltimore, MD

Introduction

The hippocampus formation, although prominently implicated in schizophrenia pathogenesis, has been overlooked in large-scale genomics efforts in the schizophrenic brain.

Methods

We performed RNA-seq in hippocampi and dorsolateral prefrontal cortices (DLPFCs) from 551 individuals (286 with schizophrenia) across the human lifespan. Expression was quantified with annotation-based (gene, exon, transcript) and agnostic methods (exon-exon junctions) on hg38 Gencode v25. RNA degradation was adjusted for using the quality surrogate variable (qSVA) framework adapted for multiple brain regions.

Results

We identified substantial regional differences in gene expression and found widespread developmental differences that were independent of cellular composition. We identified 48 and 245 differentially expressed genes (DEGs) associated with schizophrenia within the hippocampus and DLPFC, with little overlap between the brain regions. 124 of 163 (76.6%) of schizophrenia GWAS risk loci contained eQTLs in any region. Transcriptome-wide association studies (TWAS) in each region identified many novel schizophrenia risk features that were brain region-specific. Last, we identified potential molecular correlates of *in vivo* evidence of altered prefrontal-hippocampal functional coherence in schizophrenia.

Conclusions

These results underscore the complexity and regional heterogeneity of the transcriptional correlates of schizophrenia and offer new insights into potentially causative biology. The little overlap among DEGs in DLPFC and hippocampus and with TWAS-associated genes demonstrate that schizophrenia DEGs likely reflect consequences of this disorder. Furthermore, the decreased coherence between DLPFC and hippocampus in schizophrenia molecularly validate previous neuroimaging findings. Finally, we have created an extensive eQTL and expression resource as well as provided TWAS weights for human brain across development and in schizophrenia that will facilitate future analyses.

Resources

eQTL browser, raw and processed data: http://eqtl.brainseq.org/phase2/
Code: https://github.com/LieberInstitute/brainseq_phase2/ and
https://github.com/LieberInstitute/qsva_brain/

In press: https://doi.org/10.1016/j.neuron.2019.05.013
Pre-print: https://doi.org/10.1101/426213
Twitter summaries: https://twitter.com/andrewejaffe/status/1135932160622387200 and https://twitter.com/fellgernon/status/1135944857707401217

A- A+

PgmNr 56: Single-cell RNA-sequencing reveals cell-type-specific levels of aneuploidy in mammalian nervous system.

Authors:
T. Li [1]; J. Zhu [1]; R. Li [1, 2]

View Session Add to Schedule

Affiliations:
1) Center for Cell Dynamics, Department of Cell Biology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 2) Department of Chemical and Biomolecular Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA

There are conflicting reports on the prevalence of aneuploidy in the mammalian nervous system, with estimates ranging from <1% to 33% of neurons exhibiting aneuploidy. Recent work suggests that subsets of human neurons harbor large-scale genome alterations, including copy-number variations and chromosomal changes.

To investigate the extent and functional relevance of neuronal aneuploidy, we leveraged three recent large-scale single-cell RNA-sequencing (scRNA-seq) datasets (Tabula Muris, Allen Brain Atlas, and mousebrain.org) to search for evidence of chromosome-level genomic alterations in mouse, macaque, and human cells. Specifically, we developed an integrated computational pipeline to de-noise sparse scRNA-seq data, ascertained aneuploidy status based on the statistical distributions of gene expression levels for each chromosome, and quality-controlled aneuploidy results by random permutation of de-noised matrices. This resulted in a comprehensive survey of more than 212K neurons and glial cells in total (194K mouse, 1K macaque, and 17K human).

We found that across the three species, there are significant variations in the prevalence of aneuploidy across different brain regions (0 - 14%), consistent with the widely varying results from previous studies. Importantly, cell type plays a key role, with vast majority of chromosome changes appearing in glutamatergic neurons. We observed diverse karyotype changes, but identified human chr22 and mouse chr18 to be hotspots for chromosomal gains. Additionally, aging is positively correlated with aneuploidy rate in humans, and genetic markers of aneuploidy are enriched in aging and neurodegenerative diseases, synaptic transmission, mitosis, cytoskeleton, and metabolic processes. Finally, in the mouse PNS, we found that myenteric plexus of the small intestine has the highest levels of aneuploidy (11%).

Overall, we have established a novel computational pipeline that can systematically identify aneuploidy using large-scale scRNA-seq datasets. Our results suggest that chromosomal changes are differentially regulated in different cell types and different regions of the mammalian nervous system which, given the recent evidence of ongoing adult neurogenesis in the human brain, can have profound implications in further elucidating the molecular mechanisms and disease relevance of neuronal aneuploidy.

A- A+

PgmNr 57: Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells.

Authors:
H. Tilgner

View Session Add to Schedule

Affiliation: Brain and Mind Research Institute, Weill Cornell Medicine, New York City, NEW YORK.

Full-length RNA sequencing (RNA-Seq) has been applied to bulk tissue, cell lines and sorted cells to characterize transcriptomes, but applying this technology to single cells has proven to be difficult, with less than ten single-cell transcriptomes having been analyzed thus far. Although single splicing events have been described for ≤200 single cells with statistical confidence, full-length mRNA analyses for hundreds of cells have not been reported. Single-cell short-read 3' sequencing enables the identification of cellular subtypes, but full-length mRNA isoforms for these cell types cannot be profiled. We developed a method that starts with bulk tissue and identifies single-cell types and their full-length RNA isoforms without fluorescence-activated cell sorting. Using single-cell isoform RNA-Seq (ScISOr-Seq), we identified RNA isoforms in neurons, astrocytes, microglia, and cell subtypes such as Purkinje and Granule cells, and cell-type-specific combination patterns of distant splice sites. We used ScISOr-Seq to improve genome annotation in mouse Gencode version 10 by determining the cell-type-specific expression of 18,173 known and 16,872 novel isoforms.

A- A+

PgmNr 58: Global Biobank Meta-analysis Initiative: Powering genetic discovery across human diseases.

Authors:
W. Zhou [1]; B.M. Neale [1]; M.J. Daly [1,2]; on behalf of Global Biobank Meta-analysis Initiative

View Session Add to Schedule

Affiliations:
1) Broad Institute, Massachusetts General Hospital, Harvard Medical school, Boston, Massachusetts.;
2) Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Helsinki, Finland

With electronic health records and questionnaires linked to genomic data, biobanks provide unprecedented opportunities for systematically understanding genetic and environmental contributions to complex diseases. The Global Biobank Meta-analysis Initiative aims to create a framework to jumpstart global biobank collaboration. The benefits include better power for genome-wide association studies (GWASs), GWASs of understudied diseases, the opportunity for cross-validation, improvements in fine-mapping, and potential to explore subgroup analyses. With scale comes the opportunity to develop collaborations around rare or understudied traits as well as deepen genetic analysis of longitudinal phenotypes and trajectories. 18 biobanks have committed to this project so far, including Biobank Japan, BioME (USA), BioVu (USA), China Kadoorie, Colorado Biobank (USA), deCODE Genetics (Iceland), East London Genes & Health (UK), Estonian Biobank, FinnGen (Finland), Generation Scotland, HUNT (Norway), LifeLines (Netherlands), Mexico City, Michigan Genomics Initiative (USA), Million Veteran Program (USA), Partners Biobank (USA), UCLA Precision Health Biobank (USA) and UK Biobank, bringing the total sample size to more than 2 million.

We describe here initial pilot uses of this collaborative network. We harmonized phenotype definition and performed GWAS across biobanks for pilot common endpoints: asthma, atrial fibrillation (AFib), rheumatoid arthritis (RA), glaucoma, and colorectal cancer. Meta-analysis boosted the number of genome-wide significant loci compared to the sum of loci from individual biobanks. For example, 103 loci were identified for AFib with a total of 46,176 cases across UK Biobank (UKB), FinnGen, HUNT, Partners, Generation Scotland, and Biobank Japan while 50 loci were identified from the union of individual biobanks’ results and showed consistency with established GWAS from the respective cohorts. Beyond loci, we showed high genetic correlation (0.78-1.0) for asthma, AFib, and RA between UKB and FinnGen although different phenotype curation approaches were used. In addition, a preliminary genome-wide survival analysis has been successfully conducted in one of the biobanks, anticipating the value of this approach in the network.

We use exemplar phenotypes to explore the impact of phenotype harmonization, imputation panel, and statistical methods towards the development of a more definitive, global resource for human genetics with mature analytic expertise.

A- A+

PgmNr 59: Biclustering of a large gene to phenome catalog of associations unveils hidden connections between complex traits and genes.

Authors:
M. Pividori; A. Barbeira; H. Im

View Session Add to Schedule

Affiliation: Department of Medicine, The University of Chicago, IL, USA

Background. The growing number of genome-wide association results and large scale biobanks with deep genome and phenome coverage provide unprecedented opportunities to interrogate the biology of complex phenotypes and diseases. However, this large and continuously growing amount of studies also poses the challenge of interpretation and summarization, where important hidden relationships might be lost in a deluge of new results. Novel data analytic and visualization methods to find hidden connections are urgently needed to take full advantage of the data and accelerate the pace of discovery in biomedical sciences.

Methods. We performed a PrediXcan analysis on 4077 phenotypes from the UK Biobank based on publicly available GWASs, where evidence of association across tissues was aggregated using MultiXcan. To discover hidden patterns in this large association matrix (4077 traits and 19910 genes), we applied BiMax, a biclustering algorithm to simultaneously partition traits and genes into biologically interpretable sets. A bicluster is defined as a subset of traits that are jointly associated to a subset of genes. The method used here supports overlapping biclusters, which allows genes to be linked to multiple traits as well as traits linked to multiple genes. We also propose a visualization method to analyze this network of connections for a particular trait.

Results. We created a large catalog of gene-level results and found expected patterns of associations such as biclusters with anthropometric traits and known genes including *TFAP2B*, which was linked to skeletal abnormalities; we also found “Age started wearing glasses”, different eye measurements and blood pressure clustered together, with gene *TXLNB* implicated in muscle weakness diseases. Another interesting bicluster included high cholesterol, cholelithiasis (gallstones in the gallbladder) and dietary traits such as processed meat, fish and sodium intake. This bicluster included genes *RASIP1*, *FUT1*, *FUT2* and *IZUMO1* from a gene-dense region in chromosome 19 (within 100kb) which contains *FGF21* associated with dietary macronutrient intake. We also applied our visualization method to analyze how trait “Medication: insulin” was connected to other traits and genes (http://hakyimlab.org/miltondp/ashg2019/). The plot revealed a set of expected relationships between several autoimmune diseases, dietary traits, medications and genes, providing a single view of the complex interplay between them.

A- A+

PgmNr 60: Multi-trait genome-wide analyses of the brain imaging phenotypes in UK Biobank.

Authors:
C. Wu

View Session Add to Schedule

Affiliation: Department of Statistics, Florida State University, Tallahassee, Florida.

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with an impressive number of complex traits and diseases by analyzing a single trait each time. An interesting observation has been that many genetic variants are associated with multiple, sometimes seemingly unrelated traits. To improve statistical power and offer new biological insights, several methods for multi-trait association test have been proposed.

While appealing, most existing studies analyze less than ten traits jointly. However, deep phenotyping data from epidemiological studies and electronic health records are becoming rapidly available. For example, GWAS of 3,144 brain image-derived phenotypes (IDPs) have been carried out to provide insights into the genetic architecture of brain structure and function (Elliott et al. Nature 2018, 562: 210–216).

We introduce a new method called aMAT for multi-trait analysis of summary statistics from GWAS of an arbitrary number of traits. aMAT has several compelling features that make it potentially useful in many settings. First, by taking the potential singularity of the trait correlation matrix into account, aMAT yields well-controlled Type I error rates when analyzing hundreds of traits. Second, aMAT offers robust statistical power over a wide range of scenarios by combining the testing results from a class of tests data-adaptively. Third, aMAT applies the cross-trait LD score regression and thus captures all relevant sources of estimation error on estimating the trait correlation matrix. Finally, aMAT is computationally efficient because the p-value can be calculated analytically.

We conduct extensive simulations, confirming that aMAT yields well-controlled Type I error rates and achieves robust statistical power across a wide range of scenarios. We apply aMAT to summary statistics for a group of volume related imaging phenotypes in UK Biobank. aMAT identifies 28 lead SNPs spanning in 24 distinct loci, 13 of which are missed by any individual univariate GWAS. Bioinformatic analyses show that the linked genes are enriched in volume related gene sets such as hippocampal subfield CA4 volume. Finally, four additional groups of traits have been analyzed and provided similar conclusions. In summary, our results showcase the power of multi-trait analysis of summary statistics from GWAS of a large number (e.g. hundreds) of related traits to gain insights into the genetic basis of complex traits.

A- A+

PgmNr 61: Evaluating the age-of-onset-dependent genetic architecture of complex disorders in the UK Biobank.

Authors:
Y.-C.A. Feng [1,2,3,4]; T. Ge [1,2,4]; C.-Y. Chen [1,2,3,4]; J. Smoller [1,2,3,4]; B. Neale [1,3,4]

View Session Add to Schedule

Affiliations:
1) Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; 2) Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA; 3) Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; 4) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02138, USA

Genome-wide association studies have revealed that most of the complex traits have an underlying polygenic component. Twin and SNP heritability (h^2) analyses have also shown that the variance explained by genetic variation changes across the lifespan for many traits. However, there has been comparatively less investigation of the $h^2$of age of onset (AOO) for disease phenotypes. The polygenic model suggests there may exist a correlation between AOO and susceptibility to a disease, with individuals that have a higher common variant polygenic risk may develop disease earlier (e.g. depression). Here, we provide a deep genetic investigation into AOO across the European-ancestry subset of UK Biobank (N=361,140) and then extend these analyses to estimate the genetic correlation (r_g) between case-control status and AOO. For each disorder of interest, we examined both self-report (SR) and hospital in-patient (HIP) AOO and analyzed cases where AOO was available. We then performed a case-control GWAS, and within cases, a GWAS for AOO. We estimated r_g between the two traits using LD score regression. Self-reported AOO (0-60+) was consistently younger than that based on in-patient records (30-75+). Interestingly, we found that for disorders with sufficient sample size, many of them showed a strong, negative r_g between the risk of developing the disease and the age of developing the disease, consistent with the prediction of the polygenic model. Genetic correlation results for a range of results include: osteoarthritis (SR: r_g =-0.81, P=9e-4; HIP: -0.49, P=9e-4), asthma (SR: -0.55, P=9e-28; HIP: -0.33, P=0.04), hypertension (SR: -0.79, 2e-55; HIP: -0.72, P=8e-31), MI/CHD (HIP: -0.64, P=0.006), depression (HIP: -0.55, P=0.03), and breast cancer (HIP: -0.63, P=0.04). These results suggest that an earlier AOO correlates with a heavier dose of polygenic risk but the extent of this overlap varies by disease. The high r_g between the two orthogonal information—susceptibility and AOO—also suggests meta-analyzing the two types of GWAS may boost power for loci identification. We will present analysis and characterization of all the common disease phenotypes in the UK Biobank where sufficient data are available. These results enhance our understanding of the genetic architecture of AOO across the phenome and show the potential for gaining power by incorporating AOO into primary discovery GWAS, which may ultimately benefit downstream analysis such as polygenic risk score prediction.

A- A+

PgmNr 62: Exome-by-phenome-wide gene-burden association analyses using electronic health record phenotypes.

Authors:
J. Park [1,2]; S.M. Damrauer [3]; J. Chen [4]; M.D. Ritchie [1,5]; D.J. Rader [1,2,6]; Regeneron Genetics Center

View Session Add to Schedule

Affiliations:
1) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.; 2) Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; 3) Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; 4) Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; 5) Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; 6) Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Background: By coupling large-scale exome sequencing with electronic health records (EHR), genome-first approaches can enhance our understanding of the contribution of rare genetic variants to disease. Previously, we aggregated rare loss-of-function variants in a candidate gene into a 'gene burden' for association with EHR phenotypes to identify both known and novel clinical implications for the gene in human disease. However, this methodology has not yet been applied on an exome-wide level, and the clinical ontologies of rare loss-of-function variants in many genes have yet to be described.
Methods: We leveraged exome sequencing data in the Penn Medicine Biobank (N=11,451) to address on an exome-wide scale the association of each gene's rare loss-of-function variants with diverse EHR phenotypes using a phenome-wide association study (PheWAS) approach. We collapsed rare (≤0.1% minor allele frequency) predicted loss-of-function variants (frameshift insertions/deletions, gain/loss of stop codon, splice site disruption) for 1518 sufficiently-powered genes to perform gene-burden PheWAS.
Results: We identified 12 exome-by-phenome-wide significant genes (p-threshold=0.5/(1518 genes*1000 binary phenotypes)≈3E-08). Of these, known positive-control associations included *TTN* (cardiomyopathy, p=7.83E-13), *MYBPC3* (hypertrophic cardiomyopathy, p=3.48E-15), *CFTR* (cystic fibrosis, p=1.05E-15), and *CYP2D6* (adverse effects due to opiates/narcotics, p=1.50E-09). Notably, of the eight previously undescribed associations, we identified two novel associations with hypertrophic cardiomyopathy (*BBS10*, p=2.89E-08; *GDAP2*, p=1.46E-09) and a novel association with severe manifestations of diabetes mellitus (*RGS12*, p=3.36E-08). Additionally, echocardiography and serum laboratory measurements revealed that *BBS10* and *GDAP2* are associated with distinct types of cardiac hypertrophy, and that *RGS12* is associated with diabetic metabolic profiles (*e.g.* increased serum glucose).
Conclusion: Our study presents eight gene-disease associations not previously described, of which we suggest novel roles for *BBS10* and *GDAP2* in two distinct types of hypertrophic cardiomyopathy, and a novel role for *RGS12* in diabetes mellitus leading to severe outcomes (*e.g.* ophthalmic, renal, neurological manifestations). Furthermore, we show the value of aggregating rare, predicted loss-of-

function variants into gene burdens on an exome-wide scale for association with EHR phenotypes to identify novel gene ontologies.

A- A+

PgmNr 63: Trans-ethnic mega-biobank polygenic risk score analysis involving 676,000 individuals identified blood pressure and obesity as causal drivers affecting human longevity.

Authors:
S. Sakaue [1,2,3]; M. Kanai [1,2,4,5,6,7,8]; J. Karjalainen [4,5,6,8]; M. Akiyama [1,9]; M. Kurki [4,5,6,8]; N. Matoba [1]; A. Takahashi [1,10]; M. Hirata [11]; M. Kubo [12]; K. Matsuda [13]; Y. Murakami [14]; M. Daly [4,5,6,8]; Y. Kamatani [1,15]; Y. Okada [1,2,16]; FinnGen Project

View Session Add to Schedule

Affiliations:
1) Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; 2) Department of Statistical Genetics, Osaka University Graduate School of Medicine, Osaka, Japan; 3) Department of Allergy and Rheumatology, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan; 4) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 5) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 6) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA; 7) Department of Biomedical Informatics, Harvard Medical School, Boston, MA; 8) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; 9) Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan; 10) Department of Genomic Medicine, Research Institute, National Cerebral and Cardiovascular Center, Suita, Japan; 11) Laboratory of Genome Technology, Institute of Medical Science, the University of Tokyo, Japan; 12) RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; 13) Department of Computational Biology and Medical Sciences, Graduate school of Frontier Sciences, The University of Tokyo, Tokyo, Japan; 14) Division of Molecular Pathology, the Institute of Medical Sciences, the University of Tokyo, Tokyo, Japan; 15) Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan; 16) Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Osaka, Japan

Polygenic risk scores (PRSs) have successfully shown their predictive ability to point those with several-fold higher inherited risk of a given disease. While nation-wide biobanks have contributed in this context, trans-biobank meta-analysis is warranted to achieve enough statistical power (i.e. sample size) towards clinical application of PRSs. The previous PRS studies mainly focused on identifying individuals with inborn risks of a specific disease. However, a strategy to utilize PRSs as instruments to prioritize acquired factors affecting health outcomes has not been developed. To address this, we collaborated with three trans-ethnic nation-wide biobanks involving 675,898 individuals (FinnGen [FG] in Finland, UK Biobank [UKBB] in UK, and BioBank Japan [BBJ] in Japan), and sought to investigate the clinical phenotypes which are causally associated with longevity by utilizing PRSs as instrumental variables. To maximize statistical power, we constructed population-specific PRSs by introducing a recently developed statistical method of 10-fold leave-one-group-out (LOGO) meta-analysis. In LOGO, (i) a whole cohort was randomly split into ten sub-groups for which GWASs were separately conducted, (ii) for each sub-group, we calculated PRSs based on the meta-analysis results of the nine sub-group GWASs, and (iii) the PRSs were used for association test in the one withheld sub-group, and the statistics were further meta-analyzed across 10 sub-groups. Using LOGO, trans-ethnic meta-analyses revealed that the high systolic blood pressure was trans-ethnically a causal driver for the shorter lifespan (P for lifespan=3.9×10^{-13}, and P for parental lifespan=2.0×10^{-86}). We also found that PRS of obesity-related traits had strikingly heterogeneous effects on lifespan between European and Asian populations (P_{FG}=1.5×10^{-8}, P_{UKBB}=$1.7\times10^{-11}$and P_{BBJ}=0.094), which is concordant with epidemiology. Our study is epoch-making in showing the potential of the PRS study in the genetics-driven identification of causal drivers for health outcomes, which was enabled by collaboration with trans-ethnic, large-scale, deep-phenotyped and followed-up biobanks. This study would be expected to contribute to the improvement of healthcare in that the identified causal factors can be modified by medical intervention.

A- A+

PgmNr 64: Large-scale genome-wide association studies identify novel susceptibility loci for myocardial infarction.

Authors:
J.A. Hartiala [1]; Y. Han [1]; Q. Jia [1]; Z. Kurt [2]; P. Huang [1]; N.C. Woodward [1]; J. Gukasyan [1]; D.A. Trégouët [3]; N.L. Smith [4,5,6]; M. Seldin [7]; C. Pan [7]; M. Mehrabian [7]; A.J. Lusis [7]; P. Bazeley [8]; A.A. Quyyumi [10]; M. Scholz [11,12]; J. Thiery [13,12]; W. März [14,15,16]; L.J. Howe [17]; F.W. Asselbergs [17,18,19]; R.S. Patel [17,20]; L.P. Lyytikäinen [21,22,23]; M. Kähönen [24,25]; T. Lehtimäki [21,22]; T.V.M. Nieminen [26]; J.O. .Laurikka [27,28]; X. Yang [2]; W.H.W. Tang [8]; S.L. Hazen [9]; H. Allayee [1]; The GENIUS-CHD Consortium

View Session Add to Schedule

Affiliations:
1) Departments of Preventive Medicine and Biochemistry & Molecular Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033; 2) 2)Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA 90095; 3) 3)Institut National pour la Santé et la Recherche Médicale (INSERM) UMR_S 1219, Bordeaux Population Health Research Center, University of Bordeaux, France; 4) Department of Epidemiology, University of Washington, Seattle, WA; 5) Seattle Epidemiologic Research and Information Center, Office of Research and Development, Department of Veterans Affairs, Seattle WA;; 6) Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle WA; 7) Departments of Medicine, Human Genetics, and Microbiology, Immunology, & Molecular Genetics, David Geffen School of Medicine of UCLA, Los Angeles, CA 90095; 8) Departments of Cardiovascular Medicine and Cellular & Molecular Medicine, and Center for Clinical Genomics, Cleveland Clinic, Cleveland, OH 44195; 9) Departments of Cardiovascular Medicine and Cellular & Molecular Medicine, Cleveland Clinic, Cleveland, OH 44195; 10) Division of Cardiology, Department of Medicine, Emory Clinical Cardiovascular Research Institute, Emory University School of Medicine. 1462 Clifton Rd NE, Suite # 513, Atlanta, GA 30322; 11) Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany; 12) LIFE Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany; 13) Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, University Hospital, Leipzig, Germany; 14) Vth Department of Medicine, Medical Faculty Mannheim, Heidelberg University, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany; 15) SYNLAB Academy, SYNLAB Holding Deutschland GmbH, Mannheim, Germany; 16) Clinical Institute of Medical and Chemical Laboratory Diagnostics, Medical University of Graz, Graz, Austria; 17) Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, UK; 18) Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands; 19) Health Data Research UK and Institute of Health Informatics, University College London, London, UK; 20) Bart's Heart Centre, St Bartholomew's Hospital, London, EC1A2DA, UK; 21) Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland; 22) Department of Clinical Chemistry, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere 33014, Finland; 23) Department of Cardiology, Heart Center, Tampere University Hospital, Tampere 33521, Finland; 24) Department of Clinical Physiology, Tampere University Hospital, Tampere 33521, Finland; 25) Department of Clinical Physiology, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere 33014, Finland;; 26) Department of Internal Medicine, Päijät-Häme Central Hospital, Lahti, Finland; 27) Department of Cardio-Thoracic Surgery, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere 33014, Finland; 28) Department of Cardio-Thoracic Surgery, Heart Center,

Tampere University Hospital, Tampere 33521, Finland

To further define the genetic architecture of myocardial infarction (MI), we performed a meta-analysis of genome-wide association study (GWAS) data with ~7.8 million SNPs in 61,505 MI cases and 577,745 controls from the UK Biobank and the CARDIoGRAM+C4D Consortium and compared these results to those from a similar meta-analysis for coronary artery disease (CAD). In total, we identified eight novel loci significantly associated with MI at the genome-wide threshold ($p=5.0x10^{-8}$). Of these loci, four were also significantly (*AHDC1* and *FHL5*) or suggestively (*PDLIM5* and *GPSM1*) associated with CAD as well. Most notably, the other four loci (*SLC44A3, IL1F10, PDE1A,* and *COX5A-RPP25*) were significantly associated with MI but yielded only weak associations with CAD. To investigate the specificity of the association signals at these latter four loci further, we stratified the 33 ,086 CAD cases from UK Biobank into those with (n=17,505) or without (n=15,581) MI. This analysis revealed preferential association of the *SLC44A3* (p=0.019), *PDE1A* ($p=3.0x10^{-5}$), and *COX5A-RPP25* ($p=1.7x10^{-3}$) loci with MI in the presence of coronary atherosclerosis. Association of the *SLC44A3* locus with MI in the presence of CAD was further replicated in 13,925 subjects from the GeneBank, Emory Cardiovascular Biobank, ANGES/FINCAVAS, LURIC, LIFE-Heart, and UCORBIO cohorts (OR=1.16, 95% CI 1.09-1.23; $p=3.3x10^{-6}$). Taken together, these results identify several novel susceptibility loci for MI and CAD, and support the concept that some of the genetic determinants of plaque rupture and thrombus formation may be distinct from those that contribute to development of coronary atherosclerosis.

A- A+

PgmNr 65: Whole genome sequencing association analysis of stroke and its subtypes in a multi-ethnic population from Trans-Omics for Precision Medicine (TOPMed).

Authors:
Y. Hu [1]; J. Haessler [1]; P. Auer [2]; K. Wiggins [3]; A. Moscati [4]; A. Beiser [5,6]; N. Heard-Costa [5]; L. Raffield [7]; J. Chung [8]; S. Marini [9,10]; C. Anderson [9,10,11]; J. Rosand [9,10,11]; H. Xu [12]; L. Lange [13]; A. Correa [14]; S. Seshadri [5]; S. Rich [15]; R. Do [4,16]; R. Loos [4,17]; J. Bis [3]; T. Assimes [18]; B. Silver [19]; S. Liu [20]; R. Jackson [21]; S. Smoller [22]; B. Mitchell [11]; M. Fornage [23]; A. Reiner [1,24]; C. Kooperberg [1]; the TOPMed Stroke Working Group

View Session Add to Schedule

Affiliations:
1) Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; 2) School of Public Health, University of Wisconsin–Milwaukee, Milwaukee, WI; 3) Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA; 4) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; 5) Department of Neurology, Boston University School of Medicine, Boston, MA; 6) Department of Biostatistics, Boston University School of Public Health, Boston, MA; 7) Department of Genetics, University of North Carolina, Chapel Hill, NC; 8) Department of Medicine, Boston University School of Medicine, Boston, MA; 9) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 10) Medical and Population Genetics, Broad Institute, Cambridge, MA; 11) Department of Neurology, Massachusetts General Hospital, Boston, MA; 12) Department of Medicine, University of Maryland School of Medicine, Baltimore, MD; 13) Department of Medicine, University of Colorado, Denver, CO; 14) Department of Pediatrics and Medicine, University of Mississippi Medical Center, Jackson, MS; 15) Center for Public Health Genomics, University of Virginia, Charlottesville, VA; 16) Department of Genetics and Genomic Sciences, The Icahn School of Medicine at Mount Sinai, New York, NY; 17) The Mindich Child Health and Development Institute, The Icahn School of Medicine at Mount Sinai, New York, NY; 18) Department of Medicine, Stanford University, Stanford, CA; 19) Department of Neurology, University of Massachusetts Medical School, Worcester, MA; 20) Department of Epidemiology, School of Public Health, Brown University, Providence, RI; 21) Division of Endocrinology Diabetes and Metabolism, The Ohio State University, Columbus, OH; 22) Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, NY; 23) Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX; 24) Department of Epidemiology, University of Washington, Seattle, WA

Stroke is the second leading cause of death and a leading cause of long-term disability worldwide. Previous genome-wide association studies (GWAS) have identified 52 loci containing common variants for stroke in predominantly European populations. Using whole genome sequencing (WGS) data in a multi-ethnic population from the TOPMed (freeze6) Program, we aimed to identify novel variants, especially those of lower frequency or ancestry-specific rare variants, for all stroke (AS), ischemic stroke (IS), hemorrhagic stroke (HS) and their subtypes: large artery (LAS), cardioembolic (CES), small vessel (SVS); intracerebral (ICH) and subarachnoid (SAH). Data were available on 34,000 participants (6,855 cases and 27,145 controls), consisting of 22,351 European, 7,890 African American, 2,618 Hispanic, 850 Asian, 54 Native American and 237 other ancestry. Over 30 million variants with minor allele count (MAC)>20 were examined using the Scalable and Accurate Implementation of

GEneralized mixed model (SAIGE) approach, adjusting for age, sex, study, ancestry, relatedness, and the first 10 principal components. Ancestry-specific analyses were performed in European and African ancestry populations. TOPMed European-ancestry association results for IS were meta-analyzed with data from 28,408 UK Biobank (UKBB) participants (4,500 cases). Variants with *P*<5E-9 were considered as genome-wide significant (although multiple hypothesis testing was not corrected). In the ancestry-combined TOPMed analyses, two novel loci achieved genome-wide significance: *FAM173B* for AS, HS and ICH (MAF=5%, *P*<2E-10), and *13q33* for LAS (MAF=0.1%, *P*=4E-9). Three novel loci reached suggestive significance (*P*<5E-8): *RAP1GAP2* and *AUTS2* for IS (MAF=2.2% and 0.3%, *P*<4E-8), and *7p22* for HS (MAF=0.1%, *P*=4E-8). In addition, 47/52 known loci were available for validation in TOPMed, and 21 were at least nominally associated with at least one stroke type/subtype (*P*<0.05). In the African-ancestry analyses, *TEX13C* was identified as a novel locus for CES (MAF=3.5%, *P*=2E-8), with the top hit being monomorphic in Europeans. In the TOPMed and UKBB meta-analyses, one novel locus *GRP* (MAF=0.6%) and one reported locus *PITX2* were suggestively associated with IS (*P*<4E-8). Our findings reinforce that WGS in ancestrally diverse populations with large sample sizes are needed for capturing rare and ancestry-specific variants missed by GWAS, and that replication and functional annotation of the novel findings are needed.

A- A+

PgmNr 66: Genetic studies in the eMERGE network and UK Biobank offer new insights into pleiotropy across cardiovascular diseases and central nervous system disorders.

Authors:
Zhang[1,2]; Y. Veturi[2]; A. Verma[2]; T.G. Drivas[2]; W.K. Chung[3]; D. Crosslin[4]; J.C. Denny[5]; D. Fasel[6]; H. Hakonarson[7]; S. Hebbring[8]; G.P. Jarvik[4]; I. Kullo[9]; E.B. Larson[10]; S.A. Pendergrass[11]; L. Rasmussen-Torvik[12]; D. Schaid[13]; P. Sleiman[7]; J.W. Smoller[14]; I.B. Stanaway[4]; W. Wei[15]; C. Weng[16]; M.D. Ritchie[2]

View Session Add to Schedule

Affiliations:
1) Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 2) Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 3) Department of Pediatrics, Columbia University, New York, NY; 4) Departments of Medicine and Genomic Sciences, University of Washington School of Medicine, Seattle, WA; 5) Department of Medicine, Vanderbilt University, Nashville, TN; 6) Department of Biomedical Informatics, Columbia University, New York, NY; 7) Center for Applied Genomics, Children's Hospital of Philadelphia, PA; 8) Center for Human Genetics, Marshfield Clinic, Marshfield, WI; 9) Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN; 10) Kaiser Permanente Washington Health Research Institute, Seattle, WA; 11) Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA; 12) Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL; 13) Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Rochester, MN; 14) Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA; 15) Department of Biomedical Informatics in School of Medicine, Vanderbilt University, Nashville, TN; 16) Department of Biomedical Informatics, Columbia University, New York, NY

Cardiovascular diseases and central nervous system disorders (such as coronary artery disease, heart failure, Alzheimer’s disease and multiple sclerosis) have a significant impact on mortality rate worldwide and frequently co-occur in patients. Improving disease treatment would benefit from knowledge of the relationship between multiple diseases in these categories. The contribution of pleiotropy to these conditions remains largely unknown given that previous studies generally focused on single diseases for genetic variant discovery. Here we use Phenome-wide association studies (PheWAS) and multi-trait joint association studies to identify pleiotropic genetic variants. A comprehensive set of phenotypes was curated for eMERGE network and UK Biobank using International Classification of Diseases codes from the electronic health records. After quality control in the eMERGE network, we used 61 cardiovascular diseases and 28 central nervous system disorders in 43,015 European-ancestry adults and 7 million SNPs. We performed PheWAS and multi-trait joint analyses (via MultiPhen) for these 89 diseases using all SNPs. Both methods were able to identify previously known disease-associated SNPs, such as Alzheimer’s disease (rs429358) and coronary artery disease (rs1333049). Among 565 Bonferroni significant (α=0.05) SNPs identified by the PheWAS, 359 were Bonferroni significant in multi-trait joint analyses. We conducted replication studies using 294,753 European-ancestry adult samples from UK Biobank. 82 SNPs had a significant

association in both PheWAS and MultiPhen; these SNPs were mapped to *HLA, LPA, CDKN2B-AS1, NECTIN2, TOMM40, APOE* and *APOC1* regions. Since the rejection of the null hypothesis for multi-trait joint analysis fails to provide exact associated diseases, we further performed a formal statistical test for pleiotropy using sequential multivariate analyses. We characterized novel SNPs that showed significance across cardiovascular diseases and central nervous system disorders. For instance, rs56131196 (downstream *APOC1*), which has been previously shown to be associated with Alzheimer's disease, has implicated its potential pleiotropic influence on dementia, abdominal aortic aneurysm, and ischemic heart disease in eMERGE network and UK Biobank. Our research discovered novel pleiotropy across cardiovascular diseases and central nervous system disorders, which have the potential for assisting in disease prevention and drug repositioning in future research.

A- A+

PgmNr 67: Gene expression and genetic variation of ERG is associated with inflammation in endothelial cells and risk of coronary artery disease in humans.

Authors:
J. Gukasyan [1,2]; Q. Jia [1,2]; N. Woodward [1,2]; R. Zhu [1,2]; Y. Han [1,2]; L.K. Stolze [3]; Z. Kurt [4]; X. Yang [4]; C.E. Romanoski [3]; J.A. Hartiala [1,2]; H. Allayee [1,2]

View Session Add to Schedule

Affiliations:
1) Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA.; 2) Biochemistry & Molecular Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA.; 3) Department of Cellular and Molecular Medicine, University of Arizona College of Medicine, Tucson, AZ.; 4) Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA.

Recent genome-wide association studies (GWAS) support the premise that genetic factors acting at the level of the endothelium can play causal roles in the pathogenesis of coronary artery disease (CAD). The erythroblast transformation-specific related gene (ERG) represents one such candidate factor. ERG encodes a transcription factor highly expressed in endothelial cells, and has been implicated in promoting anti-inflammatory pathways, endothelial homeostasis, and angiogenesis, suggesting it would have a protective effect on the development of CAD. To our knowledge, there is no direct link between ERG and risk of CAD in humans. Given the importance of the endothelium in CAD pathogenesis, we hypothesized that ERG can play a protective role in atherogenesis. Human aortic endothelial cells (HAECs) were isolated from aortic explants of 149 heart transplant donors and incubated for 4hrs with or without 40mg/mL oxidized phospholipids (oxPL). SiRNA knockdown of ERG was carried out in HAECs from three independent donors for 4hrs. Transcript levels of candidate inflammatory genes were measured in both experiments. Genetic association of the ERG locus with risk of CAD was based on the results of a meta-analysis in >500,000 subjects from UK Biobank and CARDIoGRAM+C4D Consortium. ERG mRNA levels were significantly and consistently downregulated in HAECs after incubation with oxPL in subjects of various ethnicities (P=4.5E-19). Furthermore, *ERG* expression was inversely correlated with expression of pro-inflammatory genes, including *IL8* (P= 3.1E-09) and *IL6* (P= 2E-04), and positively correlated with athero-protective gene NOS3 (P=1.2E-03). These relationships were independent of oxPL treatment and observed in both sexes. Direct perturbation of *ERG* in HAECs with siRNA validated the inverse relationship between expression of *ERG* and the same inflammatory genes. Evaluation of the ERG locus on chromosome 21q22.2 revealed several SNPs in high linkage disequilibrium that were strongly associated with risk of CAD. The lead SNP, rs2836621, is located ~20kb upstream of ERG and yields an association p-value of 5.2E−06, which exceeds the Bonferroni-corrected threshold for testing 2713 SNPs in an 800kb interval (p=0.05/2713=1.8E-05). We provide functional *in vitro* and genetic evidence in humans that ERG could potentially influence the development of CAD through inflammatory mechanisms at the level of the vessel wall. Further support for this concept requires additional studies.

A- A+

PgmNr 68: Investigating the mechanisms underlying genetic risk for coronary artery disease utilizing endothelial cells.

Authors:
L. Stolze [1]; M. Whalen [2]; A. Conklin [2]; A. Solomon [2]; M. Kaikkonen-Määttä [3]; C. Romanoski [2]

View Session Add to Schedule

Affiliations:
1) Genetics Graduate Interdisciplinary Program, The University of Arizona, Tucson, AZ; 2) Dept of Cellular and Molecular Medicine, The University of Arizona, Tucson, AZ; 3) A.I.Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland

Coronary Artery Disease (CAD) has been the subject of numerous Genome-wide Association Studies (GWAS) with the goal of explaining the genetic component to disease presence and outcome. Approximately 160 genetic loci have been associated with CAD, however, the majority of these loci are in non-protein coding regions of the genome with unknown function. The Genotype Tissue Expression (GTEx) consortium has utilized hundreds of samples from 53 different tissues to annotate expression Quantitative Trait Loci (eQTLs) or variants that are associated with modulations in gene expression. This approach has proven invaluable; however, tissues are comprised of multiple cell-types and were all collected without any stimuli, meaning some environment-dependent eQTLs could be missed. To identify signals that are only resolvable in a singular cell type, we performed eQTL analysis on Human Aortic Endothelial Cells (HAECs) from a diverse population of 53 de-identified heart donors. Additionally, to begin to explore the effects of stimuli on eQTL effect or presence, we cultured these cells both with and without treatment by pro-inflammatory cytokine Interleukin 1 Beta. Through this process, we have discovered 69,501 variants with associations to gene expression, 15,755 of which are unique from GTEx. Similar to CAD associated variants, the majority of eQTLs discovered are intergenic suggesting function through regulatory elements. Therefore, we collected chromatin accessibility, transcription factor binding, and histone modification data from our HAEC population. Variants were discovered that are associated with changes in each data set resulting in 44,028 variants associated with a molecular trait. These molecular QTLs were enriched in low p-value eQTLs suggesting mechanisms through which expression is changing. We then compared our eQTLs and molecular QTLs to genetic variants with a genome-wide significant p-value from the latest CAD GWAS resulting 17 candidate SNPs that are associated with the expression of a gene, at least one molecular trait, and CAD. Through layering of QTL analyses in a single cell-type, we can mechanistically explain disease associated variants and further understand the underlying pathways of disease.

A- A+

PgmNr 69: Prioritization of genomic loci for coronary artery disease using targeted CRISPR screens for endothelial dysfunction.

Authors:
F. Wuennemann [1]; T. Fotsing Tadjo [1]; M. Beaudoin [1]; K.S. Lo [1]; G. Lettre [1,2]

View Session Add to Schedule

Affiliations:
1) Montreal Heart Institute, Montréal, Québec, Canada.; 2) Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada

Coronary artery disease (CAD) is one of the leading causes of mortality worldwide and remains a relevant population health problem, despite lipid-lowering treatments such as statins. The limited success of treating CAD by controlling classical risk factors such as LDL-cholesterol levels or high blood pressure highlights multifaceted pathophysiological mechanisms that contribute to disease. Furthermore, half of the identified risk loci from genome-wide association studies (GWAS) for CAD are not associated with traditional risk factors, pointing towards other, currently unknown factors that mediate risk. Endothelial cells form the inner lining of blood vessels such as the coronary arteries and they have important roles in the prevention of CAD such as selective barrier function, secretion of vasoactive molecules, response to hemodynamic forces and inflammation and regulation of homeostasis and thrombosis among others. Despite their critical role in CAD biology, little is known about the implication of endothelial dysfunction in CAD development. Here, we tested previously identified GWAS loci associated with CAD using targeted CRISPR perturbation screens (CRISPR-Cas9, CRISPRa, CRISPRi) in endothelial cells (immortalized human aortic endothelial cells (teloHAEC)). We targeted 93 previously identified GWAS SNPs for CAD, identified all proxies in LD with the original sentinel variants (r^2 > 0.8 in European-ancestry populations) and designed sgRNA in a 100bp window around all SNPs. Following infection, antibiotic selection and induction of an inflammatory response using TNF-a, cells were sorted into top and bottom 10% based on cell-cell adhesion markers (SELE, PECAM1, ICAM1, VCAM1) as a readout of endothelial activation. We identified 17 CAD loci that significantly impacted adhesion molecule presentation at the cell surface, suggesting endothelial dysfunction as the mechanism underlying the GWAS association signal. Our results highlight the potential to utilize human genetic evidence to guide the design of CRISPR perturbation screens in an effort to characterize the underlying pathophysiology of common diseases such as CAD.

A- A+

PgmNr 70: Ceramide synthase *TLCD3B* as a novel gene associated with human recessive cone-rod dystrophy.

Authors:
R.E. Bertrand [1,2]; K.H. Xiong [3,2]; C. Thangavel [3,2]; J. Wang [3]; R. Ba-Abbad [4]; R.T. SimÕes [5]; K.J. Carss [6,7]; F.L. Raymond [7,8]; K. Wang [2]; Y. Li [2]; F.B.O. Porto [9]; A.R. Webster [10]; G. Arno [10]; R. Chen [2,11]

View Session Add to Schedule

Affiliations:
1) Department of Biochemistry, Baylor College of Medicine, Houston, TX.; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3) Department of BioSciences, Rice University, Houston, TX; 4) Genetics Department, Moorfield Eye Hospital, London, UK; 5) Insituto de Ensino e Pesquisa, Santa Casa de Minas Gerais, Brazil; 6) Department of Haematology, University of Cambridge, Cambridge, UK; 7) NIHR BioResource-Rare Diseases, Cambridge University Hospitals, Cambridge, UK; 8) Department of Medical Genetics, University of Cambridge, Cambridge, UK; 9) INRET Clínica e Centro de Pesquisa, Belo Horizonte, Minas Gerais, Brazil; 10) UCL Institute of Ophthalmology, London, UK; 11) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

Inherited retinal dystrophies (IRD) constitute a genetically heterogeneous group of disorders that lead to irreversible visual loss. Identification and studying the underlying molecular causes of IRD are important for patient diagnosis and therapy development. Whole exome or genome sequencing (WES/WGS) has identified homozygous mutations in *TLCD3B* in four patients with IRD from three unrelated families diagnosed with cone-rod dystrophy, providing the first link between *TLCD3B* and IRD. To further confirm the human genetics finding, *Tlcd3b* knockout mice were generated using CRISPR/Cas9. The phenotype of the mutant retina was characterized using Electroretinography (ERG), histology, and immunohistochemistry using individual retinal cell type markers. Consistent with the phenotype observed in patients, the *Tlcd3b*$^{KO/KO}$ mice exhibited photoreceptor dysfunction and degeneration by 7 months-of-age. Significant reduction of the cone photoreceptors' light responses, thinning of the outer nuclear layer (ONL), and loss of cone-photoreceptors across the retina were observed. These results indicate that the *Tlcd3b*$^{KO/KO}$ mice recapitulate the human phenotype and can serve as a good model to study the mechanism of degeneration and develop targeted therapies. Our finding provides the first link between mutations in a ceramide synthase gene and human retinal degeneration diseases. Previous studies in mice have shown a link between increased ceramide and retinal degeneration but previous ceramide synthase knockout models have had little to no retinal phenotype. The *Tlcd3b* knock out model is the first *in vivo* model that demonstrates ceramide is essential for survival and function of cone photoreceptor cells.

A- A+

PgmNr 71: *SSBP1* mutations cause a complex optic atrophy spectrum disorder with mitochondrial DNA depletion.

Authors:
F. Ullah [1]; D. Dotto [2]; I. Meo [3]; P. Magini [4]; M. Gusic [5]; A. Maresca [6]; L. Caporali [6]; F. Palombo [6]; F. Tagliavini [6]; E. Baugh [7]; B. Macao [8]; Z. Szilagyi [8]; C. Peron [3]; M. Gustafson [9]; C. Morgia [10]; P. Barboni [6]; M. Carbonelli [6]; M. Valentino [6]; R. Liguori [9]; V. Shashi [11]; J. Sullivan [11]; S. Nagaraj [12]; E. Davis [1]; M. Seri [3]; M. Falkenberg [8]; H. Prokisch [5]; N. Katsanis [1]; V. Tiranti [3]; T. Pippucci [4]; V. Carelli [4]

View Session Add to Schedule

Affiliations:
1) Center for human disease modeling, Duke university, Durham, NC.; 2) Unit of Neurology, Department of Biomedical and NeuroMotor Sciences, University of Bologna, Italy.; 3) Unit of Medical Genetics and Neurogenetics, Fondazione IRCCS Istituto Neurologico C. Besta, Milan, Italy.; 4) Medical Genetics Unit, Sant’Orsola-Malpighi University Hospital, Bologna, Italy.; 5) Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany.; 6) IRCCS Istituto delle Scienze Neurologiche di Bologna, UOC Clinica Neurologica, Bologna, Italy.; 7) Institute for Genomic Medicine, Columbia University, New York, NY.; 8) Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, University of Gothenburg, Gothenburg, Sweden.; 9) Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709 ?USA.; 10) Department of Ophthalmology, Studio Oculistico dAzeglio, Bologna, Italy.; 11) Division of Medical Genetics, Department of Pediatrics, Duke University School of Medicine, Durham, NC, USA.; 12) Division of Nephrology, Department of Pediatrics, Duke University School of Medicine, Durham, NC, USA.

Inherited optic neuropathies include complex phenotypes, mostly driven by mitochondrial dysfunction. Here, we report an optic atrophy spectrum disorder in adults, characterized by retinal macular dystrophy and kidney insufficiency leading to transplantation. These features are associated with mitochondrial DNA (mtDNA) depletion without accumulation of multiple deletions. Using whole-exome sequencing in four families with dominant and one with recessive inheritance, we identified mutations affecting the mitochondrial single strand binding protein (SSBP1). We show that *SSBP1* mutations in patient-derived fibroblasts variably affect its protein amount and alter multimer formation, but not the binding to ssDNA. *SSBP1* mutations impaired mtDNA; nucleoids and 7S-DNA amounts; as well as mtDNA replication. The variable degree of mtDNA depletion in cells is consistent with severity of mitochondrial dysfunction, including respiratory efficiency, OXPHOS subunits and complex amount and assembly. Additionally, mtDNA depletion and cytochrome c oxidase-negative fibers were found in biopsies of kidney and skeletal muscle tissues from affected individuals. Furthermore, transient suppression of *ssbp1* in zebrafish resulted in a significantly reduced optic nerve size at 2 days post fertilization (dpf). Using this phenotype readout, we performed *in vivo* complementation studies to show that all missense variants identified in affected individuals confer a reduction of *SSBP1* function, suggesting haploinsufficiency in humans. Notably, optic nerve defects were undetectable in *ssbp1* F0 mutants at 2 dpf; however, we observed growth restriction and lethality in F0s at 3 weeks post fertilization, with a concomitant reduction of mitochondrial number, supporting further the mitochondrial underpinnings of disease. This previously unrecognized disorder of mtDNA maintenance implicates *SSBP1* mutations as novel cause of human pathology.

A- A+

PgmNr 72: Identification of *de novo* missense variants in *WDR37* as a novel cause of human disease: Delineation of the phenotype and functional studies.

Authors:
L. Reis [1,6]; E. Sorokina [1,6]; S. Thompson [1]; S. Muheisen [1]; M. Velinov [2]; C. Zamora [3]; A. Aylsworth [4]; E. Semina [1,5]

View Session Add to Schedule

Affiliations:
1) Department of Pediatrics, Medical College of Wisconsin, Milwaukee, WI.; 2) Department of Human Genetics; New York State Institute for Basic Research in Developmental Disabilities, 1050 Forest Hill Road, Staten Island, NY 10314; 3) Department of Radiology, Division of Neuroradiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; 4) Departments of Pediatrics and Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; 5) Departments of Ophthalmology and Cell Biology, Neurobiology and Anatomy, Medical College of Wisconsin, Milwaukee, WI 53226; 6) These authors contributed equally to this work

---

Exome sequencing is a powerful tool for new gene discovery in human syndromes, particularly when combined with functional studies in animal models. Using exome sequencing in individuals with undiagnosed ocular syndromes and matchmaker database analysis, we have identified *de novo* missense variants predicted to affect the N-terminal region of WDR37 (c.356C>T p.Ser119Phe, c.389C>T p.Thr130Ile, c.374 C>T p.Thr125Ile, and c.386C>G p.Ser129Cys) in four unrelated individuals. The clinical features of those affected were similar, characterized by ocular anomalies (corneal opacity/Peters anomaly, coloboma or microcornea), dysmorphic facial features (microcephaly, thin upper lip, broad nasal bridge, and ear anomalies), significant neurological impairment with structural brain defects and seizures, poor feeding, poor post-natal growth, and variable skeletal, cardiac, and genitourinary defects. One patient died in early childhood. WDR37 is a member of the WD40 repeat domain family and its specific functions are currently unknown. Immunocytochemistry and Western blot studies showed similar protein levels and localization in the cytoplasm for both wild-type and mutant WDR37. CRISPR-Cas9 mediated genome editing generated zebrafish mutants with missense and frameshift alleles, p.Ser129Phe, p.Ser129Cys (which replicates one of the human variants), p.Ser129Tyr, p.Lys127Cysfs, and p.Gln95Argfs. Heterozygous missense variants were associated with poor growth and larval lethality in zebrafish, while heterozygotes with frameshift alleles survived to adulthood and were bred to generate a complete loss-of-function model for phenotypic characterization. RNAseq analysis on zebrafish embryos carrying the p.Ser129Phe missense variant identified significant upregulation of cholesterol biosynthesis pathways. Yeast two-hybrid screening using both wild-type and mutant proteins was undertaken to identify protein interactions that are disrupted by the missense variants. This study identifies *WDR37* as an important factor in vertebrate development and provides insight into its molecular functions.

A- A+

PgmNr 73: *De novo* variants in *WDR37* are associated with epilepsy, colobomas, dysmorphism, developmental delay, intellectual disability, and cerebellar hypoplasia.

Authors:
O. Kanca [1]; J.C. Andrews [1]; P.T. Lee [1]; C. Patel [2]; S. Braddock [3,4]; A.M. Slavotinek [5]; J.S. Cohen [6]; C.S. Gubbels [7]; K.A. Aldinger [8]; J. Williams [9]; M. Indaram [10]; A. Fatemi [6]; T.W. Yu [7]; P.B. Agrawal [11]; G. Vezina [12]; B. Gangaram [5]; J. Wynn [13]; R. Hernan [13]; G. Mychaliska [22]; W.K. Chung [13]; T.C. Markello [14,15]; W.B. Dobyns [8,16,17]; D.R. Adams [14,15]; W.A. Gahl [14,15]; M.F. Wangler [1,18,19]; S. Yamamoto [1,18,19,20]; H.J. Bellen [1,18,19,20,21]; M.C.V. Malicdan [14,15]; Undiagnosed Disease Network

View Session Add to Schedule

Affiliations:
1) Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas.; 2) Genetic Health Queensland, Royal Brisbane and Women’s Hospital, Brisbane, QLD 4029, Australia; 3) Department of Pediatrics, Cardinal Glennon Children's Medical Center, St. Louis, MO, 63104, USA; 4) Division of Pediatric Medical Genetics, Saint Louis University Hospital, St. Louis, MO, 63104, USA; 5) Department of Pediatrics, University of California, San Francisco, CA, 94143-2711, USA; 6) Division of Neurogenetics and Hugo W. Moser Research Institute, Kennedy Krieger Institute, Baltimore, MD, USA; 7) Division of Genetics and Genomics, Boston Children's Hospital/Harvard Medical School/Broad Institute of Massachusetts Institute of Technology and Harvard, Boston, Massachusetts; 8) Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, 98101, USA; 9) Paediatric Department, Bundaberg Hospital, Bundaberg, QLD 4670, Queensland, Australia; 10) Department of Ophthalmology, University of California, San Francisco, CA, 94143-2711, USA; 11) Division of Newborn Medicine and Genetics and Genomics, Manton Center for Orphan Disease Research, Harvard Medical School, Boston MA 02115; 12) Division of Diagnostic Imaging & Radiology, Children’s National Health System, 111 Michigan Ave. NW, Washington, DC 20010; 13) Department of Pediatrics and Medicine, Columbia University, New York, NY, 10032, USA; 14) Office of the Clinical Director, National Human Genome Research Institute, NIH, Bethesda, MD, 20892-1851, USA; 15) NIH Undiagnosed Diseases Program, Common Fund, Office of the Director, NIH, Bethesda, MD, 20892, USA; 16) Department of Pediatrics (Genetics), University of Washington, Seattle, WA, 98195, USA; 17) Department of Neurology, University of Washington, Seattle, WA, 98195, USA; 18) Program in Developmental Biology, Baylor College of Medicine, Houston, TX, 77030, USA; 19) Jan and Dan Duncan Neurological Research Institute, Texas Children’s Hospital, Houston, TX, 77030, USA; 20) Department of Neuroscience, Baylor College of Medicine, Houston, TX, 77030, USA; 21) Howard Hughes Medical Institute, Baylor College of Medicine, Houston, TX, 77030, USA; 22) Section of Pediatric Surgery, University of Michigan, Ann Arbor, MI, 48109

WDR40 repeats (WDR) are among the most abundant protein domains and protein- protein interaction motifs in eukaryotes. Currently there are 60 WDR containing proteins associated with diseases reported in OMIM. In this study, we have identified 5 probands that have missense *de novo* variants in an evolutionarily well conserved WDR protein, WDR37. All of the probands exhibit developmental delay, intellectual disability, epilepsy, colobomas, CHARGE syndrome-like facial dysmorphism and cerebellar hypoplasia phenotypes.

Despite the high level of conservation throughout the entire protein across species, there is very little knowledge about the function of *WDR37* and its orthologs. In order to gain more insight in WDR37 function we identified the orthologous gene in *Drosophila melanogaster, CG12333* (hereafter named *wdr37*). We generated null alleles for this previously unstudied gene by replacing the coding region of the gene with transcriptional activator GAL4. *wdr37* homozygous mutant animals are viable and fertile, without overt morphological phenotypes. Careful analysis of behavioral phenotypes of *wdr37* mutants identified that they exhibit bang sensitivity, which is a phenotype associated physiologically and pharmacologically with epilepsy. Additionally, the flies lost their grip during climbing to the walls of vials and during copulation. All the phenotypes associated with *wdr37* loss can be rescued by expressing the human reference cDNA, indicating that the gene is functionally conserved between flies and humans. Interestingly the variant cDNAs could not rescue the phenotypes, suggesting that the variants impair the function of the gene.

In summary, we identified five probands that have *de novo* missense variants that map to a small region of *WDR37* and who exhibit overlapping neurological phenotypes. The analysis of the gene in fly model suggests that the gene is functionally conserved across evolution and the variants impair the function of the gene.

A- A+

PgmNr 74: Biallelic VPS35L pathogenic variants cause 3C/Ritscher-Schinzel-like syndrome through dysfunction of retriever complex.

Authors:
K. Kato [1,2,3]; Y. Oka [4]; H. Muramatsu [2]; F. Vasilev [5]; T. Otomo [5]; H. Oishi [6]; Y. Kawano [3]; Y. Nakazawa [4]; T. Ogi [4]; Y. Takahashi [2]; S. Saitoh [1]

View Session Add to Schedule

Affiliations:
1) Department of Pediatrics and Neonatology, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan; 2) Department of Pediatrics, Nagoya University Graduate School of Medicine, Nagoya, Japan; 3) Department of Pediatrics, Toyota Memorial Hospital, Toyota, Japan; 4) Department of Genetics, Research Institute of Environmental Medicine, Nagoya University, Nagoya, Japan; 5) Department of Pathophysiology and Metabolism, Kawasaki Medical School; 6) Department of Comparative and Experimental Medicine, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan

---

Background: 3C/Ritscher-Schinzel-syndrome is characterized by congenital cranio-cerebello-cardiac dysplasia, where *CCDC22* and *WASHC5* are accepted as the causative genes. In combination with the retromer or retriever complex, these genes play a role in endosomal membrane protein recycling. We aimed to identify the gene abnormality responsible for the pathogenicity in siblings with a 3C/Ritscher-Schinzel-like syndrome, displaying cranio-cerebello-cardiac dysplasia, coloboma, microphthalmia, chondrodysplasia punctata, periventricular nodular heterotopia, proteinuria, and complicated skeletal malformation.

Methods: Exome sequencing was performed to identify pathogenic variants. According to the pedigree, we focused on recessively inherited traits. Cellular biological analyses and generation of knockout mice were carried out to elucidate the gene function and pathophysiological significance of the identified variants.

Results: We identified compound heterozygous pathogenic variants (NM_020314.5: c.1097dup; p.Cys366Trpfs*28 and c.2755G>A; p.Ala919Thr) in the *VPS35L* gene, which encodes a core protein of the retriever complex. Immunoprecipitation and immunoblot analyses indicated that the endogenous VPS29 protein, which is another member of retriever complex, was not co-immunoprecipitated with the VPS35L-A919T mutant while it was co-immunoprecipitated with the wild-type VPS35L. Furthermore, the frameshift variant induced nonsense-mediated mRNA decay, thereby confirming biallelic loss of function of VPS35L. In addition, *VPS35L* knockout cells showed decreased autophagic function in nutrient-rich and starvation conditions, as well as following treatment with Torin 1. We also generated *Vps35l*-deficient mice by CRISPR/Cas9 system. Heterozygous mutant mice appeared normal and fertile. However, homozygous mutant mice ($Vps35l^{-/-}$) were embryonic lethal at an early developmental stage between E7.5 and E10.5.

Conclusions: Our results suggest that biallelic loss-of-function variants in *VPS35L* underlies 3C/Ritscher-Schinzel-like syndrome. Furthermore, VPS35L is necessary for autophagic function, and essential for early embryonic development. The data presented here provide a new insight into the critical role of the retriever complex in fetal development.

A- A+

PgmNr 75: Identification of 46 novel recessive candidate genes for intellectual disability and visual impairment in 350 consanguineous families.

Authors:
S.E. Antonarakis [1,2,3]; S.A. Paracha [4]; S. Imtiaz [5]; A. Nazir [4]; Y.M. Waryah [6]; P. Makrythanasis [1,7]; S. Qureshi [4]; S. Saeed [5]; J. Khan [4]; E. Falconnet [1]; M. Guipponi [2]; C. Borel [1]; M.A. Ansari [5]; Z. Iqbal [8]; E. Frengen [9]; E. Ranza [1,2,10]; F.A. Santoni [1,11]; K. Gul [5,12]; J. Ahmed [4]; M.T. Sarwar [4]; A.M. Waryah [6]; M. Ansar [1]

View Session Add to Schedule

Affiliations:
1) Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland; 2) Service of Genetic Medicine, University Hospitals of Geneva, Geneva, Switzerland; 3) iGE3 Institute of Genetics and Genomics of Geneva, Geneva, Switzerland; 4) Institute of Basic Medical Sciences, Khyber Medical University, Peshawar, Pakistan; 5) Department of Genetics, University of Karachi, Karachi, Pakistan; 6) Molecular Biology and Genetics Department, Medical Research Center, Liaquat University of Medical and Health Sciences, Jamshoro, Pakistan; 7) Current address, Biomedical Research Foundation of the Academy of Athens, Athens, Greece; 8) Department of Neurology, Oslo University Hospital, Oslo, Norway; 9) Department of Medical Genetics, Oslo University Hospital and University of Oslo, Oslo, Norway; 10) Current address, Medigenome, Swiss Institute of Genomic Medicine, Geneva, Switzerland; 11) Current address, Department of Endocrinology Diabetes and Metabolism, University Hospital of Lausanne, Switzerland; 12) Department of Biosciences, Faculty of Life Sciences, Mohammad Ali Jinnah University, Karachi, Pakistan

---

Consanguinity, practiced in a substantial fraction of human populations, reveals numerous rare recessive phenotypes because of the extensive regions of homozygosity by decent. The average genomic homozygosity is 253Mb in the offspring of consanguineous parents; this is 10fold higher than that of 25Mb in outbred individuals, and this provides the opportunity to identify the high impact genomic variants that cause autosomal recessive phenotypes. In Pakistan, the frequency of consanguineous marriages approaches 70%. To bridge the gap between the 1800 known and the estimated >9000 recessive gene-phenotypes, we have initiated a Swiss-Pakistani project to identify novel recessive candidate genes for two phenotypes: Intellectual Disability (ID) and Visual Impairment (VI). We have collected samples from 1265 individuals of 197 ID, and 1809 individual of 236 VI families of first cousin marriages with at least two affecteds. Exome sequence of one affected and genotyping of the whole family (parents, all affected and unaffected siblings) has been completed in 166 ID and 184 VI families to date. The likely causative gene/variant in known genes was found in 64% of the VI and 32% in the ID families. Thus, there are more unknown “recessive genes” for ID. In 18% of the VI families we have identified 25 novel candidate genes, and in 27% of the ID families 24 such candidates (including *CDC25B*, *PSMB1*, *IQSEC1*, *SLC6A6*; additional candidates will be presented in the conference). International genematching identified additional families in 20% of the candidate genes. Through this study, we already have contributed some novel genes for visual impairments, *MARK3* (PMID: 29771303), *DNMBP* (PMID: 30290152) and for ID, *LINGO1* (PMID: 28837161), *FBXL3* (PMID: 30481285), *KALRN* (PMID: 27421267) and *DYNC1I2* (PMID: 31079899) in multiple patients/families along with evidence by using animal models like drosophila and zebrafish.

Careful evaluation of the phenotypes is mandatory to assess the possibility of two or more causative genes in certain families, and to minimize false negative results. International databases from consanguineous individuals are needed to facilitate the assignment of pathogenicity to homozygous variants.

A- A+

PgmNr 76: Analyses of 1,268 breakpoints from balanced chromosomal abnormalities reveals patterns of three-dimensional reorganization associated with human developmental anomalies.

Authors:
C. Lowther [1,2,3,25]; M.M. Mehrjouy [5,25]; M. Bak [5,25]; R.L. Collins [1,2,3,5]; H. Brand [1,2,3]; B.B. Currall [1,2]; K. O'Keefe [1,2]; Z. Dong [6]; K.W. Choy [6]; E.S. Wilch [7]; O.A. Clark [7]; T. Kammin [7]; S.L.P. Schilit [7,8]; M.B. Rasmussen [5]; A.S.C. Fonseca [9]; T. Varilo [10]; J.F. Mazzeu [11]; K.T. Abe [12]; A. Lindstrand [13]; C. Sismani [14]; K. Õunap [15]; C. Schluth-Bolard [16]; S. Temel [17]; E.C. Liao [18,19,20]; C.C. Morton [3,7,21,22]; J.F. Gusella [1,2,3,23]; P. Jacky [24]; I. Bache [5,26]; N. Tommerup [5,26]; M.E. Talkowski [1,2,3,26]; Developmental Genome Anatomy Project (DGAP), International Breakpoint Mapping Consortium (IBMC)

View Session Add to Schedule

Affiliations:
1) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; 2) Department of Neurology, Harvard Medical School, Boston, MA, USA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Boston, MA, USA; 4) Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences, Harvard Medical School, Boston, MA, USA; 5) Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark; 6) Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, China; 7) Departments of Obstetrics and Gynecology and of Pathology, Brigham and Women’s Hospital, Boston, MA, USA; 8) PhD Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA; 9) Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, Brazil; 10) Department of Medical Genetics, University of Helsinki, Helsinki, Finland; 11) Faculdade de Medicina, Universidade de Brasília, Brasilsi, Brazil; 12) SARAH Network of Rehabilitation Hospitals, Brasília, Brazil; 13) Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; 14) Department of Cytogenetics and Genomics, Cyprus School of Molecular Medicine, Nicosia, Cyprus; 15) Department of Clinical Genetics, University of Tartu, Tartu, Estonia; 16) Service de Génétique, Hospices Civils de Lyon, Bron, France; 17) Department of Medical Genetics, Uludag University, Bursa, Turkey; 18) Center for Regenerative Medicine, Massachusetts General Hospital and Harvard Medical School, MA, USA; 19) Division of Plastic and Reconstructive Surgery, Massachusetts General Hospital, Boston, MA, USA; 20) Harvard Stem Cell Institute, Cambridge, MA, USA; 21) University of Manchester, Manchester Academic Health Science Center, Manchester, UK; 22) Harvard Medical School, Boston, MA, USA; 23) Molecular Neurogenetics Unit, Center for Human Genetic Research, Department of Neurology, Massachusetts General Hospital, Boston, MA, USA; 24) NW Permanente, PC, Emeritus, Portland, OR; 25) Equally contributing authors; 26) Co-senior authors

---

Balanced chromosomal abnormalities (BCAs) represent a significant risk factor for developmental disorders, but direct disruption of a known disease gene only explains a small fraction of cases. To identify novel mechanisms by which BCAs confer risk for human congenital anomalies, we analyzed 1,268 breakpoints from sequence-resolved simple BCAs in 634 individuals, including cases with congenital anomalies (n=387) and the first large-scale sequencing of BCAs in controls without discernible disease phenotypes (n=247). After categorizing all protein-coding genes into five tiers

based on evidence for association with disease (e.g., Tier 1 = known dominant disease gene; Tier 5 = no association), we find that ~13.2% of cases harbored a BCA disrupting a Tier 1 gene compared to just 2.0% of controls (OR=7.1, p=2.26E10^{-7}). To explore novel non-coding mechanisms, we removed cases and controls with a direct disruption of a Tier 1 gene from the remaining analyses. We found no significant difference in the distance between BCA breakpoints and genes between cases and controls, indicating that proximity to genes alone was not sufficient to distinguish positional effects in cases. However, when testing for enrichment of breakpoints within non-overlapping 1 Mb bins we found three genome-wide significant loci harboring a significant accumulation in cases, none of which included control breakpoints, suggesting an etiological basis for the association rather than a mechanistic hotspot for BCAs. Given that genomes are organized into hierarchical structures, we hypothesized that disruption of topological associating domains (TADs), which are critical for cis-regulatory interactions, could act as a source of non-coding risk. Indeed, layering three-dimensional regulatory structures onto our breakpoint atlas revealed seven genome-wide significant TADs in cases and none in controls. These genome-wide significant TADs included three established long-range positional effect (*MEF2C*, *FOXP1*, and *SOX9*) and four novel (2q14.3, 3p21.1-p14.3, 8q24.3, and 17p11.2) loci. Ongoing CRISPR/Cas9 functional dissection of the most significant TAD regions suggests that alteration to these 3D structures result in diverse and unpredictable chromatin contact changes and gene expression patterns. Collectively, these data suggest a rich landscape of complex pathogenic mechanisms associated with BCAs in human developmental disorders.

A- A+

PgmNr 77: Long read single molecule sequencing identifies putative fetal hemoglobin modifier loci in Africans with sickle cell anemia.

Authors:
P. Lurie [1]; B. Tayo [2]; G. Ayodo [4]; S. Obaro [5]; T. Akingbola [3]; N. Hall [8]; R. Harris [1,7]; B. Henn [9]; S. Gopalan [10]; Q. Meng [7]; S. Jhangiani [7]; R. Cooper [2]; G. Lettre [6]; K. Worley [1,7]; A. Wonkman [12]; F. Sedlazeck [7]; N. Hanchard [1,11]

View Session Add to Schedule

Affiliations:
1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Department of Public Health Sciences, Loyola University Chicago Stritch School of Medicine, Maywood, IL.; 3) Department of Hematology, University of Ibadan, College of Medicine, University College Hospital, Ibadan, Nigeria.; 4) School of Health Sciences, Jaramogi Oginga Odinga University of Science and Technology, Bondo Kenya; 5) Department of Pediatrics, University of Nebraska Medical Center, Omaha, NE; 6) Montreal Heart Institute and Department of Medicine, Faculty of Medicine, Universite de Montreal, Montreal, Quebec, Canada; 7) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 8) Pediatrics-Nutrition, Baylor College of Medicine, Houston, TX; 9) Department of Anthropology, University of California Davis, Davis, CA; 10) Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY; 11) USDA/ARS Children's Nutrition Research Center, Baylor College of Medicine, Houston, TX; 12) Division of Human Genetics, University of Cape Town, Cape Town, South Africa

INTRODUCTION: Sickle cell Anemia (SCA) - one of the most common single-gene Mendelian disorders in the world – is caused by recessive inheritance of a single missense mutation in the beta globin gene (*HBB*). *HBB* is located within the 60kb beta-globin cluster on chromosome 11, along with other developmentally-expressed beta-globin genes, including the fetal beta globin gene (*HBF*). Continued expression of *HBF* positively modifies SCA and has been associated with cis-acting factors in the beta-globin cluster; however, the genomic complexity of the region has limited identification of specific contributing loci. This information is particularly lacking for African populations, where the extensive genetic diversity is poorly represented in current human genome references.

METHODS: To better define the region and identify putative candidate modifiers of *HBF* in SCA, we used long-range PCR to amplify ~10kb windows across the beta-globin cluster in 40 individuals with SCA from Nigeria, Cameroon, and Kenya. Amplicons were subsequently subject to long read single-molecule sequencing using the PacBio I sequel system. Resulting reads were aligned to hg38 and Sniffles software was used to call structural variants (SVs).

RESULTS: In these African SCA individuals we identified 72 structural rearrangement events across the beta globin cluster, totaling 81kb, including 227 insertions, 15 deletions, and 76 inversions. 99% of these SVs are not described in the Database of Genomic Variants or the genome aggregation database (gnomAD), and each sample had, on average, 8 SVs. We observed nine unique SVs, ranging in size from 31bp to 1,662bp, that overlapped the same genomic locus and occurred in >10 samples.

These SVs were predicted to disrupt putative ENCODE transcription factor binding sites for transcription factors CEBPB, GATA1, and SMARCA4, which have all been previously shown to regulate fetal hemoglobin production *in cis*.

CONCLUSION: We have identified common, novel SVs in the beta-globin gene cluster of African SCA individuals that are putative candidates for native cis-acting regulators of *HbF*. Additional studies to validate these SVs in our expanded African SCA cohorts and demonstrate their functional impact on fetal hemoglobin are ongoing. The current results demonstrate the power of using long-read single molecule sequencing in diverse populations to uncover hidden genomic sequence complexity, as well as identify viable candidate disease modifiers.

A- A+

PgmNr 78: DeBreak: Deciphering the exact breakpoints of structural variants using long sequencing reads.

Authors:
Z. Chong; Y. Chen

View Session Add to Schedule

Affiliation: University of Alabama at Birmingham, Birmingham, Alabama.

Structure Variants (SVs) are a common and important type of genetic variations and are usually disease-causing mutations. The Third-Generation-Sequencing (TGS) techniques, such as PacBio and Oxford Nanopore, generate real-time long single-molecular reads and thus have great potential in accurately detecting SVs. There are only a few available tools and their performance is far from optimal. Here, we developed DeBreak which can comprehensively characterize all forms of SVs using TGS data.

DeBreak treats short and long events separately as they often have different features. After mapping raw reads to the reference genome with customized minimap2, DeBreak detects short SVs from CIGAR strings and longer SVs according to the orientation of two segments of the same read. For larger insertions that cannot be spanned in a single read, DeBreak detects insertion breakpoint candidates and performs local de novo assembly to generate longer and more accurate contigs to rescue them. For each detected SV event, DeBreak applies partial order alignment near its breakpoint in order to resolve the exact breakpoint position.

To benchmark the performance of DeBreak, we generated a 50X PacBio *in silico* dataset by spiking in both homozygous and heterozygous events and compared it with Sniffles and pbsv. DeBreak achieved an overall sensitivity and precision of 96.9% and 99.8% for the total simulated events, which is 5.3% higher than alternative tools in terms of balanced accuracy. In Oxford Nanopore *in silico* dataset, DeBreak also exceeded the other SV callers.

When applied on a real dataset, DeBreak consistently performed better than Sniffles and pbsv. In the sample HG002 from Genome In a Bottle (GIAB), DeBreak reached a sensitivity of 93.2% in deletion detection and 90.7% for insertion detection, while the best performance of the other tools only achieved 89.5% and 80.6%, respectively. After applying its large-insertion-rescue module, DeBreak rescued 205 missed insertion events and thus further increased its sensitivity by 3.4%. In addition, DeBreak achieved a genotyping accuracy of 91.8%, which is 15.5% and 46.3% higher than pbsv and Sniffles, respectively. DeBreak also performed well on the PacBio sequencing reads of the SKBR3 cell line, which means it can also be applied on cancer genomes. The higher balanced accuracy of DeBreak will be beneficial to the community for discovering both germline and disease-associated SVs.

Availability: https://github.com/Maggi-Chen/DeBreak.

A- A+

PgmNr 79: Application of long read genome sequencing to patients with undiagnosed diseases.

Authors:
C.M. Reuter [1,2]; A.N. Raja [2]; D.B. Zastrow [1,2]; R. Ungar [3]; J.N. Kohler [1,2]; D.E. Bonner [1,2]; S. Sutton [2]; L. Fernandez [1,2]; M. Majcherska [1,2]; C. McCormack [1,2]; S. Marwaha [1,2]; S. Utiramerur [4]; N.I.H. Undiagnosed Diseases Network [5]; P.G. Fisher [1,6]; J.A. Bernstein [1,7]; E.A. Ashley [1,2,3]; M.T. Wheeler [1,2]

View Session Add to Schedule

Affiliations:
1) Stanford Center for Undiagnosed Diseases, Stanford, CA; 2) Division of Cardiovascular Medicine, School of Medicine, Stanford, CA; 3) Department of Genetics, Stanford School of Medicine, Stanford, CA; 4) Clinical Genomics Program, School of Medicine, Stanford, CA; 5) NIH Undiagnosed Diseases Network, Office of the Director and the National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 6) Department of Neurology, School of Medicine, Stanford, CA; 7) Department of Pediatrics - Medical Genetics, School of Medicine, Stanford, CA

Long read (LR) genome sequencing has the potential to identify structural variants (SVs) such as deletions, insertions and inversions not detectable on short read (SR) sequencing, enabled by reads which may span many thousands of base pairs. We explored the added value of LR sequencing in an undiagnosed pediatric patient with microcephaly, developmental delay, Dandy-Walker variant, Raynauds, Horner syndrome, seizures, and dysmorphic features. Previous SR exome and genome sequencing were unrevealing. We performed proband-only LR genome sequencing with Oxford Nanopore Technologies PromethION platform. We used Minimap2 for alignment (Li et. al., 2018) and Sniffles for SV calling (Sedlazeck et. al., 2018). We compared SVs from LR with SVs from SR (GATK, CNVkit). SV calls unique to the LR sequencing platform were annotated using AnnotSV (Geoffroy et al 2018). We selected a subset of SVs to further curate based on clinical relevance to the patient. Curation included visualization in Integrative Genomics Viewer (IGV) to investigate possible sequencing artifacts. A total of 25665 SVs were identified with the LR SV calling algorithm (15576 deletions, 5 duplications, 9057 insertions, 499 inversion, 528 translocations, 34 inverted duplications). Of these, 25.7% (6442/25665) were unique to the LR data. Of LR unique calls, 48 were considered pathogenic (Class 5) and 3209 were considered likely pathogenic (Class 4) by AnnotSV. Twenty-six Class 5 SVs were further curated based on overlap with a gene. Screening for sequencing artifacts or false-positive LR SV calls with IGV supported accurate calls in 5/26 SVs. The remainder mapped to a region of the genome with an unannotated reference sequence (8/26 SVs), encompassed a repeat region (1/26 SVs), or were questionable by eye in IGV (12/26 SVs). One hundred and eleven Class 4 SVs that overlapped an exonic region of a gene were further curated. IGV supported accurate calls in 35/112 SVs. The remainder encompassed a repeat region (25/112 SVs) or could not be visualized in IGV (52/112 SVs). Using parental SR genome sequencing data to infer inheritance, 6 heterozygous candidate SVs were flagged for additional clinical correlation with the patient. Our data demonstrate that LR genome sequencing enables detection of unique SV calls not readily identified in SR data, which may have clinical relevance for patients with undiagnosed diseases. Future work will streamline identification and prioritization of true positive SVs.

A- A+

PgmNr 80: Integration of optical genome mapping and sequencing technologies for identification of structural variants in disorders/differences of sex development (DSD).

Authors:
E. Vilain [1, 2]; S. Bhattacharya [1]; M. Almalvez [1]; E.C. Delot [1, 2]; H. Barseghyan [1, 2]

View Session Add to Schedule

Affiliations:
1) Center for Genetic Medicine Research, Children's Research Institute, Children's National Medical Center, Washington, DC 20010, USA; 2) Genomics and Precision Medicine, School of Medicine and Health Sciences, George Washington University Washington, DC 20052, USA

Genomic technologies such as chromosomal microarrays, exome sequencing, and DSD-targeted panels have helped increase diagnostic rates. Despite this success, about half of the patients with DSD still remain without a firm diagnosis. Short-read sequencing (SRS) has the capability of reliably identifying single nucleotide variants (SNVs), as well as small insertions and deletions (INDELs). However, due to its methodological limitations (*i.e.* utilization of short reads), it fails to sensitively identify large structural variants (SVs), such as deletions, insertions, inversions, translocations and large copy number variants. While SVs currently represent a rare cause of DSD, they could be under-recognized due to the lack of available diagnostic technology.

To overcome these limitations, we used novel optical genome mapping (OM) technology, combined with whole-genome SRS. OM captures a pattern of fluorescent labels in long DNA molecules (>150 kbp) in nanochannel arrays for *de novo* genome assembly and SV calling. We have demonstrated the utility of OM for clinical diagnosis: in a series of patients diagnosed with Duchenne muscular dystrophy with large SVs involving the *DMD* gene, we successfully identified all major SV types, ranging from 13 kbp to 5.1 Mbp. We have now applied OM and WGS to 13 undiagnosed DSD cases. On average, we identified ~2000 insertions, ~1000 deletions, ~250 inversions and 0-1 translocations per individual, showing that, while less common than single nucleotide variants, SVs account for a substantial fraction of genetic variation. The Database of Genomic Variants, internal dataset frequency, and data from 200 healthy individuals were used to filter out common variants. Each proband carried 2-5 rare *de novo* SVs.

One of the defining features of OM is its capability to generate *de novo* genome assembly with contiguous maps spanning entire chromosome arms. However, SV breakpoint resolution is limited by the intervals between fluorescent dye labeling motifs with error rates ranging from 3-5 kbp. Using the available WGS data we realigned short reads around the SV breakpoints predicted by OM, decreasing breakpoint uncertainty to a couple of hundred base pairs.

Here, we show the ability of OM to detect large structural variants otherwise missed by SRS or chromosomal microarrays with some occurring in known DSD genes. We developed methods to integrate and annotate SRS and OM data to provide the ability to survey variants ranging from SNVs to large SVs.

A- A+

PgmNr 81: Developmental genome anatomy project (DGAP): Are there important clinical insights still to be learned from classical cytogenetics?

Authors:
R. Pina-Aguilar [1, 2]; O. Altiok-Clark [1, 2]; B.B. Currall [2,3]; C. Lowther [2,3]; K. Nalbandian [1,4]; J.F. Gusella [1,2,3]; E. Liao [2,3]; M.E. Talkowski [2,3,5]; C.C. Morton [1,2,5,6]

View Session Add to Schedule

Affiliations:
1) Brigham and Women's Hospital, Boston, MA; 2) Harvard Medical School, Boston, MA; 3) Massachusetts General Hospital, Boston, MA; 4) Massachusetts College of Pharmacy and Health Sciences University, Boston, MA; 5) The Broad Institute, Boston, MA; 6) University of Manchester, Manchester, UK

The Developmental Genome Anatomy Project (DGAP) was established in 1999 to investigate apparently balanced chromosomal abnormalities (BCAs) as a powerful means to identify single gene etiologies in developmental phenotypes. In the past two decades DGAP has made a multitude of novel gene discoveries in developmental disorders and continues to expand our understanding of the properties of BCAs and their functional consequences beginning with cases identified by traditional cytogenetic analyses. Various techniques including karyotyping, FISH, Sanger and massively parallel sequencing have been employed in DGAP. Long-insert genome sequencing (liGS) was established as a cost-effective sequencing method that can refine breakpoints and detect copy number changes involved in some chromosomal rearrangements. DGAP has recruited 284 families from 19 countries. 192 BCAs are de novo, 32 are familial cases co-segregating with the phenotype and in 60 the origin is unknown. Prenatal cases (3 from in vitro fertilization) constitute 7% (20) of DGAP probands; 19 BCAs are de novo and one inherited with intrafamilial variability in the phenotype. Rearrangements include simple translocations (208; 73%) and inversions (49; 17%), complex rearrangements (22; 8%), insertions (4; 1%), and a ring (1; 0.35%). Using the referral karyotype, there were 724 breakpoints of which the most common were on chromosomes 2 (n=50 probands (7%)) and 1 (40, (6 %)); 32 were in the X chromosome (4%) and 5 in the Y chromosome (0.69%). Recurrent breakpoints from the referral karyotypes were at Xp11 (n=13), 2p21 (9) and 5q15 (8). In the majority of cases with similar karyotype breakpoints, there is no recurrent affected gene. However, specific clinical phenotypes predict pathogenicity of the BCAs by a gain of function mechanism, such as overgrowth disorders. The DGAP cohort has shown that classical cytogenetics remains a rich resource of insights into novel human disorders, provides knowledge of the clinical relevance of genome architecture and establishes that the future of cytogenetics is genome sequencing with structural rearrangement analysis (such as liGS). As the current state of clinical testing is dominated by panels/exome sequencing and microarrays, clinicians should be aware that BCAs are cryptic for those technologies.

A- A+

PgmNr 82: The genomic formation of human populations in Central and South Asia.

Authors:
V. Narasimhan; N. Patterson; D. Reich; on behalf of the Central and South Asia Ancient DNA Study

View Session Add to Schedule

Affiliation: Department of Genetics, Harvard Medical School, Boston, MA, 02139

To elucidate the extent to which the major cultural transformations of farming and the distribution of languages in Eurasia were accompanied by movement of people, we report genome-wide ancient DNA data from 523 individuals spanning the last 8000 years. Movements of people following the advent of farming resulted in genetic gradients across Eurasia that can be modeled as mixtures of seven deeply divergent populations. A key gradient formed in southwestern Asia beginning in the Neolithic and continuing into the Bronze Age (BA), extending as far east as the desert oases of Central Asia in peoples of the Bactria Margiana Archaeological Complex (BMAC). This supports the idea that the archaeologically documented dispersal of domesticates was accompanied by the spread of people from multiple centers of domestication. The main BMAC population carried no ancestry from Steppe pastoralists and did not contribute substantially to later South Asians. However, Steppe pastoralist ancestry appeared in outlier individuals at BMAC sites by the turn of the 2ndmillennium BCE. Using data from ancient individuals from the Swat Valley of northern South Asia, we show that Steppe ancestry then integrated south in the first half of the 2ndmillennium BCE, contributing up to 30% of the ancestry of modern South Asians. The Steppe ancestry in South Asia has the same profile as that in BA Eastern Europe, tracking a movement of people that affected both regions and that likely spread the unique shared features shared between Indo-Iranian and Balto-Slavic languages. The primary ancestral population of modern South Asians is a mixture of people related to early Holocene populations of Iran and South Asia that we detect in outlier individuals from two sites in cultural contact with the Indus Valley Civilization (IVC), making it plausible that it was characteristic of the IVC. After the IVC's decline, this population mixed with northwestern groups with Steppe ancestry to form the "Ancestral North Indians" (ANI) and with southeastern groups to form the "Ancestral South Indians" (ASI) whose direct descendants live today in tribal groups in southern India. Mixtures of these two post-IVC groups--the ANI and ASI--drive the main gradient of genetic variation in South Asia today. We reveal a parallel series of events leading to the spread of Steppe ancestry to both Europe and South Asia, thereby documenting movements of people that were likely conduits for the spread of Indo-European languages.

A- A+

PgmNr 83: The fine-scale genetic structure of the Han Chinese population and a reference panel of 100K individuals for genotype imputation and association studies.

Authors:
Wang [1]; C. Liu [1]; X. Ma [1]; Y. Gao [1,2]; X. Zhang [1,2]; Y. Pan [1]; C. Zhang [1]; Y. Wang [1]; S. Xu [1,2]

View Session Add to Schedule

Affiliations:
1) CAS-PICB, Shanghai, China; 2) School of Life Science and Technology, ShanghaiTech University, Shanghai, China

A reference panel facilitates complex diseases mapping using population-based association studies, which has been well-established and demonstrated its power for populations of European ancestry but lacks for Han Chinese, the largest ethnic group in East Asia and in the world. Here, we construct a reference panel of 100,000 Han Chinese individuals (Han100K), with whole-genome deep-sequenced or high-density genome-wide single-nucleotide variants (SNV) genotyped or imputed. The Han100K is the first and currently only available reference panel specific to the Han Chinese population and accounting for the fine-scale genetic structure. Overall, the main difference is between northern and southern subgroups, while higher divergence is observed among southern subgroups compared with northern subgroups. The population structure of the Han Chinese population is likely driven by gene flow from the geographically neighboring groups, followed by an isolation by distance pattern among Han Chinese sub-groups. Indeed, considerable ancestries are identified in Han Chinese to be derived from non-Han Chinese groups, indicating prevalent and complex gene-flow from surrounding ethnic minority groups. Generally, the northern Han Chinese have been influenced more by northern Chinese minorities, and southern Han Chinese more by southern minorities. With the high-quality individual-level variant data provided, we further create a database with a remote computer server that allow researchers to carry out ancestry inference, imputation and association studies in Han Chinese or populations of ancestry closely related to Han Chinese.

A- A+

PgmNr 84: Recent population history inferred from more than 5,000 high-coverage South Asian genomes.

Authors:
J. Wall [1]; J. Robinson [1]; S. Belsare [1]; A. Bhaskar [2]; R. Gupta [3]; J. Tom [4]; T. Bhangale [4]; R.K. Rai [5]; A. Butterworth [6]; J. Danesh [6]; V. Mohan [7]; A. Ghosh [8]; A. Barik [5]; A. Chowdhury [5]; D. Saleheen [9]; S. Kathiresan [10]; E. Stawiski [3]; A. Peterson [3]

View Session Add to Schedule

Affiliations:
1) Inst Human Gen, Univ California San Francisco, San Francisco, California.; 2) News Feed Division, Facebook, Menlo Park, CA; 3) Genomic Medicines Division, MedGenome, Foster City, CA; 4) Dept of Bioinformatics, Genentech, South San Francisco, CA; 5) Society for Health and Demographic Surveillance, Birbhum, West Bengal, India; 6) Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 7) Madras Diabetes Research Foundation, Chennai, Tamil Nadu, India; 8) GROW Research Laboratory, Narayana Nethralaya Foundation, Bengaluru, India; 9) Dept of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA; 10) Preventative Cardiology, Massachusetts General Hospital, Boston, MA

South Asia contains hundreds of different ethnic or caste groups, many of which are thought to be mostly or completely endogamous. However, the age of this extreme population structure (and the underlying caste system) is unknown, with estimates ranging from hundreds to thousands of years. We analyzed high-coverage whole-genome sequence data from 6,610 individuals, including 1,812 from Pakistan, 500 from Bangladesh, 1,356 from urban South India and 1,180 from the Birbhum district of West Bengal. We used these data to estimate recent changes in population size and split times between caste groups. We do not observe the huge excess of extremely rare variants that has been observed in multiple studies of European and African-American populations. This observation cannot be fully explained by recent inbreeding: simulations suggest that the estimated levels of consanguinity (7.7% are the offspring of 1st cousin marriages, and 27.8% are the offspring of 2nd cousin marriages) will have a modest effect on the site frequency spectrum. Inbreeding with longstanding endogamy though may mostly explain our results.
Next, we developed a novel method for estimating the genome-wide average divergence time between a single individual and a focal group. This method focuses on extremely rare variants, which should be the most informative about very recent demographic events, and is robust to demographic events affecting the particular individual studied. We focused this work on samples from Birbhum district, West Bengal due to the presence of additional metadata on caste and religion. We used 704 general-caste individuals from Birbhum as the focal group, and estimated divergence times for all other individuals. Mean divergence times ranged from ~2,600 years for the Santal, an Austro-Asiatic language speaking tribal group, to 850 years for “scheduled castes” (i.e., Dalits), 625 years for Bangladeshis and 225 years for “Other Backward Castes” (OBC) individuals. The recent divergence times for OBC individuals confirms that this category is more of a political construct than a long-lived social grouping, while the other divergence times suggest a substantial amount of gene flow between groups. Finally, we extended our approach to thousands of other genomes from around the world. We show how patterns of rare variation can be used to detect asymmetrical migration, and document evidence for more migration from East Asia into Bengal than the converse.

A- A+

PgmNr 85: Fine-scale population structure and demographic history of British Pakistanis and its implications for disease risk.

Authors:
E. Arciero [1]; T. Tsismentzoglou [2]; D. Mason [3]; N. Small [4]; E. Sheridan [2]; R. Trembath [5]; M.E. Hurles [1]; D.A. van Heel [6]; J. Wright [3]; M. Iles [2]; H.C. Martin [1]

View Session Add to Schedule

Affiliations:
1) Human Genetics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, United Kingdom; 2) Leeds Institute for Data Analytics, Worsley Building, University of Leeds, Leeds, United Kingdom; 3) Bradford Institute for Health Research, Bradford, United Kingdom; 4) University of Bradford, Bradford, United Kingdom; 5) School of Basic and Biomedical Sciences, King's College London, London, United Kingdom; 6) Centre for Genomics and Child Health, Queen Mary University of London, London, United Kingdom

---

The Pakistani population is very diverse and contains different ethnic subgroups due to geographic origin, social organisation in kinship networks (*biraderi*) and language. We have analysed genetic data from > 4000 British Pakistani individuals recruited to the Born in Bradford and East London Genes & Health (ELGH) cohorts, by far the largest sample of Pakistani individuals yet used to address population genetic questions. We used a combination of SNP-chip and exome sequence data to explore the demographic history and genetic structure in the British Pakistani population, and the impact of the *biraderi* marriage system on autozygosity and on the frequency of rare damaging variants. The Bradford Pakistanis, although relatively homogeneous compared to the full spectrum of genetic variation in Pakistan, show signatures of population substructure within them, with most clustering according to their *biraderi* group rather than geographical origin. The 14 self-defined *biraderi* groups in Bradford show different proportions of ancestral components, revealing different genetic affinity both to other Pakistani and Indian populations and between themselves. The Pathan are the most distinct group followed by the Jatt, Choudhry, Bains and Rajput, which form two distinct subgroups. The Kashmiri, Syed, Sheikh, Mughal are more genetically similar to each other. Pakistanis within ELGH either cluster with the Bradford samples or represent potential further groups. Analysis of these groups suggests differing degrees of population bottleneck in historical times with Bains and Qasabi having the smallest estimated historical effective population sizes. We found that these bottlenecks have led to founder events for Mendelian disease-causing alleles in various groups. Examples are stop-gained and missense variants in genes such as *SAG*, *CPS1* and *PGM1* that are linked with different congenital diseases. Consistent with the high frequency of consanguineous marriage in the community, Bradford British-Pakistani individuals show high levels of autozygosity, with a mean of ~3.9% of the genome autozygous and ~25% of individuals having > 2% autozygosity. We are currently investigating the extent to which endogamy versus consanguinity contributes to disease in ~800 developmental disorder patients with Pakistani ancestry. This study highlights how social organisations such as the *biraderi* impact both the distribution of genetic variation and potentially the risk of rare genetic disorders.

A- A+

PgmNr 86: Identity-by-descent mapping of 87,131 genomes reveals the structure and recent genealogical history of Han Chinese populations.

Authors:
A. Lan; X. Yao; S. Tang; L. Wang; K. Kang; H. Weng; X. Wu; G. Chen

View Session Add to Schedule

Affiliation: WeGene, Shenzhen, Guangdong, China

Despite a number of studies on genetic polymorphism have widened our understanding of the evolutionary history of east Asian populations, the recent demography of the planet's largest ethnic group, Han Chinese, has yet to be revealed. Here we constructed a network of over 40 million identity-by-descent (IBD) segments from 87,131 genotyped Han Chinese individuals, and revealed the recent population structure and genealogical patterns. Through IBD clustering, the key factors influencing population patterns were disclosed: major geographic barriers, such as the Yangtze River, reduce gene flows; the historical and politicized migrations, such as the ‘Chuang Guandong’ immigration wave during the hundred-year period beginning in the last half of the 19th century, were captured; cultural and linguistic characteristics may drive noticeable separation of subgroups in a region (e.g. the Canton, Hakka and Chaoshan subgroups in Guangdong province). Our study demonstrated a high-resolution structure and dynamics of the Han Chinese populations in the last hundreds of years, which were driven by geographical, political and cultural factors.

A- A+

PgmNr 87: A paleogenomic reconstruction of the deep population history of Central and South America.

Authors:
N. Nakatsuka [1,2]; C. Posth [3,4]; I. Lazaridis [1]; C. Barbieri [3,5]; P. Skoglund [6]; T. Lamnidis [3]; S. Mallick [1]; N. Rohland [1]; J. Sandoval [7]; R. Burger [8]; E. Tomasto-Cagigao [9]; C. Méndez [10]; G. Politis [11]; G. Valverde [12]; A. Cooper [12]; B. Llamas [12]; W. Haak [3]; D. Kennett [13,14]; A. Strauss [15,16,17]; J. Krause [3,4]; D. Reich [1,18]; L. Fehren-Schmitz [19,20]

View Session Add to Schedule

Affiliations:
1) Department of Genetics, Harvard Medical School, Boston, MA, USA; 2) Harvard-MIT Division of Health Sciences and Technology, Boston, MA, USA; 3) Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany; 4) Institute for Archaeological Sciences, Archaeo- and Palaeogenetics, University of Tübingen, Tübingen, Germany; 5) Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland; 6) Francis Crick Institute, London NW1 1AT, UK; 7) Centro de Genètica y Biología Molecular, Facultdad de Medicina, Universidad de San Martín de Porres, Lima, Peru; 8) Department of Anthropology, Yale University, New Haven, CT, USA; 9) Department of Humanities, Pontifical Catholic University of Peru, San Miguel, Peru; 10) Centro de Investigación en Ecosistemas de la Patagonia, Coyhaique, Chile; 11) INCUAPA-CONICET, Facultad de Ciencias Sociales, Universidad Nacional del Centro de la Provincia de Buenos Aires, Olavarría, Argentina; 12) Australian Centre for Ancient DNA, School of Biological Sciences and The Environment Institute, Adelaide University, Adelaide, SA, Australia; 13) Department of Anthropology, The Pennsylvania State University, University Park, PA, USA; 14) Institutes for Energy and the Environment, The Pennsylvania State University, University Park, PA, USA; 15) Departamento de Genètica e Biologia Evolutiva, Universidade de São Paulo, São Paulo, Brazil; 16) Museu de Arqueologia e Etnologia, Universidade de São Paulo, São Paulo, Brazil; 17) Centro de Arqueologia Annette Laming Emperaire, Miguel A Salomão, Lagoa Santa, MG, Brazil; 18) Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA; 19) UCSC Paleogenomics, University of California, Santa Cruz, Santa Cruz, CA, USA; 20) UCSC Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA

We generated new genome-wide ancient DNA from 113 individuals forming four parallel time transects in Belize, Brazil, the Central Andes, and the Southern Cone, each dating to at least ~9,000 years ago (BP). We find that the common ancestral population to Central and South Americans radiated rapidly from just one of the two early branches that contributed to Native Americans today. We document two previously unappreciated streams of gene flow between North and South America. One affected the Central Andes by 4,200 BP and thoroughly mixed into all groups there by 1,500 BP. The other explains an affinity between the oldest North American genome associated with the Clovis culture and the oldest Central and South Americans from Chile, Brazil and Belize. However, this was not the primary source for later South Americans, as the other ancient individuals derive from lineages without specific affinity to the Clovis associated genome, suggesting a population replacement that began at least 9,000 BP and was followed by substantial population continuity in multiple regions.

We not only summarize our recent study of the early population structure of Central and South

America (Posth, Nakatsuka et al. Cell 2018) but also report a deep resolution analysis of the Central Andes region (Peru, Bolivia and North Chile), including many archaeological cultures without prior genomic data, notably the Moche, Wari, Tiwanaku, Inca and Chanka. We find that the population structure that distinguishes the Indigenous inhabitants of the North vs. South highlands today is directly related to the structure that was already developing about 6,000 BP, and also to the structure that distinguished highlands peoples from those of other regions beginning at least around 9,000 BP. We document bi-directional gene flow between the North and South highlands, as well as between the highlands and the coast. We also report changes and population heterogeneity within the Titicaca Basin and Cusco regions involving immigration from groups originating from elsewhere in the Central Andes coinciding with the rise of the Tiwanaku and Inca polities. On a continent-wide scale, our study reveals long-range gene flow from the Southern Central Andes into Argentina as well as between the Central Andes and Northwest Amazon Basin. Lastly, we observe a sacrifice in southern Argentina of a child with North Peruvian Coast ancestry during the Inca period.

PgmNr 88: Network centrality measures of colocalized genes and phenotypes in UK Biobank and 48 tissues from the Genotype-Tissue Expression Project.

Authors:
G. Rocheleau [1,2]; A. Dobbyn [1,3]; E. Stahl [1,3]; H.-H. Won [4]; R. Do [1,2]

View Session Add to Schedule

Affiliations:
1) Genetics & Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 2) Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 3) Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 4) Samsung Advanced Institute for Health Sciences and Technology (SAIHST), Sungkyunkwan University, Seoul, South Korea

Introduction
Genome-wide association studies (GWAS) have identified thousands of loci associated with various human diseases and traits. The majority of these studies have examined pleiotropy with other related traits and diseases. However, it is unknown how these loci are similar in terms of the number and type of genetic association phenotypes that they share, and whether this can reveal shared biology between genes.

Material and Methods
We applied coloc2, a Bayesian colocalization approach which estimates the posterior probability that a gene is causal in both GWAS and expression quantitative trait loci (eQTL) studies. Two sets of GWAS summary association statistics were considered as input for coloc2: the UK Biobank association results for binary and continuous traits (Neale lab); and SAIGE binary phenotype association results (Zhou et al., Nat Genet 2018) using PheCodes based on International Classification of Diseases codes in UK Biobank. Associated eQTLs for 48 tissues were downloaded from the current release (V7) of the Genotype-Tissue Expression project. For each tissue, genes with posterior probability of colocalization > 0.80 were included in a bipartite network of genes and phenotypes. Diverse network centrality measures of importance, such as eigenvector, betweenness or PageRank, were computed in the gene network projection for each tissue separately, and for all tissues aggregated.

Results
For traits such as diabetes and cardiovascular diseases, we confirmed loci with high network centrality measures were previously suggested as being causal by other studies and further identified novel genes were likely to be implicated in specific diseases and traits. Many colocalized genes were expressed in unsuspected tissues for some disease, emphasizing the importance of aggregating results over a wide range of different tissues. The gene network revealed connections between many loci through shared genetic association phenotypes.

Conclusion
Network centrality measures applied to colocalization of genetic association phenotypes and eQTL data in diverse tissues and cell types identified central genes in specific diseases, traits and tissues, and uncovered unknown biological connections between genes.

A- A+

PgmNr 89: Leveraging gene co-expression pattern to infer trait-relevant tissues in genome-wide association studies.

Authors:
L. Shang [1]; X. Zhou [1,2]

View Session Add to Schedule

Affiliations:
1) Biostatistics, University of Michigan, Ann Arbor, Michigan.; 2) Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan.

Genome-wide association studies (GWASs) have identified many SNPs associated with common diseases. However, understanding the biological functions of these identified SNPs associations requires identifying disease-relevant tissues or cell types. Several computational methods, such as LDSC and RolyPoly, have been recently developed to integrate omics studies into GWASs for inferring disease-relevant tissues or cell types. However, existing methods have so far failed to account for an important biological feature of gene expression data: genes are interconnected with each other and are co-regulated together. Such gene co-expression pattern occurs in a tissue specific or cell type specific fashion and may contain invaluable information for inferring disease-tissue relevance. Indeed, one key hypothesis in the recent omnigenic model also states that tissue-specific gene network underlies the etiology of various common diseases. Here, we develop a network method to incorporate tissue-specific gene co-expression networks constructed from either bulk or single cell RNA sequencing studies into GWAS data to facilitate the inference of trait-relevant tissues or cell types. Our method relies on a covariance regression network model to express gene-level effect sizes for the given GWAS trait as a function of the tissue-specific adjacency matrix, and with a composite likelihood-based inference algorithm, is scalable to accommodate hundreds of thousands of genes. We refer to our method as CoCoNet and validate its performance through extensive simulations. We apply our method for an in-depth analysis of four neurological disorders (schizophrenia, bipolar disorder, bipolar disorder/schizophrenia, Alzheimer’s disease)and four autoimmune diseases (primary biliary cholangitis, Crohn’s disease, ulcerative colitis, and inflammatory bowel disease). These in-depth analyses provided critical evidence supporting the recently proposed glutamate hypothesis for schizophrenia, and also revealed that SNPs associated with Crohn’s disease can lead to distinct gene co-expression regulation in the tissues harboring disease symptoms (such as in colon and skin).In addition, our results provide important empirical evidence supporting a key hypothesis of the omnigenic model that trait-relevant gene co-expression network underlies disease etiology.

A- A+

PgmNr 90: Calculating principled gene priors for genetic association analysis.

Authors:
L. Thakur; J. Flannick

View Session Add to Schedule

Affiliation: Genetics/Metabolics, Boston Children's Hospital, Boston, MA.

It has long been appreciated that prior knowledge of a gene’s function affects the likelihood it will harbor disease-associated variants. Today, an enormous number of datasets hold information that could in principle affect a gene’s “prior” for disease association, including pathway databases, functional annotations (e.g. UniProt), literature corpora, and model organism studies. Nearly all researchers use “gene annotations” within these datasets to assign *ad hoc* qualitative priors when interpreting experimental results. A logical question is therefore: to what extent do different annotations quantitatively affect a gene’s prior, and how can multiple annotations be combined into a principled prior estimate? We propose a new method to (a) estimate the effect of each gene annotation on the prior odds of association, based on excess per-SNP heritability nearby genes with the annotation; and (b) combine multiple annotations into an integrated prior odds estimate, accounting for overlap among them. Our method explicitly models how annotations impact a gene’s prior, permitting estimates of quantities such as (a) the statistical significance of a prior odds estimate, (b) which of a set of overlapping annotations has the highest effect on a prior, or (c) the fraction of variance in gene priors that can be explained by a set of gene annotations. We apply this method to assess how 2236 gene annotations from the knockout mouse project (KOMP) affect priors for type 2 diabetes (T2D) and 17 related quantitative traits. All 12 gene annotations corresponding to KO mouse glycemic phenotypes increase the prior odds of T2D association, with effects ranging from 2.0-fold (“increased insulin sensitivity”) to 6.2-fold (“increased insulin secretion”). Among the five annotations with the strongest, statistically significant increases on gene prior were “increased white fat cell size” (9.1-fold, p=0.043) and “pancreatic islet hyperplasia” (8.4-fold, p=0.019). When all annotations are combined, the genes with the highest T2D priors demonstrate enrichment for rare coding variant T2D associations in an analysis of 45,231 human exomes. We describe a public web-portal for accessing the prior estimates for each gene and dissecting the relative weighting of each gene annotation in each prior estimate. This resource can be used to prioritize genes for further experimental study, or to increase the power of association studies through re-weighting of association statistics.

A- A+

PgmNr 91: Transcriptomic-LD-score regression (tLDSC).

Authors:
C. Chatzinakos [1]; N.G. Harnett [1]; B.C. Bowlby [2]; L.D. Nickerson [1]; K.J. Ressler [1]; S.A. Bacanu [2]; N.P. Daskalakis [1]

View Session Add to Schedule

Affiliations:
1) Psychiatry, Harvard Medical School/ McLean, Belmont, MA.; 2) VIPBG, VIRGINIA COMMONWEALTH UNIVERSITY, Richmond, VA

Background: Biological intermediate phenotypes (BIPs), like imaging, are used in genetics to interpret disease risk. A well-established method to estimate heritability and genetic overlap of Genome Wide Association Studies (GWAS) of disease traits and BIPs is Linkage Disequilibrium (LD)-score regression. However, this correlation-based method agnostically aggregates signals over all SNP markers without identifying driver variants/genes. To increase the biological resolution of the genetic link between trait and BIP, we developed an alternative method, “transcriptomic-LD-score regression,” that introduces a molecular intermediate phenotype step in the calculation of the correlation.
Methods: We applied this method in GWAS data from the Psychiatric Genomics Consortium (PGC) traits and UK BioBank (UKBB), a comprehensive health study, that includes neuroimaging, from a large sample of individuals between the ages of 40-69. Our method first aggregated the GWAS SNP signals, for traits and for the heritable, Neuro-BIPs (NBPIs), at the transcriptomic-level using S-PrediXcan software and then estimated transcriptomic heritability, genetic covariance and genetic correlation
Results: We identified approximately 400 NBIPs (out of 3144) with significant heritability, adjusting for multiple comparisons. The analysis demonstrated that a greater percentage of structural MRI (19%; reflecting brain morphometry) and diffusion MRI (37%; reflecting structural connectivity) NBIPs were more heritable compared to resting state-functional MRI NBIPs (0.3%; reflecting brain functional connectivity).
We applied our method to all heritable NBIPs to correlate them with the GWAS from bipolar disorder (BPD), major depressive disorder (MDD), schizophrenia (SCZ), and posttraumatic stress disorder (PTSD) using brain-specific gene expression of over 10 brain regions. The heritability estimates using our method are on average smaller compared to LD-score regression given that we measure heritability using brain-specific gene expression which contains much less genetic information. Importantly, we observed that our transcriptomic-LD-score regression method detected larger effects and greater numbers of significant correlations (20% and 60% for BPD, 24% and 35% for MDD, 18% and 20% for PTSD, 30% and 27% for SCZ, respectively) compared to LD-score regression.
Conclusion: Our new transcriptomic-LD-score regression increased the link of BIPs with traits compared to standard LD-score regression.

A- A+

PgmNr 92: A powerful statistical framework to leverage tissue-specific gene expression regulation in identifying complex disease-associated genes.

Authors:
W. Liu [1]; M. Li [2]; W. Zhang [2]; G. Zhou [1]; J. Wang [1]; X. Wu [3]; H. Zhao [1,2,3]

View Session Add to Schedule

Affiliations:
1) Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT; 2) Department of Biostatistics, School of Public Health, Yale University, New Haven, CT; 3) Department of Molecular Cell Biology, Genetics and Developmental Biology, Yale Uni-versity, New Haven, CT

Recently, many methods such as PrediXcan and FUSION have been developed to identify novel disease-associated genes using gene expression as a mediating trait linking genotypes and diseases. However, the tissue-specificity of gene expression regulation is often poorly modeled, leading to increases in tissue-similarities after gene expression imputation and power loss in identifying tissue-specific associated genes in relevant tissues. Here, we introduce a new method named T-GEN (**T**ranscriptome-mediated identification of disease-associated **G**enes with **E**pigenetic a**N**notation) to identify disease genes leveraging tissue-specific epigenetic information. Built on the assumption that SNPs with active epigenetic signals are more likely to be expression quantitative trait loci (eQTL), T-GEN is able to impute expression for more genes (2.6%~55.3%) without much increase of tissue-similarities (3.3% compared to 165% in PrediXcan) after imputation. Applying T-GEN to 208 traits from the LD Hub, we were able to identify more disease-associated genes (7.7%~102%) compared to existing methods. In the tissue with the most-enriched heritability for 102 traits, T-GEN identified the largest number of associated genes compared to previous methods. More specifically, applying T-GEN to late-onset Alzheimer's disease (LOAD) from the LD Hub (n=54,162), we identified 96 associated genes at 15 loci and further replicated 48 genes in an external GWAS dataset (GWAX, n=114,564). Compared to existing methods like PrediXcan (79 genes at 10 loci) and FUSION (81 genes at 11 loci), T-GEN identified the largest number of associated genes/loci (96 genes at 15 loci). Furthermore, most genes identified by PrediXcan (78%) and FUSION (83%) were also identified by T-GEN while ~30% of T-GEN genes were uniquely identified by T-GEN. In addition to identifying new genes in known loci, like *ADRA1A* (p=1.85e-8) in *PTK2B* locus, T-GEN identified two loci (*COG4*, p=1.35e-7 and *TMEM135*, p=1.80e-8) located beyond 1MB of known GWAS loci. Further pathway enrichment analysis identified apoptosis-related network, statin pathway and ApoE-related inflammation and atherosclerosis, and statin pathway was again identified as enriched pathway (p=0.067) in the GWAX data. Through leveraging epigenetic annotation, T-GEN increases the power of tissue-specific disease gene identification with a focus on tissue-specific gene expression imputation, and provides novel insights into the molecular mechanisms of complex human traits.

A- A+

PgmNr 93: Ensemble of colocalization methods improves causal gene prioritization in simulations and GWAS.

Authors:
M.J. Gloudemans [1]; A.S. Rao [2,3]; B. Liu [4]; B. Balliu [5,6]; D. Calderon [1]; J.K. Pritchard [4,7,8]; E. Ingelsson [2,9]; S.B. Montgomery [5,7]

View Session Add to Schedule

Affiliations:
1) Biomedical Informatics Training Program, Stanford School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA; 2) Cardiovascular Institute, Stanford School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA; 3) Department of Bioengineering, Stanford University, Stanford, CA 94305, USA; 4) Department of Biology, School of Humanities and Sciences, Stanford University, Stanford, CA 94305, USA; 5) Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA; 6) UCLA Department of Biomathematics, 621 Charles Young Drive South, 5215 Life Sciences Los Angeles, CA 90095-1766; 7) Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; 8) Howard Hughes Medical Institute, Stanford University, Stanford, California 94305, USA; 9) Department of Medicine, Stanford University, Stanford, CA 94305, USA

A multitude of GWAS follow-up tools have emerged to connect molecular-level genetic effects (measured using eQTL, sQTL, and other QTL studies) and organism-level effects (measured using GWAS). Some of these methods, such as eCAVIAR, COLOC, and enloc, are described as “colocalization” methods. However, other methods based on transcriptome imputation (TWAS, PrediXcan), Mendelian randomization (SMR, GSMR), or stepwise regression (RTC) have also elucidated causal genes, providing an extensive but often confusing palette of software for GWAS follow-up. As a result, validation experiments must sift out true causal genes from a pile of false positives.

To determine when each of these methods is most appropriate, we integrated all eight in a single comparative pipeline, along with a naïve eQTL lookup method and our own ensemble method unifying all others. We gauged these methods’ performance and sensitivity to hyperparameters on a set of loci with simulated shared and non-shared genetic effects under varying LD patterns, number of causal SNPs per locus, and effect sizes of causal variants. While every method outperformed the naïve eQTL lookup under typical parameter settings, the ensemble approach most effectively distinguished true from spurious colocalizations.

Application of these approaches to recently published GWAS and eQTL datasets also showed poor consistency. The median overlap of colocalized genes between two methods was only 31%, which complicates prioritization of candidate genes. For example, in a recent colocalization analysis of coronary artery disease GWAS loci with coronary smooth muscle cell eQTLs, only 5 of 95 known GWAS loci colocalized with eQTLs in either of two methods (eCAVIAR and SMR), and the methods agreed on only a single causal gene. To systematically assess these discrepancies, we combined results from 60 GWAS. Adding top results from a second method expanded the set of candidate genes by 70% on average, but by integrating all methods in an ensemble approach, we achieved increased enrichment for evolutionarily conserved (phyloP) and predicted functional SNPs (Eigen), two proxies for disease risk.

In conclusion, published colocalization methods show low agreement but complement one another. Our work demonstrates how combining disparate methods will better prioritize disease mechanisms and drug targets.

A- A+

PgmNr 94: Determining genome-wide significance thresholds in biobanks with thousands of phenotypes: A case study using the Michigan Genomics Initiative.

Authors:
A. Annis [1]; A. Pandit [1]; E. Schmidt [1,2]; L. Fritsche [1]; P. VandeHaar [1]; C. Brummett [1]; S. Kheterpal [1]; V. Blanc [1]; E. Kaleba [1]; M. Boehnke [1]; S. Zöllner [1]; M. Zawistowski [1]; G. Abecasis [1]

View Session Add to Schedule

Affiliations:
1) Biostatistics, University of Michigan, Ann Arbor, Michigan.; 2) Wellcome Sanger Institute, Hinxton CB10 1SA, UK.

Genome-wide association studies (GWAS) test millions of genetic variants for association. To account for the large numbers of tests, a p-value of 5e-8 is traditionally required for statistical significance. Biobanks now provide the unprecedented opportunity to analyze thousands of phenotypes by linking genetic data and electronic health records. The large number of phenotypes and their correlation structure, which are specific to each biobank, further add to the multiple testing burden, obscuring how to properly distinguish true genetic associations when simultaneously assessing the results of thousands of GWAS.

To address this question, we performed a two-way replication experiment between independent biobank cohorts: the Michigan Genomics Initiative (MGI), which comprises >64,000 patients enrolled at Michigan Medicine, and the UK Biobank (UKB). We identified 854 independent associations in the MGI GWAS with p<5e-8 across 602 traits that were also present in UKB. Only half of these associations had consistent direction of effect in UKB; of these only 204 had p<0.05, giving an overall replication rate of just 23.8%. The replication rate, however, increased dramatically with increased stringency of the discovery p-value from the MGI analysis. We replicated ~75% of associations with MGI p<5e-10 and ~100% with MGI p<5e-20. We repeated this experiment in the opposite direction, finding that 876 of 2,736 independent associations (32.0%) with p<5e-8 in UKB replicated in MGI. Applying a more stringent discovery p-value, we replicated ~50% of association signals with UKB p<5e-10 and ~80% of association signals with UKB p<5e-20. Taken together, these results indicate that the typical 5e-8 significance threshold produces large numbers of spurious results when testing across thousands of phenotypes. Determining the precise significance threshold depends on properties of the respective biobank.

We present a computationally efficient permutation strategy to compute significance thresholds tailored to the number and correlation structure of traits in a biobank, as well as the sample size and case-control ratios for each trait. We show that when applied to the MGI dataset, our permutation strategy identifies significance thresholds that maintain desired type I error rates based on replication in UKB. Importantly, this method can be implemented in any biobank to provide customized significance thresholds that account for distinct properties of the respective dataset.

A- A+

PgmNr 95: A survey of epigenetic variation in >23,000 individuals identifies many disease-relevant epimutations and novel CGG expansions.

Authors:
A.J. Sharp [1]; P. Garg [1]; B. Jadhav [1]; O. Rodriguez [1]; N. Patel [1]; A. Martin-Trujillo [1]; M. Jain [2]; H. Olsen [2]; B. Paten [2]; B. Ritz [3]

View Session Add to Schedule

Affiliations:
1) Dept of Genetics & Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA.; 2) UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA; 3) Department of Epidemiology, Fielding School of Public Health and Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

There is growing recognition that epimutations, most often recognized as promoter hypermethylation events, are associated with a number of human diseases. While projects such as ExAC have provided a catalog of rare coding variation, little information exists on the prevalence and distribution of rare epigenetic variation. Here we report results of a survey of Illumina 450k array methylation profiles from 23,116 individuals, representing the largest cohort of methylomes ever assembled. Using a robust outlier approach, we identified a total of 13,938 epimutations at 4,481 distinct loci, each of which comprised multiple concordant CpGs. These included potentially inactivating promoter methylation events at 410 OMIM disease genes, including seven considered to be pathogenic and clinically actionable by the ACMG. For example, we observed promoter hypermethylation of *BRCA1* and *LDLR* at population frequencies of ~1 in 3,000 and ~1 in 6,000 respectively, suggesting that epimutations are an under-appreciated phenomenon and likely underlie a significant fraction of human genetic disease. Using expression data from 7,786 samples, we confirmed that many epimutations are associated with outlier gene expression, and thus potentially exert phenotypic effects.

To gain insight into mechanisms underlying the origin of epimutations, we performed three complementary analyses: (i) Using 700 MZ twin pairs, we observed that 30% of epimutations show discordant methylation patterns between identical twins, indicating that some epimutations are likely sporadic and/or occur post-zygotically. (ii) Contrastingly, using phased SNP data from 3,572 samples, we observed a significant excess of shared haplotypes among carriers of some epimutations, indicating that a different subset segregate in the population secondary to underlying sequence mutations. (iii) Finally, we identified 35 loci where rare gains of methylation coincide with the presence of unstable CGG repeats. Using long read sequencing, we validated that these epimutations are caused by novel CGG expansions, identifying the molecular defect underlying most of the known folate-sensitive fragile sites in the human genome.

Our study provides (i) a catalog of rare epigenetic changes in the human genome, (ii) identifies many novel hypermethylated CGG repeat expansions, and (iii) indicates that epimutations represent a subset of pathogenic alleles at many disease loci, which would be missed by purely sequence-based approaches.

A- A+

PgmNr 96: Open access to dbGaP new aggregated allele frequency for variant interpretation.

Authors:
L. Phan; H. Zhang; W. Qiang; E. Shekhtman; E. Moyer; E. Ivanchenko; D. Revoe; D. Shao; R. Villamarin; Y. Jin; M. Kimura; M. Feolo; J. Wang; N. Sharopova; M. Bihan; A. Sturcke; M. Lee; N. Popova; L. Hao; W. Wu; C. Bastiani; M. Ward; V. Lyoshin; K. Kaur; J.B. Holmes; B.L. Kattman

View Session Add to Schedule

Affiliation: NIH, NLM/NCBI, Bethesda, Maryland.

NCBI database of Genotypes and Phenotypes (dbGaP) contains the results of over 1,200 studies investigating the interaction of genotype and phenotype. The database has over two million subjects and hundreds of millions of variants along with thousands of phenotypes and molecular assay data. This unprecedented volume and variety of data promise huge opportunities to identify genetic factors that influence health and disease. With this possibility, NIH has recently updated the Genomic Summary Results (GSR) access restriction to allow responsible sharing and use of the dbGaP GSR data (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html).

In fulfilling the updated GSR policy and to improve variant interpretation for health and disease, NCBI has undertaken the challenging task to compute allele frequency for variants in dbGaP across approved un-restricted studies and provide the data as ‘open-access’ to the public. The work involved harmonizing and normalizing heterogeneous data and file formats either from GWAS chip array or direct sequencing. Using dbSNP and dbGaP workflows the data were QA/QC and were transformed to standard VCF format as input into an automated pipeline to aggregate, remap and cluster to existing dbSNP rs, and compute allele frequency. Allele frequencies are calculated for 12 major populations including European, Hispanic, African, Asian, and others that were computed using GRAF-pop (Jin et al., 2019).

The initially released data (Q2-2019) included MAF for about 500M sites with data in dbSNP and +20M novel sites from +150 thousand subjects across more than 60 studies. dbGaP MAF data are consistent with MAF data previously reported in GnomAD for the same variants. Moreover, dbGaP has frequency data for novel and existing variants in dbSNP and ClinVar but not reported in 1000Genomes, GnomAD, ExAC, or TopMed. The data volume will grow and can potentially reach over a billion variants from millions of subjects combined across all dbGaP studies. New studies will be added to future dbSNP build release for ‘de novo’ allele frequency calculation across all studies. This presentation will describe the available resources (Web, FTP, and API) and how researchers, clinicians, and developers can incorporate these data into their workflows and applications to understanding human variation and disease.

Acknowledgments:
Work at NCBI is supported by the NIH Intramural Research Program and the National Library of Medicine.

A- A+

PgmNr 97: Patterns of mitochondrial DNA variation across 15,000 individuals in the gnomAD database.

Authors:
N.J. Lake [1,2]; K. Laricchia [3,4]; G. Tiao [3,4]; S. Pajusalu [1,5,6]; L. Gauthier [3]; M. Shand [3]; J. Soto [3]; J. Emery [3]; D.G. MacArthur [3,4]; V.K. Mootha [3,4,7]; S.E. Calvo [3,4]; M. Lek [1,3]; gnomAD Consortium

View Session Add to Schedule

Affiliations:
1) Department of Genetics, Yale School of Medicine, New Haven, CT, USA; 2) Murdoch Children’s Research Institute, Royal Children’s Hospital, Melbourne, Australia; 3) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Massachusetts General Hospital, Boston, MA, USA; 5) Department of Clinical Genetics, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia; 6) Department of Clinical Genetics, United Laboratories, Tartu University Hospital, Tartu, Estonia; 7) Howard Hughes Medical Institute, Chevy Chase, MD, USA

The human mitochondrial DNA (mtDNA) encodes 37 genes, including 13 protein-coding, 2 ribosomal RNA (rRNA) and 22 transfer RNA (tRNA) genes. Pathogenic variants in the mtDNA cause mitochondrial disease, a clinically heterogenous group of disorders that can manifest as severe disease in childhood or in early adulthood. Analysis of large mtDNA population databases can greatly aid variant interpretation for clinical use.

Driven by the hypothesis that regions invariant across the human population may reflect an essential role of a region in mitochondrial function, we examined mtDNA variation in ~15,000 WGS samples from gnomAD. Variants were called using a newly developed mitochondrial-mode version of Mutect2, and all variants relative to the reference mtDNA were considered. In total, more than 11,000 unique variants were identified at over 9,000 unique mtDNA positions in gnomAD. To capture variation within populations not in gnomAD, we also included over 4,000 haplogroup variants in PhyloTree Build 17 in our analyses; over 10% of these were not present in gnomAD.

We then identified mtDNA regions that lacked any variation in our dataset. As expected, we observed variation at the majority of control region, intergenic, and synonymous sites. In contrast, non-synonymous, rRNA, and tRNA sites show substantial constraint with the majority of positions invariant across all humans studied (although low-heteroplasmy variation might be present). We dive deeply into rRNA, tRNA, and protein-coding regions showing particularly high constraint, and examine their overlap with known functional domains.

These positions invariant across the human population will be of utility for clinical interpretation of variants detected in rare disease. Indeed, we show enrichment of known pathogenic variants within the invariant positions. In sum, analysis of large population data have potential to facilitate mtDNA variant interpretation and provide functional insight into mtDNA regions.

A- A+

PgmNr 98: Public platform with 42,291 exome control samples enables association studies without genotype sharing.

Authors:
M. Artomov [1,2]; A.A. Loboda [2,3]; M.N. Artyomov [4]; M.J. Daly [1,2,5]

View Session Add to Schedule

Affiliations:
1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Broad Institute, Cambridge, MA; 3) ITMO University, St. Petersburg, Russia; 4) Department of Immunology and Pathology, Washington University in St. Louis, St. Louis, MO; 5) Institute for Molecular Medicine Finland, Helsinki, Finland

---

Introduction: Acquiring a sufficiently powered cohort of control samples can be time consuming or, sometimes, impossible. Accordingly, an ability to leverage control samples that were already collected and sequenced elsewhere could dramatically improve power in genetic association studies. Majority of the genotyped and sequenced human DNA samples to date are subject to strict data sharing regulations, large-scale sharing of, in particular, control samples is extremely challenging. We developed a method allowing selection of the best-matching controls in an external pool of samples that is compliant with personal genotype data protection restrictions.
Materials and Methods: We provide a web platform that stores 42,291 exome sequencing samples available for control selection and a complimentary R-package to be used on a user side for generation of anonymous data from case genotypes that will be uploaded to the web-platform.
Results: Our approach uses singular value decomposition of the matrix of case genotypes to rank external controls by similarity to cases without disclosing any individual-level data. We demonstrate that this recovers an accurate case-control association results for both ultra-rare and common variants independently of the sequncing platforms. We implemented framework for meta-analysis of multiple ancestries and sequencing platforms as a single step from the user side. Finally, we provide a free access to a database of 42,291 exomes to be used as external controls that enables association studies for case cohorts lacking control subjects and facilitates data sharing among projects with strict regulations for individual level data access.
Discussion: We present a freely accessible resource with a large-scale control database enabling association studies for “case-only” sequencing data with carefully selected controls and controllable error rates.

A- A+

PgmNr 99: Fast identity by descent detection across 500,000 UK Biobank samples reveals recent evolutionary history and population structure.

Authors:
J. Nait Saada [1]; A. Gusev [2,3]; P.F. Palamara [1]

View Session Add to Schedule

Affiliations:
1) Department of Statistics, University of Oxford, Oxford, Oxfordshire, United Kingdom; 2) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA; 3) Brigham & Women's Hospital, Division of Genetics, Boston, MA 02215, USA

Detection of Identical-By-Descent (IBD) segments provides a fundamental measure of genetic relatedness and plays a key role in a wide range of genomic analyses. We developed a new method, called FastSMC, that enables accurate biobank-scale detection of IBD segments transmitted by common ancestors living up to several hundreds of generations in the past. FastSMC combines a fast heuristic search for IBD segments with accurate coalescent-based likelihood calculations, and enables estimating the age of common ancestors transmitting IBD regions. We used coalescent simulation to verify that FastSMC outperforms the accuracy of existing methods in detecting IBD regions within 25, 50, 100, 150 and 200 generations (e.g. area under precision-recall curve improvement within the past 100 generations of 10% over RefinedIBD, 11% over GERMLINE and 80% over RaPID, after fine-tuning all methods), while requiring only marginally more time than GERMLINE, the most scalable method. We applied FastSMC to 487,409 phased British samples from the UK Biobank. We detected the presence of ~217 billion IBD segments transmitted by shared ancestors within the past 50 generations, obtaining a fine-grained picture of genetic relatedness within the past two millennia in the UK. We reconstructed region-specific effective population size within the past 50 generations, detecting substantially smaller recent effective size in the North of the country. After excluding close relatives (≤3rd degree cousins), the sharing of recent ancestry remained highly predictive of geographic co-localization, enabling us to estimate the birth coordinates of a random sample with an average error of 91km (K-nearest-neighbors using closest 5 individuals), a 68% improvement over standard genomic correlation. We sought evidence of recent positive selection by identifying loci with unusually high density of coalescence times within the past 50 generations. We detected 12 genome-wide significant signals, including 5 loci with previous evidence of positive selection (e.g. *LCT*, *HLA* and *LDLR*) and 7 novel loci (including *MRC1*, associated cardiovascular disease). Furthermore, the DRC statistic along the genome was significantly correlated with summary association statistics for LDL ($p < 0.01$ after adjusting for MAF and LD), consistent with recent evolutionary pressure on the trait. These results underscore the presence of subtle population structure and the widespread action of natural selection during recent millennia.

A- A+

PgmNr 100: Genome-wide rare variant analysis for thousands of phenotypes in 70,000+ exomes.

Authors:
E.T. Cirulli [1]; S. White [1]; R.W. Read [2,3]; G. Elhanan [2,3]; W.J. Metcalf [2,3]; K.A. Schlauch [2,3]; J.J. Grzymski [2,3]; J. Lu [1]; N.L. Washington [1]

View Session Add to Schedule

Affiliations:
1) Helix, San Carlos, California.; 2) Desert Research Institute, Reno, Nevada; 3) Renown Institute of Health Innovation, Reno, Nevada

Large-scale human genetic analyses have thus far focused on common variants, but the development of large cohorts of deeply phenotyped individuals with exome sequence data has now made comprehensive analyses of rare variants possible. Defining the effects that rare variants can have on human phenotypes is essential to advancing our understanding of human health and disease. Yet rare variant analysis presents its own challenges and requires its own methodology. Here, we analyzed the effects of rare (MAF<0.1%) variants on >3,000 phenotypes in 50,000 exome-sequenced individuals from the UK Biobank and performed replication as well as joint analyses with >1,000 phenotypes in >20,000 individuals from the Healthy Nevada cohort, exome sequenced at Helix. Our analyses of non-benign coding and loss of function (LoF) variants identified 78 gene-based associations that pass our strict statistical significance threshold (p<5x10-9). Of the 40 discovery associations whose phenotypes were represented in the replication cohort, 98% showed effects in the expected direction, and 48% attained formal replication significance (p<0.001). In addition, 6 significant associations were identified in our meta analysis of both cohorts together. These are associations where carrying any rare coding or LoF variant in the gene is associated with an enrichment for the phenotype in question, as opposed to GWAS-based associations with single variants. Among the results, we confirm known associations of *PCSK9* and *APOB* variation with LDL levels; we extend knowledge of variation in the *TYRP1* gene, previously associated with blonde hair color only in Solomon Islanders, to blonde hair color in individuals of European ancestry; and we make the novel discovery that *STAB1* variation is associated with blood flow in the brain. Importantly, our results do not suffer from the test statistic inflation that is often seen with rare variant analyses of biobank-scale data because of our rare variant-tailored methodology, which includes a step that optimizes the carrier frequency cutoff for each phenotype based on prevalence. We’ve made our results available for download and for interactive browsing in an app (https://ukb.research.helix.com). This comprehensive analysis of the effects of rare variants on human phenotypes marks one of the first steps in the next big phase of human genetics, where large, deeply phenotyped cohorts with next generation sequence data will elucidate the effects of rare variants.

A- A+

PgmNr 101: Whole genome sequence association analysis of body mass index in 45,159 TOPMed participants.

Authors:
Z. Li [1]; X. Li [1]; H. Zhou [1]; J. Brody [2]; M. Graff [3]; L. Lange [4]; K. North [3]; X. Lin [1,5]; TOPMed Anthropometry-Adiposity Working Group

View Session Add to Schedule

Affiliations:
1) Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts.; 2) Cardiovascular health research unit, University of Washington, Seattle,Washington; 3) Epidemiology, UNC Gillings School of Global Public Health, Chapel Hill, NC; 4) Biomedical Informatics and Personalized Medicine, School of Medicine, University of Colorado, Aurora, CO; 5) Statistics, Harvard University, Cambridge, MA

Introduction

Obesity is heritable, predisposes to many diseases and is commonly defined using body mass index (BMI). A large number of common and low frequency variants have been identified to be associated with BMI using GWAS. However, these common variants only explain a small fraction of heritability and a vast majority of variants in the human genome are rare. Because GWAS did not well captured rare variants (RVs), rare variant association analysis have so far not been very successful for detecting BMI-associated RVs. Ongoing large-scale whole genome sequencing (WGS) studies, such as the multi-ethnic NHLBI Trans-Omics Precision Medicine (TOPMed) Initiative, enable assessment of associations between BMI and rare variants (RVs) across the genome.

Hypothesis

Rare variant (RV) aggregations are associated with BMI.

Methods

We performed analysis of TOPMed Freeze 5 data from deep whole genome sequencing (>30X coverage) of 45,159 individuals across 20 TOPMed studies. After QC, a total of 326M variants were identified, 314M of which were rare variants (MAF<1%). For rare variant set association analysis, we applied our newly developed variant-Set Test for Association using Annotation infoRmation (STAAR) method, which dynamically incorporates multi-faceted variant function annotations to perform both whole genome gene-centric functional category based analysis and genetic region analysis using sliding windows across the genome.

Results

For RV-set gene-centric analysis, STAAR identified 2 genome-wide significant associations with BMI, including putative Loss-of-Function RVs in *GNA14* and missense RVs in *CREBRF*. After conditioning on known BMI-associated variants, the association of putative Loss-of-Function RVs in *GNA14* remained significant. In sliding window analysis, one region in the gene *FANCM* achieves genome-wide significance defined using Bonferroni correction for the number of tests . This association remains significant after adjusting for known BMI-associated variants.

Summary

Several novel RV-sets associated with BMI were identified using the TOPMed WGS Freeze 5 data. By incorporating variant functional annotation, STAAR empowers RV association analysis in WGS.

A- A+

PgmNr 102: Somatic variation observed across tissues in healthy individuals.

Authors:
S. Vattathil [1]; T.J. Comi [1,2]; L. Chen [1]; D.T. Akey [1]; J.M. Akey [1]

View Session Add to Schedule

Affiliations:
1) Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ; 2) Research Computing, Office of Information Technology, Princeton University, Princeton NJ

Theoretical calculations and the concept of genetic robustness predict that somatic variation accumulates over time even in healthy individuals. Recent empirical studies have confirmed this prediction, albeit for a limited number of tissues. These results emphasize that studying somatic variation in healthy individuals not only can reveal details of normal aging, but also is critical for correctly interpreting mutations in diseased tissue samples to make personalized diagnoses and treatment decisions. Specific outstanding questions include how frequently cancer-linked mutations occur without eliciting a disease phenotype, and how genetic fidelity changes over a lifetime. In an effort to address these questions and others, as part of the enhanced Genes, Tissues, and Expression (eGTEx) Project, we performed high-coverage (~150x) exome sequencing of 496 tissue samples collected through rapid autopsy from 23 donors aged between 21 and 69 years. The study comprises 10 to 32 (median 22) samples per donor from a wide range of body sites including skin, gastrointestinal organs, and brain. A subset of samples was sequenced in duplicate to assess technical variation. We produced a refined set of variant calls including somatic single-nucleotide variants (SNVs) and large chromosomal imbalances for each sample using a combination of published germline and somatic variant calling tools and custom workflows designed to exploit the availability of many samples per donor and minimize the impact of technical artifacts and other potential sources of false calls. Most samples have matching RNA sequence data generated by the GTEx project, which we assessed for support of mutations discovered from the DNA data. As expected, the observed somatic mutation burden in terms of number of loci and variant allele fraction is substantially lower than that in tumor samples. The mutation count varies widely across samples, and consistent with recent reports, the highest somatic mutation burden is found in sun-exposed skin and esophagus samples. Most SNVs are specific to a single sample, suggesting that they arose or expanded later in life. This study expands our understanding of normal somatic variation and its relevance to medical diagnostics. The results suggest that additional research aimed at characterizing normal somatic variation, especially in tissues with high renewal capacity and extrinsic mutagenic factors, will further resolve this subject.

A- A+

PgmNr 103: Parental somatic mosaicism for CNV deletions: A need for more sensitive and precise detection methods in clinical diagnostics settings.

Authors:
Q. Liu [1]; T. Wilson [1]; J.A. Rosenfeld [1,2]; C.M. Grochowski [1]; C.A. Bacino [1,2,3]; S.R. Lalani [1]; A. Patel [1,2]; A. Breman [4]; J.L. Smith [1,2]; S.W. Cheung [1]; J.R. Lupski [1,3,5,6]; W. Bi [1,2]; P. Stankiewicz [1]

View Session Add to Schedule

Affiliations:
1) Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX; 2) Baylor Genetics, Houston, TX; 3) Texas Children's Hospital, Houston, TX; 4) Dept. of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN; 5) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 6) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

Growing evidence implicates the importance of somatic mosaicism in the etiology of several human genetic conditions. Both affected and unaffected carriers of a pathogenic mosaic variant can transmit it to their children if it is also present in germline cells. In 2014, we reported identification of low-level (<10%) parental somatic mosaicism for copy number variant (CNV) deletions in 4% of 100 unrelated families and emphasized the importance of mosaicism detection for genetic counseling for recurrence risk (PMID: 25087610). In addition, investigation of genome-wide mutation rates and spectra in multi-sibling families using whole genome sequencing data revealed that in parental germ line, 3.8% of mutations were mosaic for SNVs, resulting in 1.3% of mutations being shared by siblings (PMID: 26656846). By querying the chromosomal microarray analysis (CMA) database at Baylor Genetics, we randomly selected an additional 46 unrelated family trios in whom CNV deletions (pathogenic, likely pathogenic, or variant of uncertain significance) were determined by CMA to be apparently *de novo* events in the affected children. Using junction-specific PCR and Sanger sequencing, we characterized the CNV deletion junctions in the affected children. Subsequently, we screened the parental blood samples and have identified four (8.7%) parents (three fathers and one mother) with suspected somatic mosaicism (for chr2:148,774,690-148,921,364, chr16:89,526,612-89,561,693, chr21:37,618,772-37805220, and chrX:152,973,280-153,033,394 (hg19)). Using droplet digital PCR with primers specific either for the deletion junction (jct) or mapping within the deleted region (del), we have assessed the levels of mosaicism in blood at 18.2% (jct), ~ 5% (del), undetectable (del), and ~ 15% (del), respectively. To validate these data and to determine the distribution of the detected mosaicism in various tissues representing ectoderm, endoderm, and mesoderm, we are now testing additional parental samples with other highly sensitive and quantitative methods, including blocker displacement amplification (BDA) (PMID: 29805844) and PCR amplicon-based next generation sequencing. Our results further imply that more sensitive and precise methods are needed in clinical diagnostic settings for testing parental somatic mosaicism and for better assessment of potential disease recurrence risk.

A- A+

PgmNr 104: Genomic profiling of surgically excised brain tissue uncovers somatic mosaicism in seizure disorders.

Authors:
D.C. Koboldt [1]; K.E. Miller [1]; E. Crist [1]; K.M. Leraas [1]; T.A. Bedrosian [1]; C.E. Cottrell [1]; V. Magrini [1]; D.R. Boué [2]; C.R. Pierson [2]; J. Leonard [3]; R.K. Wilson [1]; A. Ostendorf [4]; E.R. Mardis [1]

View Session Add to Schedule

Affiliations:
1) Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, Ohio, USA; 2) Department of Pathology and Laboratory Medicine, Nationwide Children's Hospital, Columbus, Ohio, USA; 3) Division of Neurosurgery, Nationwide Children's Hospital, Columbus, Ohio, USA; 4) Division of Neurology, Nationwide Children's Hospital, Columbus, Ohio, USA

Epilepsy is one of the most common conditions associated with neurological disorders, affecting as many as 1 in 26 individuals in the United States alone. While 20-30% of cases are caused by acquired conditions (stroke, head injury, or tumor), the majority are believed to have an underlying genetic etiology. Gene discovery in neurological disorders has implicated hundreds of genes, many of which have been incorporated into comprehensive clinical genetic testing panels. Despite these advances, the molecular basis of disease remains elusive for almost half of epilepsy cases.

A growing number of disorders have been associated with somatic mosaicism, in which genetic alterations are only found in a subset of cells in the body. Next-generation sequencing (NGS) enables the detection of mosaic variants at fairly low allele frequencies (~1-5%), provided that these mutations are present in the tissue specimen(s) sampled. Obtaining access to affected tissue, however, is particularly challenging for neurodevelopmental disorders. As a Level 4 epilepsy center, our institution offers advanced epilepsy surgery services, including evaluation using intracranial electrodes to identify epileptogenic zones and surgical excision of affected tissue. We recently initiated an IRB-approved study to perform genomic profiling of excised tissue and matched normal (blood) samples from patients with refractory epilepsy.

To date, we have performed whole-exome sequencing (WES) on 55 brain tissue samples from 16 patients and detected mosaic variants in 4/16 (25%). Two of these harbor pathogenic mutations in disease genes consistent with clinical features. In a 3-year-old male with cryptogenic West syndrome, WES uncovered a 2-bp deletion in *SLC35A2*, NM_005660.3:c.634_635delTC (p.Ser212LeufsTer9). Using deep targeted sequencing of *SLC35A2* coding regions (~2,500x depth), we detected this mutation in 12 brain tissue specimens at allele frequencies ranging from 4.8% to 19.5% and confirmed its absence from blood. In a 1-year-old male with left hemimegalencephaly, we identified two distinct mosaic indels in *PTEN* in left frontal lobe tissue. Both variants are frame shift indels present in a subset of cells (34.1% and 11.8%), and were confirmed via RNA-seq of same specimen.

Our results demonstrate the power of somatic NGS studies to identify mosaic pathogenic variants in surgically excised tissues, and suggest that somatic mosaicism represents an important etiology of neurological disorders.

A- A+

PgmNr 105: Mosaic copy number variants are associated with autism spectrum disorder.

Authors:
M. Sherman [1,2,3]; R. Rodin [3,4]; G. Genovese [3,5]; C. Dias [4,6]; C. Walsh [3,4]; P. Park [7]; P.-R. Loh [2,3]

View Session Add to Schedule

Affiliations:
1) Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.; 2) Department of Genetics, Brigham and Women’s Hospital, Boston, MA.; 3) Broad Institute of MIT and Harvard, Cambridge, MA.; 4) Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA; 5) Department of Genetics, Harvard Medical School, Boston, MA; 6) Division of Developmental Medicine, Boston Children's Hospital, Boston, MA; 7) Department of Biomedical Informatics, Harvard Medical School, Boston, MA

Germline copy number variants (CNVs) are a known cause of autism spectrum disorder (ASD), but whether mosaic (post-zygotic) CNVs arising during early embryogenesis also contribute to ASD risk is unknown. Here we answer this question in the affirmative using two independent genotyping datasets of individuals with ASD (probands) and their unaffected siblings and parents: the Simons Simplex Collection (SSC) (2594 probands and 2423 siblings) and SPARK collection (8875 probands and 2931 siblings). We leveraged statistical haplotype phasing together with available parental genotypes to sensitively detect mosaic duplications, deletions and copy-number neutral loss of heterozygosity (CNN-LOH). Since genotyping was performed on blood and saliva, we filtered events appearing to arise due to clonal hematopoiesis.

Mosaic CNVs were detected infrequently in both probands and unaffected siblings (0.37% of probands and 0.28% of siblings across SSC and SPARK). Nonetheless, in both cohorts, probands carried a significant burden of large (>4 Mbp) mosaic CNVs compared to siblings ($p = 0.043$ and $p = 0.011$ in SSC and SPARK, respectively; OR=10.76, CI=1.45-79.7, $p=0.0012$ combined). Although most CNVs were deletions and duplications, five were CNN-LOH including uniparental disomies of chromosomes 1 and 2 and two other CNN-LOH events disrupting 1p and 2q. These CNN-LOH events occurred exclusively in probands and created homozygosity for predicted loss-of-function SNVs.

We found clear links between several events and observed phenotypes. A proband with a mosaic 18q deletion lacked verbal communication, consistent with Pitt-Hopkins syndrome; and a proband with mosaic CNN-LOH of 11p had macrocephaly and clinical growth abnormalities, consistent with Beckwith-Wiedemann syndrome. Furthermore, several short CNVs in probands disrupted known autism-related genes (e.g., *TRIO* and *USP7*), suggesting even small mosaic CNVs may contribute to ASD etiology.

To confirm the presence of mosaic CNVs in neurons, we whole-genome sequenced post-mortem brain tissue from 60 Autism BrainNet samples. We identified one large mosaic duplication and validated its presence in 15% of neurons using digital-droplet PCR; breakpoint analysis revealed the event was composed of multiple inversions and duplications. Our results demonstrate a clear association between mosaic CNVs and autism spectrum disorder, suggest a link between CNN-LOH and ASD, and demonstrate that mosaic CNVs may arise as complex rearrangements.

A- A+

PgmNr 106: Single-cell analysis of human preimplantation embryos reveals widespread and complex patterns of chromosomal mosaicism.

Authors:
R.C. McCoy; M.R. Starostik

View Session Add to Schedule

Affiliation: Department of Biology, Johns Hopkins University, Baltimore, Maryland.

It is estimated that 40-60% of human embryos are lost between fertilization and birth, primarily due to aneuploidy. In contrast to meiotic errors, which generate uniform aneuploidy among all embryonic cells, postzygotic mitotic errors generate chromosomal mosaicism, whereby different cells possess distinct chromosome complements. While the incidence and fitness consequences of meiotic aneuploidy are well established, those of mosaic aneuploidy remain controversial. Resolving this controversy is critical to a basic understanding of human development, as well as for guiding fertility applications such as preimplantation genetic testing for aneuploidy (PGT-A) among *in vitro* fertilized (IVF) embryos.

One critical limitation in the study of mosaicism is that most inferences have been based on bulk DNA sequencing or comparisons of multiple biopsies of a few embryonic cells. Single-cell genomic data provide a promising opportunity to quantify mosaicism on an embryo-wide scale. To this end, we analyzed published single-cell RNA sequencing data from a total of 1463 cells from 88 disaggregated human embryos, spanning the cleavage to blastocyst stages. We inferred chromosome gains and losses for each sample based on chromosome-wide up- or down-regulation of gene expression, respectively, and also integrated signatures of allelic imbalance at heterozygous single nucleotide variants (SNVs).

Our analysis revealed widespread chromosomal mosaicism across preimplantation development. Among all embryos and stages, we detected a total of 645 (44%) aneuploid cells (5% FDR), with more than half of embryos harboring at least one aneuploid cell. Mosaic aneuploidies exhibited diverse patterns ranging from gain or loss of individual chromosomes to chaotic abnormalities impacting many cells and chromosomes simultaneously. By clustering these patterns within embryos, we reconstructed histories of chromosome mis-segregation and distinguished meiotic and early mitotic errors from those occurring after lineage differentiation. We further observed a significant enrichment in the allocation of aneuploid cells to the trophectoderm versus the inner cell mass (OR = 3.97, 95% CI [2.18, 7.23], $P = 6.3 \times 10^{-6}$), thereby complicating interpretation of standard trophectoderm biopsy in the context of PGT-A. Together, our work provides a high-resolution view of aneuploidy within human IVF embryos and suggests that low-level mosaicism may represent a physiologic state of preimplantation development.

A- A+

PgmNr 107: Identification of low-level parental somatic mosaic SNVs and InDels in a large exome sequencing cohort of individuals with Mendelian disorders.

Authors:
T. Gambin [1]; J.A. Karolak [1]; Q. Liu [1]; S.N. Jhangiani [2]; Z.H. Coban Akdemir [1]; J.R. Lupski [1,2,3,4]; P. Stankiewicz [1]

View Session Add to Schedule

Affiliations:
1) Dept. of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3) Texas Children's Hospital, Houston, TX; 4) Dept. of Pediatrics, Baylor College of Medicine, Houston, TX

Mosaic variants due to somatic mutations can cause both cancer and Mendelian conditions. In families with children affected with a Mendelian disorder, detection of somatic mosaicism in the parental samples is essential to better estimate the recurrence risk. Although high depth next generation sequencing facilitates detection of low-level mosaic variants, thus far, only a few studies have reported results of systematic analyses of low-level mosaicism in large exome sequencing (ES) cohorts. We computationally analyzed ES data from 882 unrelated family trios (with the complete VCF & BAM files) recruited in the Baylor Hopkins Center for Mendelian Genomics (BHCMG) at BCM. After removal of the low quality samples and variants with minor allele frequency (MAF) >0.0001 in the 1000Genomes or local databases, we found 3,156 apparent *de novo* SNVs or indels in 768 probands. To identify the candidate low-level (<10%) somatic mosaic variants, we calculated alternate allele fraction (AAF) for all rare (<0.0001) variants in the entire trio dataset with the total depth of read coverage >20x and AAF between 0.3-0.7 in the affected patients (likely heterozygous variant), and AAF lower than 0.1 but greater than zero (likely low-level mosaic variant) in one of the parental samples. We found 76 variants in 67 unrelated family trios fulfilling these criteria, including 53 missense, 9 splicing, 5 indels, 4 stop-gain, 4 non-coding, and 1 synonymous SNVs. To assess the quality of these findings, for each genomic position of putative mosaic SNVs, we retrieved the pileup information from 7788 ES samples. We found that on average 4% of BHCMG samples (ranging from 0 to 45%) carry at least one read supporting alternative allele at the position of potential mosaic variant. The selected candidate mosaic variants are being validated using three highly sensitive orthogonal molecular assays: (i) PCR amplicon-based next generation sequencing with high coverage; (ii) droplet digital PCR (ddPCR); and (iii) blocker displacement amplification (BDA) (PMID: 29805844). Moreover, we are attempting to evaluate and identify the most reliable predictors (e.g. GC content, mapping quality, frequency of individuals with alternate reads) that can facilitate discrimination between true positive from false positive mosaic variants. These predictors will be used for building a classification tool enabling accurate detection of low-level mosaicism in ES samples.

A- A+

PgmNr 108: Loss-of-function genomic variants with impact on liver-related blood traits highlight potential therapeutic targets for cardiovascular disease.

Authors:
J.B. Nielsen [1,2]; O. Rom [1]; I. Surakka [1]; S.E. Graham [1]; W. Zhou [1,3,4,5]; L.G. Fritsche [1,6,10]; S.A. Gagliano Taliun [6,8]; C. Sidore [9]; Y. Liu [1]; M.E. Gabrielsen [10]; A.H. Skogholt [10]; B. Wolford [1,5]; W. Overton [5]; TOPMed. program [11]; S. Lee [5]; H.M. Kang [5]; F. Cucca [9,12]; O.L. Holmen [10,13]; B.O. Åsvold [10,13,14]; M. Boehnke [5,8]; S. Kathiresan [15,16,17]; G.R. Abecasis [5,8,18]; Y.E. Chen [1]; C.J. Willer [1,5,10,19]; K. Hveem [10,13]

View Session Add to Schedule

Affiliations:
1) Department of Internal Medicine: Cardiology, University of Michigan, Ann Arbor, Michigan, USA; 2) Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark; 3) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA.; 4) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.; 5) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.; 6) Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA.; 7) Department of Dermatology, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway.; 8) Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA.; 9) Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy.; 10) K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, NTNU, Norway.; 11) The NHLBI Trans-Omics for Precision Medicine (TOPMed) program.; 12) Dipartimento di Scienze Biomediche, Università degli Studi di Sassari, Sassari, Italy; 13) HUNT Research Centre, Department of Public Health and Nursing, Norwegian University of Science and Technology, Levanger, Norway.; 14) Department of Endocrinology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway; 15) Harvard Medical School, Boston, Massachusetts, USA.; 16) Center for Genomic Medicine and Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, USA.; 17) Broad Institute, Cambridge, Maryland, USA.; 18) Regeneron Pharmaceuticals, Tarrytown, New York, USA.; 19) Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA.

Cardiovascular diseases (CVD), and in particular cerebrovascular and ischemic heart diseases, are leading causes of death globally. Lowering circulating lipids is an important treatment strategy to reduce risk. However, some pharmaceutical mechanisms of reducing CVD may increase risk of fatty liver disease or other metabolic disorders. To identify potential novel therapeutic targets, which may reduce risk of CVD without increasing risk of metabolic disease, we focused on the simultaneous evaluation of quantitative traits related to liver function and CVD. Using a combination of low-coverage (5x) whole-genome sequencing and targeted genotyping, deep genotype imputation based on the TOPMed reference panel, and genome-wide association study (GWAS) meta-analysis, we analyzed 12 liver-related blood traits (including liver enzymes, blood lipids, and markers of iron metabolism) in up to 69,479 participants in the population-based Nord-Trøndelag Health Study (HUNT) in Norway. Following fine-mapping of genomic regions using step-wise conditional analyses and trans-ancestry meta-analyses in up to 203,476 people, we identified 88 likely causal protein-altering variants that were associated with one or more liver-related blood traits. We identified several loss-of-function (LoF) variants reducing low-density lipoprotein cholesterol (LDL-C) or risk of CVD without increased risk of liver disease or diabetes, including variants in known lipid genes (e.g. *APOB*, *LPL*). A novel LoF variant, *ZNF529*:p.K405X*,* was associated with decreased levels of LDL-C (P=1.3x10^{-8}) but demonstrated no association with liver enzymes or non-fasting blood glucose levels. Silencing of *ZNF529* in human hepatocytes resulted in upregulation of LDL receptor (LDLR) and increased LDL uptake in the cells, suggesting that inhibition of *ZNF529* or its gene product could be used for treating hypercholesterolemia and hence reduce the risk of CVD. Taken together, we demonstrate that simultaneous consideration of multiple phenotypes and a focus on rare protein-altering variants may identify promising therapeutic targets.

A- A+

PgmNr 109: Computational discovery of candidates for drug repositioning for preterm birth.

Authors:
B. Le [1,2]; S. Iwatani [3]; R.J. Wong [3]; D.K. Stevenson [3]; M. Sirota [1,2]

View Session Add to Schedule

Affiliations:
1) Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA; 2) Department of Pediatrics, University of California, San Francisco, San Francisco, CA; 3) Department of Pediatrics, Stanford University, Palo Alto, CA

A preterm birth (PTB) is birth before completing 37thwks of gestation. According to the World Health Organization, in 2018, an estimated 15 million babies were born prematurely, with associated complications being the leading cause of infant mortality worldwide. In the USA, progesterone (P4) is the only drug approved by the FDA for use in the prevention of recurrent spontaneous PTBs. However, studies have shown that P4 has limited success, working in approximately only 1/3rdof cases[1]. The development of new drugs is both costly and time-consuming, and in the case of pregnancy, safety concerns are especially paramount. Computational drug repositioning leverages data on existing drugs that have already been developed and evaluated for safety in order to discover novel therapeutic uses. Repositioning using gene expression data has been successful for many indications, including inflammatory bowel disease, dermatomyositis, and various cancers. Using a rank-based pattern matching strategy, we compared a differential gene expression signature for PTB[2] to the differential gene expression profiles of drugs profiled in the Connectivity Map database[3] and assigned a reversal score to each disease-drug pair. 81 drugs, including notably P4, were found to have a significantly reversed differential gene expression compared to that found for PTB. 73 of these drugs had a reversal score higher than P4. Many of these drugs have been evaluated in the context of pregnancy, with 23 drugs belonging to pregnancy categories B or C – indicating no known risk in human pregnancy. From these drugs, we selected a compound with a good safety profile and validated it in an animal inflammation model using lipopolysaccharide (LPS), which showed a reduction of fetal wastage compared with both vehicle or P4 treatment alone. These promising results demonstrate the potential effectiveness of the computational drug repurposing pipeline for identifying compounds that might be effective in preventing PTB.

[1] Norwitz, E. R. & Caughey, A. B. Progesterone Supplementation and the Prevention of Preterm Birth. *Rev. Obstet. Gynecol.* **4,** 60–72 (2011).
[2] Vora, B. *et al.* Meta-Analysis of Maternal and Fetal Transcriptomic Data Elucidates the Role of Adaptive and Innate Immunity in Preterm Birth. *Front. Immunol.* **9,** (2018).
[3] Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935 (2006).

A- A+

PgmNr 110: Novel drug targets for ischemic stroke identified through Mendelian randomization analysis of the blood proteome.

Authors:
M. Chong [1,2]; J. Sjaarda [1]; M. Pigeyre [1]; P. Mohammadi-Shemirani [1,3]; R. Lali [1]; A. Shoamanesh [1,6]; H.C. Gerstein [1,4]; G. Pare [1,2,4,5]

View Session Add to Schedule

Affiliations:
1) Population Health Research Institute (PHRI), David Braley Cardiac, Vascular and Stroke Research Institute, Thrombosis and Atherosclerosis Research Institute, Hamilton Health Sciences, Hamilton, Ontario, Canada; 2) Department of Biochemistry, McMaster University, Hamilton, Ontario, Canada; 3) Department of Medical Sciences, McMaster University, Hamilton, Ontario, Canada; 4) Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; 5) Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada; 6) Department of Medicine, Division of Neurology, McMaster University, Hamilton, Ontario, Canada

---

Background: Novel drugs are warranted for treatment of ischemic stroke. Circulating protein biomarkers with causal genetic evidence represent promising targets but no systematic screen of the proteome has been done.
Methods: First, using Mendelian Randomization (MR), we assessed causality between 653 circulating proteins vs. three ischemic stroke subtypes: large artery atherosclerosis, cardioembolic stroke, and small artery occlusion. Second, we used MR to assess whether identified biomarkers also affect intracranial bleeding, specifically, intracerebral and subarachnoid hemorrhages. Third, we expanded this analysis to 679 diseases to test a broad spectrum of side-effects associated with hypothetical therapeutic agents. For all MR analyses, summary-level data from genome-wide association studies (GWAS) were used to ascertain genetic effects on circulating biomarker levels versus disease risk. Biomarker effects were derived by meta-analysis of five GWAS (N≤20,509). Disease effects were derived from large GWAS analyses, including MEGASTROKE (N≤322,150) and UKBiobank (N≤408,961) studies.
Results: Several biomarkers emerged as causal mediators for ischemic stroke. Causal mediators for cardioembolic stroke included histo-blood group ABO system transferase (ABO), coagulation factor XI (F11), scavenger receptor class A5 (SCARA5), and tumour necrosis factor-like weak inducer of apoptosis (TNFSF12). Causal mediators for large artery atherosclerosis included ABO, cluster of differentiation 40 (CD40), apolipoprotein(a) (LPA), and matrix-metalloproteinase-12 (MMP12). SCARA5 (OR=0.78; 95% CI, 0.70-0.88; P=1.46x10^{-5}) and TNFSF12 (OR=0.86; 95% CI, 0.81-0.91; P=7.69x10^{-7}) represent novel protective mediators of cardioembolic stroke. TNFSF12 also increased risk of subarachnoid (OR=1.53; 95% CI, 1.31-1.78; P=3.32x10^{-8}) and intracerebral (OR=1.34; 95% CI, 1.14-1.58; P=4.05x10^{-4}) hemorrhages, whereas SCARA5 decreased risk of subarachnoid hemorrhage (OR=0.61; 95% CI, 0.47-0.81; P=5.20x10^{-4}). Multiple side-effects beyond stroke were identified for six biomarkers, most (75%) of which were beneficial. No adverse side-effects were found for F11, LPA, and SCARA5.
Conclusions: Causal roles for five established and two novel biomarkers for ischemic stroke were identified. Side-effect profiles were characterized to help inform drug target prioritization. In

particular, SCARA5 represents a promising target for treatment of cardioembolic stroke with no predicted adverse side-effects.

A- A+

PgmNr 111: Proteome instability is an immunogenic therapeutic vulnerability in mismatch repair deficient cancer.

Authors:
N. Sahni [1,2,3]; D. McGrail [2]; J. Garnett [2]; S. Kopetz [4]; R. Broaddus [5]; G. Mills [2]; S.Y. Lin [2]

View Session Add to Schedule

Affiliations:
1) Epigenetics and Molecular Carcinogenesis, UT MD Anderson Cancer Center, Science Park, Smithville, Texas.; 2) Systems Biology. UT MD Anderson Cancer Center, Houston, Texas; 3) Bioinformatics and Computational Biology, UT MD Anderson Cancer Center, Houston, Texas; 4) Department of Gastrointestinal Medical Oncology UT MD Anderson Cancer Center, Houston, Texas; 5) Department of Pathology UT MD Anderson Cancer Center, Houston, Texas

Deficient DNA mismatch repair (dMMR) induces a hypermutator phenotype, leaving a genomic scar known as microsatellite instability (MSI). MSI is observed in approximately 30% of endometrial cancers1, 20% of gastric cancers2, 15% of colorectal cancers3, and in a smaller fraction of other tumor types. This hypermutator phenotype is thought to produce large numbers of immunogenic neoantigens, leading to the approval of MSI status as a clinical biomarker for immunotherapy. However, more than 60% of patients with MSI tumors fail to respond to immune checkpoint therapy4. To uncover alternative therapeutic vulnerabilities for these patients, we used transcriptome signature-guided approaches to identify MLN4924 (pevonedistat), a Nedd8-activating enzyme inhibitor, as a potential therapy for dMMR/MSI cancers. We discover that destabilizing mutations from the dMMR mutation process lead to rampant proteome instability in MSI tumors, resulting in an abundance of misfolded protein aggregates. To compensate, MSI cancer cells activate a Nedd8-mediated degradation pathway to facilitate clearance of misfolded proteins, which is blocked by treatment with MLN4924. The accumulation of misfolded proteins in MSI cancer cells following MLN4924 treatment activated the unfolded protein response, promoted immune cell migration, and induced immunogenic cell death. Antitumor vaccination with MLN4924-treated cells stimulated the generation of endogenous tumor antibodies and prevented tumor incidence upon re-challenge. Based on this immunostimulation, we combined MLN4924 with PD1 blockade, finding that the combination increased recruitment of CD8+ lymphocytes and improved therapeutic efficacy beyond either treatment alone. Taken together, our results indicate that targeting proteome instability is a novel therapeutic avenue for MSI patients and may potentiate immune checkpoint blockade, potentially increasing the depth and duration of response, as well as the fraction of dMMR/MSI patients who can benefit.

A- A+

PgmNr 112: Genome wide association analysis in a *Drosophila* model of *NGLY1* deficiency identifies NKCC1 as a modifier of disease and a novel NGLY1 substrate in *Drosophila* and mouse.

Authors:
E. Coelho [1]; D. Talsness [1]; K. Owings [1]; K. Peralta [1]; G. Mercenne [2]; J. Pleinis [2]; A. Zuberi [3]; C. Lutz [3]; A. Rodan [2]; C.Y. Chow [1]

View Session Add to Schedule

Affiliations:
1) Human Genetics, University of Utah, Salt Lake City, UT; 2) Internal Medicine, University of Utah, Salt Lake City, UT; 3) The Jackson Laboratory, Bar Harbor, ME

Autosomal recessive loss-of-function mutations in *N-Glycanase 1* (*NGLY1*) cause *NGLY1* deficiency, the only known congenital disorder of deglycosylation. NGLY1 typically deglycosylates glycoproteins that have been retrotranslocated from the ER lumen to the cytoplasm, as a way of regulating the substrate. Patients with *NGLY1* deficiency present with a variety of symptoms, some of which are developmental delay, movement disorder, seizures, liver dysfunction, and alacrima. The severity and combination of symptoms vary widely between patients, which is striking given that all identified patients carry two null mutations. To understand what may be underlying inter-individual differences in severity of *NGLY1* deficiency, we performed a natural genetic variation screen. A *Drosophila* model of *NGLY1* deficiency was crossed into 200 strains from the *Drosophila* Genetic Reference Panel (DGRP), a collection of wild-derived *Drosophila* strains. The effect of natural genetic variation in modulating the loss of NGLY1 phenotype was quantified by counting the number of mutant flies that survived to adulthood. Across the 200 strains, there was a large phenotypic spectrum in the absence of *NGLY1* from nearly 100% survival to complete lethality. A genome wide association analysis identified natural polymorphisms that modified survival, which generated a list of ~60 candidate genes, including several components of the endoplasmic reticulum associated degradation (ERAD) pathway and several glycoproteins. The most significant hit with a human ortholog was *Drosophila Ncc69* (human *NKCC1/2*), a well conserved plasma membrane ion co-transporter. *Drosophila* functional genetic analysis reveals epistatic interactions between *Ncc69* and *NGLY1* for lethality and seizure phenotypes. Subsequently we have shown that NKCC1 is improperly deglycosylated in *NGLY1* null mouse embryonic fibroblasts (MEFs), indicating NKCC1 is a direct substrate. NKCC1 is only the second identified NGLY1 substrate. Further analysis in the *NGLY1* knockout MEFs reveals a significant decrease in NKCC1 function, as measured by Rb^+ ion flux. The misregulation of this ion channel in *NGLY1* deficiency patients may help to explain symptoms such as alacrima, seizures, and liver dysfunction. This study demonstrates the power of *Drosophila* to identify conserved elements of *NGLY1* deficiency and other rare diseases that may serve as targets for personalized therapies.

A- A+

PgmNr 113: Identification of post-translational regulators of MeCP2 levels as potential therapeutic targets for MECP2 duplication syndrome.

Authors:
M. Zaghlula[1,2]; J.-Y. Kim[2,3]; L. Nitschke[2,4]; H.H. Jeong[2,3]; C.E. Alcott[2,5,6]; J.-P. Revelli[2]; Z. Liu[2,3]; S.J. Elledge[7,8]; H.Y. Zoghbi[1,2,3,4,5,6,9]

View Session Add to Schedule

Affiliations:
1) Program in Translational Biology and Molecular Medicine, Baylor College of Medicine, Houston, TX.; 2) Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX.; 3) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.; 4) Integrative Program in Molecular and Biomedical Sciences, Baylor College of Medicine, Houston, TX.; 5) Program in Developmental Biology, Baylor College of Medicine, Houston, TX.; 6) Medical Scientist Training Program, Baylor College of Medicine, Houston, TX.; 7) Department of Genetics, Harvard Medical School, Boston, MA.; 8) Howard Hughes Medical Institute, Harvard Medical School, Boston, MA.; 9) Howard Hughes Medical Institute, Baylor College of Medicine, Houston, TX.

Advances in clinical sequencing continue to highlight the involvement of dosage-sensitive genes in the pathogenesis of neurological disorders. Maintenance of the levels of these proteins within a narrow range is crucial for proper brain function. Methyl-CpG-binding protein 2, *MECP2*, is one such gene: loss-of-function mutations in *MECP2* cause Rett syndrome (RTT) while duplications spanning the *MECP2* locus cause *MECP2* Duplication syndrome (MDS). Both disorders are severe and progressive; current treatments are restricted to symptom management. Importantly, normalization of MeCP2 abundance has been shown to rescue disease phenotypes in mouse models of both disorders. As specific post-translational modifications of proteins can be determinants of protein stability, we set out to identify post-translational regulators of MeCP2 that can be targeted therapeutically.

To this end, we performed cell-based siRNA and CRISPR screens in a reporter system that allows us to monitor changes in MeCP2 levels. From these screens we selected two promising candidates, the atypical kinase RIOK1 and the ubiquitin protease USP1, to perform mechanistic and genetic interaction studies.

First, we confirmed that shRNA-mediated knockdown of *RIOK1* decreases endogenous levels of MeCP2 in HEK293T cells and show nuclear interaction of these proteins in cells. Next, we developed a loss-of-function mouse model of *Riok1* and found that MeCP2 levels are reduced in heterozygous knockout mice by 15%. As this effect is relatively modest, we hypothesized that one of its orthologs, *RIOK2*, might play a compensatory role. As we found that knockdown of *RIOK2* reduces MeCP2 levels in cells, we are now following up with *in vivo* to test further hypotheses about the genetic relationship between the RIO Kinases and MeCP2.

Our studies with the deubiquitinating enzyme USP1 revealed that genetic as well as pharmacological targeting of this screen candidate are able to reduce MeCP2 stability robustly in cells. We also found that knockdown of *Usp1* in the mouse brain lowers MeCP2 levels by 30%. Ongoing studies aim to elucidate whether USP1 directly targets MeCP2 for ubiquitin proteolysis and whether the strong reduction in MeCP2 protein stability in vivo is sufficient to ameliorate disease phenotypes in a mouse model of *MECP2* Duplication Syndrome.

Overall, this screening approach is a powerful and translationally invaluable path to identify drug targets in the treatment of many dosage-sensitive disorders.

A- A+

PgmNr 114: Association of exome sequence variation with blood lipids in 170,000 individuals across multiple ancestries.

Authors:
G. Hindy [1,2]; J. Flannick [1,3,4]; P. Natarajan [1,5,6]; M. Chaffin [1]; A.V. Khera [1,5,6]; G.M. Peloso [7]; on behalf of the AMP-T2D-GENES and TOPMed Lipids Working Groups

View Session Add to Schedule

Affiliations:
1) Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA; 2) Department of Clinical Sciences, Lund University, Malmö, Sweden; 3) Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts, USA; 4) Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA; 5) Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA; 6) Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA; 7) Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA

For genetic association studies, whole-exome sequences provide two distinct advantages: a focus on rare variants in coding sequence and the ability to pinpoint a causal gene. Here, we assembled blood lipid phenotypes and exome sequences in >170,000 participants from multiple ancestries emerging from four data sources: The Myocardial Infarction Genetics Consortium (n=44,208), the AMP-T2D-GENES exomes (n=32,486), the Trans-Omics for Precision Medicine (TOPMed) project restricted to the exome (n=44,101) and UK Biobank exomes (n=51,275). Our study included individuals of African-American (n=16,507), East Asian (n=10,420), European (n=97,493), Hispanic (n=16,440), Samoan (n=1,182) and South Asian (n=30,025) ancestries. We performed extensive quality control within each data source and additionally removed duplicates and first- and second-degree relatives across the data sources. We annotated ~24.5 million variants and selected ~580,000 loss-of-function (LOF) variants and over 1.1 million missense variants predicted to be damaging. We performed association analysis within each data source by ancestry and case-status using linear mixed models incorporating relatedness for 6 traits (total, LDL, HDL, and non-HDL cholesterol and triglycerides and triglyceride-HDL ratio). Gene-based meta-analysis across data sources was performed in all subjects, by race, and by case-status using RAREMETALS. We replicated known associations with LDL cholesterol (*PCSK9*, *LDLR*, and *APOB*), triglycerides (*APOC3*, *ANGPTL3* and *ANGPTL4*), and HDL cholesterol (*CETP* and *SCARB1*). We additionally identified several novel genes associated with blood lipids including *SRSF2* with total cholesterol (-29 mg/dL, $p = 7\times10^{-8}$), *ALB* with total (33 mg/dL, $p = 1\times10^{-8}$) and LDL cholesterol (30 mg/dL, $p = 8\times10^{-9}$), *NR1H3* with HDL cholesterol (2.1 mg/dL, $p = 1\times10^{-9}$), *PLA2G12A* with HDL cholesterol (-2.3 mg/dl, $p = 8\times10^{-14}$) and triglycerides (6%, $p = 1\times10^{-8}$) and *CREB3L3* with triglycerides (12%, $p = 5\times10^{-14}$). We observed significant heterogeneity ($p < 0.002$) among different ancestries for *LDLR* and *APOB* with LDL cholesterol and *LIPG* and *ABCA1* with HDL cholesterol. Our findings demonstrate the importance of large exome sequencing studies in multiple ancestries for the discovery of novel genes associated with blood lipids that may be valuable targets for cardiovascular disease prevention.

A- A+

PgmNr 115: Trans-ethnic meta-analysis of cholesterol and triglyceride levels from 1.6 million individuals.

Authors:
S.E. Graham [1]; G.J.M. Zajac [2]; I. Surakka [1]; X. Li [3,4]; E. Marouli [5]; I. Ntalla [5]; S. Ramdas [6]; X. Yin [2]; X. Zhu [7,8]; T.L. Assimes [8,9]; C.D. Brown [6]; P. Deloukas [5]; A.P. Morris [10]; Y.V. Sun [11,12,13]; G. Abecasis [2]; G. Peloso [14]; S. Kathiresan [15,16,17,18]; C.J. Willer [1,19,20]; Million Veteran Program, Global Lipids Genetics Consortium

View Session Add to Schedule

Affiliations:
1) Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI; 2) Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI; 3) Institute for Translational Genomics and Population Sciences, Lundquist Institute, Harbor-UCLA, Torrance, CA; 4) Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA; 5) William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK; 6) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 7) Department of Statistics, Stanford University, Stanford, CA; 8) VA Palo Alto Health Care System, Palo Alto, CA; 9) Department of Medicine, Stanford University School of Medicine, Stanford, CA; 10) School of Biological Sciences, University of Manchester, Manchester, UK; 11) VA Atlanta Healthcare System, Decatur, GA; 12) Department of Epidemiology, Emory Rollins School of Public Health, Atlanta, GA; 13) Department of Biomedical Informatics, Emory School of Medicine, Atlanta, GA; 14) Department of Biostatistics, Boston University School of Public Health, Boston, MA; 15) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 16) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA; 17) Department of Medicine, Harvard Medical School, Boston, MA; 18) Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA; 19) Department of Human Genetics, University of Michigan, Ann Arbor, MI; 20) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

Elevated cholesterol greatly increases the risk of cardiovascular disease. Blood lipid levels are moderately heritable (~50%) and approximately 400 loci influencing lipid levels have been identified to date. To identify additional genetic regions influencing lipid levels, we have performed a trans-ethnic meta-analysis of cholesterol and triglyceride levels from >150 cohorts, including UK Biobank and the Million Veteran Program, as part of the Global Lipids Genetics Consortium. Approximately 20% of individuals (325,000) are of non-European ancestry. Each cohort performed imputation using 1000 Genomes Phase 3 and/or the Haplotype Reference Consortium reference panels and generated GWAS summary statistics for triglycerides and HDL, LDL, nonHDL, and total cholesterol. In addition, men and women were analyzed separately to identify sex-specific associations. We identify ~750 loci from trans-ethnic meta-analysis of 1.6 million individuals, including over 300 novel loci associated with one or more of these traits from single variant analysis. In addition, we are performing conditional analysis and group-based tests using covariance matrices collected from each cohort to identify independently associated variants and additional genes. The relatively large proportion of non-European individuals in the present study allows for comparison of local LD structure within these loci which facilitates fine-mapping and aids in the identification of likely causal variants. Prioritized genes within these regions may help to identify potential new drug targets as well as further

understand the biology underlying lipid levels. Furthermore, genetic risk scores constructed from associated variants are predictive of lipid levels and show pleiotropic associations with other diseases including coronary atherosclerosis and myocardial infarction, helping to identify individuals most at risk of cardiovascular disease and related disorders.

A- A+

PgmNr 116: Integrated statistical and molecular analysis of a missense variant in the *PROCR* gene that increases risk for venous thromboembolism but protects against coronary artery disease.

Authors:
D. Stacey [1]; L. Chen [1]; J.M.M. Howson [1]; A.M. Mason [1]; S. Burgess [2]; S. MacDonald [3]; J. Langdown [3]; H.L. McKinney [4,5,6]; K. Downes [4,5,6]; N. Farahi [7]; C. Summers [7]; J. Danesh [1]; D.S. Paul [1]; INVENT, ARIC study, INTERVAL study

View Session Add to Schedule

Affiliations:
1) MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; 2) MRC Biostatistics Unit, University of Cambridge, Cambridge, UK; 3) Department of Haematology, Addenbrooke's Hospital, Cambridge, UK; 4) Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, UK; 5) National Health Service Blood and Transplant (NHSBT), Cambridge Biomedical Campus, UK; 6) NIHR BioResource, Cambridge University Hospitals, Cambridge Biomedical Campus, UK; 7) Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK

Genome-wide association studies with increasingly large sample sizes are revolutionizing our knowledge of the genetic aetiology of complex diseases and traits. To translate this knowledge into biological understanding and to aid in the prioritization of therapeutic targets, follow-up mechanistic studies have become a major priority. The minor G allele of a missense variant (rs867186; S219G) in the *PROCR* gene, which encodes the endothelial protein C receptor (EPCR), has previously been associated ($p<5\mathrm{x}10^{-8}$) with increased venous thromboembolism (VTE) risk but decreased coronary artery disease (CAD) risk. Here we elucidate the mechanisms underlying the association between rs867186 and vascular diseases.

In a phenome-scan of rs867186, we found that the divergent pattern of genetic associations with VTE and CAD also applies more generally across venous and arterial diseases, with the G allele increasing and decreasing risk, respectively. The G allele was also associated with higher protein C (PC) and Factor VII (FVII) levels, but not other vascular intermediate traits (e.g., lipids, circulating proteins). Using a novel statistical colocalization approach, we identified rs867186 as the most likely causal variant at this locus for PC and FVII levels, as well as CAD and VTE risk (posterior probability: 0.93). Furthermore, through Mendelian Randomization analyses, we show that PC is a causal factor in both CAD ($p=4.61\mathrm{x}10^{-26}$) and VTE ($p=2.86\mathrm{x}10^{-23}$).

To further explore the impact of rs867186 on the PC pathway, we also performed a recall-by-genotype study ($n=52$ healthy volunteers). We found the minor G allele to be associated with increased plasma soluble EPCR levels ($p=3.29\mathrm{x}10^{-22}$), with a ~2-fold increase for every G allele. This is indicative of increased shedding of membrane EPCR in G allele carriers. We also replicated the association between the G allele and higher PC levels ($p=9.48\mathrm{x}10^{-5}$). Activated PC and thrombin-antithrombin complex levels were not associated with rs867186 genotype ($p>0.05$). Finally, experiments to assess a possible role for soluble EPCR in monocyte-endothelial cell adhesion are ongoing.

In conclusion, we provide first evidence indicating that the PC pathway is a key mediator of the rs867186 genotypic associations with vascular disease. We propose that this pathway is responsible for the divergent pattern of genotypic associations with venous and arterial disease through modulation of coagulation and inflammatory mechanisms, respectively.

A- A+

PgmNr 117: Kidney genes as drivers of heritable predisposition to hypertension: Multi-omic analysis of loci from genome-wide association studies.

Authors:
J.M. Eales [1]; X. Jiang [1]; X. Xu [1]; S. Saluja [1]; E. Cano-Gamez [2]; H. Guo [3]; G. Trynka [2]; A.P. Morris [4]; N.J. Samani [5]; F.J. Charchar [6]; M. Tomaszewski [1]; TRANSLATE consortium

View Session Add to Schedule

Affiliations:
1) Division of Cardiovascular Sciences, University of Manchester, Manchester, UK; 2) Wellcome Sanger Institute, Hinxton, UK; 3) Division of Population Health, Health Services Research & Primary Care, University of Manchester, Manchester, UK; 4) University of Manchester, Manchester, UK; 5) Department of Cardiovascular Sciences, University of Leicester, Leicester, UK; 6) Cardiovascular Genomics, Federation University, Ballarat, Australia

---

Blood pressure (BP) is a complex polygenic trait. Over 800 common genetic variants have been identified in genome-wide association studies (GWAS) of BP and hypertension. The target genes and mechanisms through which these variants operate remain elusive partly because the key tissue of relevance to BP regulation has not been exploited in follow-up omics-style studies. We examined the effects of 821 common BP GWAS variants on the kidney transcriptome, spliceome and epigenome using a unique collection of 430 human kidneys from the TRANScriptome of renaL humAn TissuE (TRANSLATE) collection. We identified 418 kidney genes as kidney expression partners for the BP GWAS variants. Using a human kidney single-cell dataset we detected enrichment in expression of the BP GWAS eGenes in cells from the proximal tubule, consistent with the key role of this segment of the nephron in sodium reabsorption and pressure natriuresis. Through intersection of BP GWAS loci with the kidney spliceome and DNA methylome we detected 319 and 581 BP GWAS kidney sGenes and mGenes (respectively). Through colocalisation analysis we then demonstrated that in 31% of BP GWAS loci the kidney quantitative trait locus (QTL) and GWAS association signals colocalise to shared genetic variants. Further two-stage Mendelian randomisation analysis conducted on loci with documented colocalisation signals revealed a total of 88 unique genes with a causal link to BP. Of those, 30 genes are causal to BP through renal expression, 34 - through alternatively spliced isoforms and 46 - through kidney DNA methylation. Only 10 of the kidney genes with a causal effect on BP (i.e. endothelin-1 gene – *END1* and the voltage-dependent calcium channel beta 4 subunit - *CACNB4*) are recognised contributors to human hypertension. A majority of these genes map onto pathophysiologically novel pathways such as mitochondrial respiration and energy metabolism, intracellular bio-degradation of proteins or tryptophan metabolism. Through integration of multi-omics data from a large collection of human kidney samples we colocalised GWAS association and multi-omic QTL signals in 31% of known BP GWAS loci and uncovered 98 kidney genes with a causal effect on BP.

A- A+

PgmNr 118: Mendelian randomization supports causal associations between lipids and adiposity in non-alcoholic fatty liver disease.

Authors:
K.A.B. Gawronski [1]; M. Vujkovic [2]; M. Serper [3,4,5]; D. Kaplan [3,4]; R. Carr [3,4]; K. Lee [23]; Q. Shao [6]; D. Miller [7]; P. Reaven [8]; L.S. Phillips [9,10]; C. O'Donnell [11,12]; J.B. Meigs [13,14]; P. Wilson [10]; T. Assimes [15,16]; Y.V. Sun [10,17,18]; J. Huang [19]; K. Cho [12,19]; J. Lee [15,16]; P. Tsao [15,16]; D. Rader [1,4,20,21]; C. Brown [1]; S. Damrauer [3,22]; D. Saleheen [2,3]; J. Lynch [23,24]; K. Chang [3,4]; B.F. Voight [1,3,25]

View Session Add to Schedule

Affiliations:
1) Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 2) Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 3) Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA; 4) Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 5) Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA; 6) Edith Nourse Rogers Memorial Veterans Hospital, 20024, Center for Healthcare Organization & Implementation Research, Bedford, MA; 7) Department of Health Law, Policy, and Management, Boston University School of Public Health, Boston, MA; 8) Phoenix VA Health Care System, Phoenix, AZ; 9) Department of Veterans Affairs, Atlanta Healthcare System, Decatur, GA; 10) Emory University School of Medicine, Atlanta, GA; 11) VA Boston Healthcare, Section of Cardiology and Department of Medicine, Boston, MA; 12) Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 13) Massachusetts General Hospital and Harvard Medical School, Boston, MA; 14) Broad Institute, Cambridge, MA; 15) VA Palo Alto Health Care System, Palo Alto, CA; 16) Department of Medicine, Stanford University School of Medicine, Stanford, CA; 17) Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA; 18) Department of Biomedical Informatics, Emory University, Atlanta, GA; 19) Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA; 20) Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 21) Cardiovascular Institute, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 22) Department of Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 23) VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, Salt Lake City, UT; 24) University of Massachusetts College of Nursing and Health Sciences, Boston, MA; 25) Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Nonalcoholic fatty liver disease (NAFLD) is a heritable disorder characterized by fat accumulation in the liver not caused by alcohol consumption. Although the progression of NAFLD varies, it often leads to inflammation and can result in cirrhosis and liver cancer. Approaches to prevent or treat NAFLD require knowledge of causal risk factors for disease progression. Epidemiological studies have identified several metabolic traits that are associated with NAFLD, but it is unknown if these traits directly contribute to NAFLD progression. Mendelian randomization (MR) can be used to evaluate causal hypotheses using genetic data. However, application of MR has been hindered by the lack of genome-wide association data for NAFLD. The Million Veteran’s Program (MVP) has recently

conducted the largest NAFLD genome-wide association study to date, permitting us to apply MR to evaluate the causal role of cardiometabolic traits in NAFLD etiology.

The MVP is a longitudinal health research initiative collecting genetic data that is connected to electronic health records (eHR) from US military veterans. The MVP has developed and validated an eHR-based definition of NAFLD: elevated alanine aminotransferase enzyme levels for two or more time points at least 6 months apart, and within a two-year period prior to enrollment. Individuals with other potential causes of liver disease and alcohol-use disorder were excluded. This definition resulted in 46,653 NAFLD cases and 101,701 controls of European ancestry. Using these data, we utilized MR to evaluate a preliminary set of causal hypotheses for 4 blood lipid levels and 2 adiposity-related cardiometabolic traits with eHR-defined NAFLD. Using inverse variance weighted MR, we observed that a 1-SD genetically-driven increase in body mass index (BMI), waist-hip ratio adjusted for BMI (WHRadjBMI), or reduction in high-density lipoprotein cholesterol (HDL-C) were each associated with increased risk of NAFLD (BMI: OR=1.33, 95% CI=1.24-1.42, P=8.0x10^{-16}; WHRadjBMI: OR=1.55, 95% CI=1.39-1.72, P=6.4x10^{-16}; HDL-C: OR=1.26, 95%CI=1.13-1.40, P=3.3x10^{-5}). These associations remained robust after univariate sensitivity analyses including weighted-median and Egger regression.

This is the largest study to date providing evidence supporting the hypothesis that lipid and adiposity measures are causal risk factors for NAFLD. These results indicate that genetic studies in MVP are poised to rapidly increase our understanding of NAFLD etiology.

PgmNr 119: Mendelian randomization analyses reveal a causal effect of thyroid function on cardiovascular risk factors and diseases.

Authors:
E. Marouli [1,2]; A. Kus [3,4,5]; F. Del Greco [6]; L. Chaker [3,4]; R. Peeters [3,4]; A. Teumer [7,8]; M. Medici [3,4,9]; P. Deloukas [1,2]

View Session Add to Schedule

Affiliations:
1) William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK; 2) Centre for Genomic Health, Life Sciences, Queen Mary University of London, London, UK; 3) Academic Center for Thyroid Diseases, Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; 4) Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; 5) Department of Internal Medicine and Endocrinology, Medical University of Warsaw, Warsaw, Poland; 6) Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lubeck, Bolzano, Italy; 7) Institute for Community Medicine, University Medicine Greifswald, Germany; 8) DZHK (German Center for Cardiovascular Research), partner site Greifswald, Greifswald, Germany; 9) Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands

---

Despite progress in prevention and treatment over the past two decades, cardiovascular disorders remain a leading cause of mortality worldwide. Several observational studies suggest that even minor variation in thyroid function is associated with atherosclerotic cardiovascular disease, type 2 diabetes (T2D), hypertension, dyslipidaemia, and obesity. This raises the question as to whether common forms of mild thyroid dysfunction need treatment to prevent such complications.

Through two-sample Mendelian randomization approaches, we investigated whether the relationship between thyroid function (interrogated through variation in reference range TSH and FT4, subclinical hypo- and hyperthyroidism, Hashimoto’s Disease (HD) and Graves’ disease) and cardiometabolic risk is causal, and possible mediation pathways underlying it.

Using both published GWAS results and data from the UK Biobank, we show that one standard deviation (SD) increase in TSH levels causes a 5% decrease in the risk of stroke (OR=0.95, 95% CI= 0.91 to 0.99). Multivariable MR analyses indicated that this effect is mediated through atrial fibrillation (AF). While normal range thyroid function is not associated with Coronary Artery Disease (CAD), HD was associated with a 7% increased risk of CAD (OR=1.07, 95% CI= 1.01 to 1.13). The effect of HD on CAD risk appears to be mediated via body mass index (BMI).

Looking at cardiovascular risk factors, a one SD increase in TSH levels was causally associated with: an increase of 0.05 SD units in total cholesterol serum levels (β=0.05, 95% CI= 0.02 to 0.08), a 0.03 SD unit increase in low-density lipoprotein (β=0.03, 95% CI= 0.003 to 0.02), and a 0.31 mmHg decrease in pulse pressure (PP) (β=-0.31, 95% CI= -0.54 to -0.08). In line with these findings, subclinical hyperthyroidism is causally associated with increased PP (β=0.15, 95% CI= 0.05 to 0.25). There is no evidence for a causal association between normal range FT4 levels and the tested outcomes.

These results establish that minor variation in normal range thyroid function can be a novel modifiable risk factor for stroke through its effect on AF. Furthermore, variation in thyroid function can affect cardiovascular risk *via* its effects on cholesterol levels and blood pressure. These findings pave the way to consider future adjustment of thyroid function in managing patients' risk of stroke.

A- A+

PgmNr 120: Multi-tissue analysis reveals short tandem repeats as ubiquitous regulators of gene expression and complex traits.

Authors:
M. Gymrek; S. Feupe Fotsing; C. Wang; S. Saini; R. Yanicky; S. Shleizer-Burko; A. Goren

View Session Add to Schedule

Affiliation: Univ California San Diego, La Jolla, California.

Short tandem repeats (STRs) have been implicated in a variety of complex traits in humans. However, genome-wide studies of the effects of STRs on gene expression thus far have had limited power to detect associations and elucidate the underlying biological mechanisms. Here, we leverage whole genome sequencing and expression data for 17 tissues from the Genotype-Tissue Expression Project (GTEx) to identify STRs whose repeat lengths are associated with expression of nearby genes (eSTRs). Fine-mapping analysis reveals more than 3,000 high-confidence eSTRs, which are enriched in known or predicted regulatory regions. We show eSTRs may act through a variety of mechanisms, including controlling nucleosome positioning (homopolymers), altering affinity of transcription factor binding sites (dinucleotides), and modulating DNA or RNA secondary structure (GC-rich promoter repeats). We further apply co-localization analysis to identify hundreds of eSTRs that potentially drive published GWAS signals and implicate specific eSTRs in height, schizophrenia, and blood traits. For example, we identified a dinucleotide STR near the 3’ end of the gene *RFT1* as a potential causal variant for height. To validate this finding, we imputed the STR into an independent cohort (eMERGE) and found a positive association between repeat number and height (p=0.0032). We additionally performed a dual reporter assay to test the effect of this STR *in vitro* and recapitulated the expected positive association between repeat number and expression (p=0.013). Overall, our results demonstrate that eSTRs potentially contribute to a range of human phenotypes. We expect that our comprehensive eSTR catalog will serve as a valuable resource for future studies of complex traits. Complete eSTR summary statistic data is publicly available and can be browsed interactively at webstr.gymreklab.com.

A- A+

PgmNr 121: Cellular deconvolution of GTEx tissues powers eQTL studies to discover thousands of novel disease and cell-type associated regulatory variants.

Authors:
M. Donovan [1]; A. D’Antonio-Chronowska [2]; M. D’Antonio [2]; K. Frazer [2,3]

View Session Add to Schedule

Affiliations:
1) Bioinformatics and Systems Biology Graduate Program, UC San Diego, La Jolla, California.; 2) Institute for Genomic Medicine, UC San Diego, La Jolla, California.; 3) Pediatrics, UC San Diego, La Jolla, California.

The Genotype-Tissue Expression (GTEx) resource has contributed a wealth of novel insights into the regulatory impact of genetic variation on gene expression across tissues types, however thus far not been utilized to study how variation acts at the resolution of the diverse cell types composing the tissues. Cell-type-specific gene expression signatures needed to deconvolute heterogenous tissues can be obtained by analyzing single-cell RNA-seq (scRNA-seq) generated from an analogous tissue. However, there are relatively few human scRNA-seq resources currently available, and thus only a small fraction of GTEx tissues can be deconvoluted using existing human scRNA-seq data. To address this gap, using liver as a proof-of-concept tissue, we show that mouse scRNA-seq can be used as an alternative to human scRNA-seq for the cellular deconvolution of GTEx tissues. We then used mouse scRNA-seq from the Tabula Muris resource to deconvolute 28 GTEx tissues corresponding to 14 organs (over 6,000 bulk RNA-seqs). We showed that there were between two (bladder) to seven (brain and heart) different cell types within each of the 28 GTEx tissues and that the proportion of these cell types varied substantially across samples of the same tissue type. In two deconvoluted tissues (GTEx liver and skin), we performed eQTL analyses considering estimated cellular compositions to identify cell-type-associated eQTLs. We found that considering the relative cell population distributions estimated as covariates for eQTL analyses identifies thousands of cell-type-associated genetic associations with lower effect sizes not detected at the tissue-level. To explore the functional impact of the cell-type-associated eQTLs identified in skin, we examined their overlap with GWAS lead variants for skin traits and disease. We found that cell-type-associated eQTLs in skin commonly colocalize with variants in GWAS loci for melanoma, malignant neoplasm, and infection signatures. We then functionally characterized six eQTLs significantly associated with leukocytes and colocalized with malignant skin neoplasms that regulate genes (*TCF19, ATAD3C, SERPINB9, NT5C2,* and *CD1E)* known to play a role in cancer progression or immune response. Our study provides a framework to deconvolute the cellular composition of bulk RNA-seq from GTEx, which can be implemented immediately for characterizing the functional impact of cell-type-associated genetic variation on complex traits and disease.

A- A+

PgmNr 122: Does sex-specificity of eQTLs propagate to complex traits?

Authors:
E. Porcu [1,2]; A. Claringbould [3]; K. Lepik [2,4,5]; R. Jansen [6]; L. Franke [3]; F.A. Santoni [7]; A. Reymond [1]; Z. Kutalik [2,4]; BIOS Consortium

View Session Add to Schedule

Affiliations:
1) Center of Integratve Genomics, Lausanne, Switzerland; 2) Swiss Institute of Bioinformatics, Lausanne Switzerland; 3) University Medical Centre Groningen, Groningen, the Netherlands; 4) University Center for Primary Care and Public Health, Lausanne, Switzerland; 5) Institute of Computer Science, University of Tartu, Tartu, Estonia; 6) Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Psychiatry, Amsterdam Neuroscience, the Netherlands; 7) Endocrine, Diabetes, and Metabolism Service, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland

---

In spite of the fact that the prevalence of many diseases differs between women and men, only few published genome-wide association studies (GWAS) were performed in a sex-stratified manner. To understand the genetic basis of sexual dimorphism, we investigated whether sex-biased eQTLs translate to sex-biased trait associations and searched for sex-specific expression-trait causal effects.
To investigate the possible role of eQTLs in sex differences, we performed a genome-wide analysis of sex-specific whole blood RNA-seq eQTLs from 3,500 individuals. Among pre-selected 9 million SNP-gene pairs, we identified 460 SNP-gene associations showing significantly different effects in men and women (FDR<0.05). These sex-biased eQTLs cluster in 17 genes, 11 of which are female-biased, 5 are male-biased and *ZNF718* has both types of eQTLs. Performing PheWAS analyses for the 460 significant SNPs on >700 traits, we found that sex-biased eQTLs in *CDIP1* and *PSMD5* translate into sex-specific trait-associations for trunk predicted mass. However, such examples are sporadic and sex-specific expression regulation does not systematically propagate to high-level traits. This implies that there should exist a sex-specific compensatory mechanism downstream the gene expression.
To explore this hypothesis, we applied a sex-specific transcriptome-wide Mendelian Randomization approach (TWMR) combining sex-specific summary statistics for both eQTLs and 23 complex human traits. In total, we discovered 48 gene-trait associations showing significantly different causal effect in men and women. We observe new female-specific effect of *SCAP* expression on waist-hip ratio ($P_{TWMR\text{-}females}$=6.01E-13 and $P_{TWMR\text{-}males}$=0.71) and *HCAR3* (2.50E-14 and 0.95). Notably, 37 of the 48 sex-specific associations are not identified in TWMR using sex-combined GWAS and (even larger n=?31,684) eQTL data, suggesting that those loci harbor variants exerting effects in a sex-dependent manner.
Our findings emphasize the importance of sex-stratified analyses, as they can improve power in identifying associations and are essential for a better understanding of genetic basis of sexual dimorphism hence contributing to precision medicine.

A- A+

PgmNr 123: Causal mediation analysis identifies a comprehensive map of gene-level polygenicity and pleiotropy across 43 traits in 49 tissues.

Authors:
Y. Park [1,2]; A.K. Sarkar [3]; L. He [4]; K.T. Nguyen [1,2]; N. Daskalakis [5,6]; M. Kellis [1,2]; GTEx Consortium

View Session Add to Schedule

Affiliations:
1) Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Department of Human Genetics, University of Chicago, Chicago, IL; 4) Social Science Research Institute, Duke University, Durham, NC; 5) Neurogenomics and Translational Bioinformatics Laboratory, McLean Hospital, Belmont, MA; 6) Department of Psychiatry, Harvard Medical School, Boston, MA

Genome-wide association studies of hundreds of phenotypes have revealed that (1) complex traits are polygenic, (2) pleiotropy is pervasive, and (3) 90% of the significant associations implicate non-coding regions. These features make finding causal genes for phenotypes of interest difficult. State of the art computational methods combine the evidence of GWAS and eQTL data to prioritize candidate genes. However, these methods can be confounded by gene-gene correlations (leading them to prioritize non-causal genes) and non-genetic trait-trait correlations (leading to inflated statistics within a region). To overcome these challenges, we developed CaMMEL, a method based on causal inference that adjusts for unmediated confounding effects in eQTL and GWAS summary statistics.

We used our method to jointly analyze gene expression in 49 GTEx tissues and 43 high-powered GWAS (N=13,283-1,331,010, mean 326,427) and identified ~154 protein-coding genes that causally mediate the effect of genetic variation on phenotype. Importantly, our method did not find evidence of causal mediation for 54% of significant associations.

Polygenicity mediated by multiple tissues: Total mediated effects aggregated across tissues explain 11-23% of the total genetic variation of each trait, although most enriched tissues within each trait explain average 1-2% of the total variation, suggesting that genetic effects of the human traits are manifested in multiple tissues. Nonetheless, our analysis highlights trait-specific enrichment of the tissues in 30 traits (FDR < 5%), of which examples include brain tissues in Alzheimer’s disease ($p < 1.6e-4$) and hypothyroidism ($p < 1.8e-5$), ovary in rheumatoid arthritis ($p < 5.9e-4$).

Pleiotropy at mediated gene-level: When we aggregate the mediated effects as a polygenic score of each trait-tissue combination, we recapitulated 38 of 42 known pleiotropic pairs are also correlated at mediated gene-level. Of total 73 trait-trait pairs correlated by the mediated genetic effects, 49 of them may not be so evident at the GWAS-level (correlation r=.21 vs. r=.01). At the gene-level, our meditation results suggest that genes are highly pleiotropic, exerting actions on average 13 traits (SD=9) through eQTLs in average 28 tissues (SD=12). Moreover, a strong correlation between the degree of trait- and tissue-pleiotropy (r=.87) suggest that pleiotropic genes tend to participate in a wide range of tissue and cell type-specific mechanisms.

A- A+

PgmNr 124: Identifying the genes and cell types affected by genetic risk loci for primary sclerosing cholangitis.

Authors:
E.C. Goode [1,2]; L. Moutsianas [3]; L. Fachal [1]; T. Raine [4,1]; S.M. Rushbrook [5]; C.A. Anderson [1]

View Session Add to Schedule

Affiliations:
1) Human Genetics, Wellcome Trust Sanger Institute, Cambridge, Cambridgeshire, United Kingdom; 2) Institute of Metabolic Science, University of Cambridge, Cambridge, UK; 3) Genomics England, London, UK; 4) Department of Medicine, University of Cambridge, UK; 5) Norfolk and Norwich University Hospital, Norwich, UK

Primary sclerosing cholangitis (PSC) is an inflammatory disorder of the bile duct that is highly co-morbid with Inflammatory bowel disease (IBD) and leads to liver failure. 23 loci have been associated with PSC risk, the majority in non-coding regions. We aimed to identify the affected genes, cell types and pathways in order to identify potential therapeutic targets for PSC.

We performed Bayesian tests of colocalisation across PSC loci with 13 existing eQTL datasets covering 5 tissue types, 7 immune-cell types, some in >10 activation states and with 8 other immune-mediated diseases (IMDs). We performed fine-mapping of PSC loci to identify a 95% credible set of variants. For loci with >1 variant in this set, we performed fine-mapping within colocalising eQTL data to increase power to detect single PSC causal variants.

11 of 23 PSC loci colocalised with another IMD with a posterior probability >80%. Of these, 7 colocalised with an eQTL in ≥1 cell types. Fine-mapping in PSC resolved 2 loci to single causal variants, and 5 loci to credible sets of ≤10 variants. Fine-mapping of colocalising eQTL identified a further causal variant for PSC and another 2 with ≤10 variants. For example, PSC locus 21q21 colocalised with type 1 diabetes and coeliac risk as well as an eQTL for *UBASH3A* in T-reg and CD4+T-cells. UBASH3A is known to attenuate T-cell NFκB and IL-2 signalling. Fine-mapping of eQTL data reduced the credible set from 5 to a single causal missense intronic variant, with the PSC risk allele decreasing *UBASH3A* expression.

Another locus, 21q11, colocalised with IBD risk and an eQTL for *ETS2* in monocytes and IL-4 stimulated macrophages. Fine-mapping of the eQTL reduced the credible set from 11 to 8 variants. *ETS2* is >500kb upstream of the lead PSC variant, in comparison with the previously proposed candidate gene, *PSMG1* which lies ~100kb downstream. Further evidence for a role of *ETS2* was confirmed by colocalisation between this region in IBD and *ETS2* in the same cell types. *ETS2* potentiates negative selection of T cells, but no co-localisations were found with our T-reg or CD4+ T-cell eQTL data. To address this, we are generating our own eQTL maps from PSC donors, including rare CCR9+ gut-homing T-cells, hypothesised to be pathogenic in PSC, to be presented at the meeting.

Our work confidently identifies causal variants, perturbed genes and cell types for several PSC risk loci and supports further study of eQTL in disease-specific and activated cell types.

A- A+

PgmNr 125: ADAR-mediated editing of a microRNA in bronchial epithelial cells is associated with severe asthma in children of asthmatic mothers.

Authors:
K.M. Magnaye [1]; D.K. Hogarth [2]; E.T. Naureckas [2]; S.R. White [2]; C. Ober [1]

View Session Add to Schedule

Affiliations:
1) Human Genetics, University of Chicago, Chicago, Illinois.; 2) Medicine, University of Chicago, Chicago, Illinois

One of the most consistent risk factors for childhood-onset asthma is having a mother with asthma. A possible explanation may be altered gene regulation *in utero* that is long-lasting. One such mechanism could be through miRNA-mediated gene silencing that differs between children of mothers with and without asthma. Our lab has previously shown that miRNA expression (Nicodemus-Johnson *et al. JACI* 2013; 131:1496) and correlations between miRNAs and their gene targets (unpublished) differ between adult asthma cases with and without an asthmatic mother. Another mechanism that alters miRNA-mediated gene silencing is deamination of adenosine to inosine (A-to-I), or ADAR-mediated editing, of sites within the seed regions of miRNAs. Here, we present the first genome-wide analysis of ADAR-mediated editing using miRNA-seq in primary bronchial epithelial cells (BECs) from 132 asthma cases and controls. Of the 19 A-to-I sites detected, 16 were in seed regions. Four of the 16 edited sites were observed in >10 individuals and were tested for differential editing (% A-to-I) between groups. One site at position 5 of miR-200b-3p was edited less frequently in asthma cases (n = 71) compared to controls (n = 39) ($p_{adjusted}$ = 0.0057). A-to-I editing of this site was then tested for an association with asthma severity (mild, moderate and severe) based on lung function and medication use. Editing of this site was negatively correlated with asthma severity (p = 0.0066). In particular, the severe asthma group (n = 36) had significantly less A-to-I editing of the 5th position of miR-200b-3p compared to controls (n = 39; p = 6.17×10^{-4}). Bioinformatic prediction revealed 210 *in silico* target genes for the edited miR-200b-3p, which were enriched for one pathway: IL4 signaling (p = 5.87×10^{-4}). Finally, we used RNA-seq to test for differential expression (DE) of the two editing enzymes, *ADAR* and *ADARB1*, between asthma cases and controls and between adults with and without a mother with asthma. *ADAR* was significantly downregulated in adults with a mother with asthma (p = 0.0046), but not in asthma cases overall. Together, these results suggest that maternal asthma may alter the gene regulatory landscape in BECs *in utero*, with long-lasting effects on ADAR-mediated editing of the 5th position of miR-200b-3p. We further suggest that this “exposure” results in reduced suppression of IL4 signaling in BECs of asthmatics, leading to more symptoms and more severe asthma.

A- A+

PgmNr 126: Development and psychometric analysis of instrument to measure high school genetics knowledge. GLASS: Genetics Literacy Assessment for Secondary Schools.

Authors:
K. Ormond [1,2]; J. Keyes [1]; R.J. Okamura [1,3]; S. Lee [2,4]; M. Dougherty [5]; B. Domingue [6,7]

View Session Add to Schedule

Affiliations:
1) Human Genetics/Genetic Counseling, Stanford Univ, Stanford, California.; 2) Stanford Center for Biomedical Ethics, Stanford Univ, Stanford, California; 3) Invitae Corp., San Francisco, California; 4) Division of Ethics, Department of Medical Humanities and Ethics, Columbia Univ, New York, New York; 5) Department of Pediatrics, Univ of Colorado, Aurora, Colorado; 6) Graduate School of Education, Stanford Univ, Stanford, California; 7) Stanford Center for Population Health Sciences Stanford Univ, Stanford, California

---

Background: There currently is no established educational assessment to assess the genetics literacy of students at the secondary school level. In 2015, the Genetics Literacy Assessment for Secondary Schools (GLASS) was constructed using The American Society of Human Genetics list of core concepts essential for secondary school genetics literacy. During the 2018-2019 academic year, a multi-state administration and psychometric analysis of the GLASS was completed to establish validity and reliability.

Purpose: The aim of this study was to administer the GLASS to a broad multi-state sample of 12th-grade students with varying degrees of genetics education exposure to assess the psychometric properties of the GLASS.

Methods: We recruited students between Fall 2018 and Spring 2019. We used the Rasch model as a basis for measurement to examine a variety of technical properties of the GLASS, including its dimensionality, analysis of student performance on the GLASS as a function of student demographics, and item fit analysis. We also used the Rasch model to construct a Wright map that offers information about the relative position of items and examinees on the scale.

Results: Two hundred and twenty-three 12th-grade students from Maine (n=13), Connecticut (n=22), California (n=44), Alabama (n=63) and North Carolina (n=81) completed the GLASS. Based on an interim analysis of 179 responses, the GLASS was found to have relatively high reliability (Cronbach’s alpha = 0.78). The GLASS has an appropriate distribution of relatively easy and challenging items as measured by item difficulty indices (0.229 - 0.798; mean = 0.552). Items on the GLASS had reasonable discriminatory ability as calculated by point biserial correlations (0.219 - 0.553; mean = 0.413). A standardized mean difference was calculated (SMD = 1.08) between students who had completed AP/IB/Honors biology coursework and students who had completed general biology coursework. Current results will be updated by the time of the conference to account for the additional forty-four completed assessments.

Discussion: Health-related genetics is becoming an increasingly prevalent paradigm in modern society. Based on our current study, the GLASS is a reasonably reliable and valid measure of genetics literacy that can be used to evaluate the genetics education being provided to their secondary school students in the U.S.

A- A+

PgmNr 127: Preparing medical laboratory scientists for the genomic era: A 5-year comprehensive education strategy towards professional competencies in variant interpretation.

Authors:
N.P. Thorne [1]; A. Nisselle [1,3,5]; S. Lunke [4,5]; A. Fellowes [6]; M. Martyn [1,3,5]; A. Roesley [1,6]; C. McEwen [1,3]; M. Fanjul Fernandez [4]; T.Y. Tan [3,4,5]; D. Liddicoat [1,3]; Z. Stark [4]; E. Thompson [6]; F. Maher [1,3]; F. Cunningham [1]; I. Macciocca [4]; G. Reid [3,5]; P. James [5,7]; J. Hodgson [3,5]; C. Gaff [1,2,3,5]; Diagnostic Advisory Group

View Session Add to Schedule

Affiliations:
1) Melbourne Genomics Health Alliance, Melbourne, Victoria, Australia; 2) Walter and Eliza Hall Institute, Melbourne, Victoria, Australia; 3) Murdoch Children's Research Institute, Melbourne, Victoria, Australia; 4) Victorian Clinical Genetics Service, Melbourne, Victoria, Australia; 5) The University of Melbourne, Melbourne, Victoria, Australia; 6) Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia; 7) The Royal Melbourne Hospital, Melbourne, Victoria, Australia

Variant interpretation (VI) is crucial to genomic testing, yet there is limited research into the educational needs and career pathways of medical laboratory scientists for this. Melbourne Genomics Health Alliance based in Victoria, Australia, adopted a multipronged education strategy to develop a genomic-competent diagnostic workforce. We developed a suite of VI training programs (germline and cancer) based on adult learning theory, culminating in >15,000 learning hours over 5 years to local, national and international professionals. The programs included: workplace immersion experiences (48d cross-laboratory trainees (CLT)); 1–2d professional development workshops (PDW); and two Masters-level subjects. Program evaluation used mixed methods.

At commencement, the majority of CLTs (7/10) were unfamiliar with VI. Mid-point interviews around barriers and facilitators of using immersion to acquire and maintain competency, revealed the need for an initial PDW. A group exit interview (9/10) provided insights into how CLTs implemented learnings within their professional roles, including leading and/or establishing VI processes in their laboratories. PDW participants (n=374) were Medical Scientists (36.1%), Researchers/Bioinformaticians/Students (26.2%), Clinical Geneticists (16.8%), Other Medical (17.3%), Genetic Counsellors (3.1%) and other Allied Health Professionals (0.5%). Pre-post surveys (188/374) and case assessments were analysed to determine actual versus self-assessed capability within and across workshops over time. Average self-assessed understanding of VI increased by 38.4% and 88.7% of participants anticipated incorporating their learning into their professional role. Comparisons of paired pre-post case assessment data confirmed increased capability and highlighted aspects of VI where education was most valuable.

Masters-level subjects (24 modules across six curricula; 44 students to date) were developed based on these programs, and Bloom’s taxonomy was used to align learning outcomes with desired VI competencies. The subjects use a blended learning approach (online modules plus hands-on workshops). Evaluation continues to assess effectiveness of the blended learning approach, and student feedback on content. This work provides a basis for an educational framework and competencies in VI that could be applied across the international medical laboratory scientist workforce in the genomic era.

PgmNr 128: Are hospitals prepared to implement rapid whole-genome sequencing (WGS) in NICU/PICU settings? Healthcare provider perspectives from the frontline.

Authors:
M. Brown [1]; A.M. Li-Rosi [1]; S. Hussain [1]; R. Veith [2]; A. Jorgenson [2]; D. Salyakina [1]; J. McCafferty [1]; K. Schain [1]; A. Quittner [1]; Nicklaus Children's Personalized Medicine Clinical Team

View Session Add to Schedule

Affiliations:
1) Nicklaus Children's Hospital, Miami, Florida.; 2) Children's Hospitals and Clinics of Minnesota

Background: The Personalized Medicine Initiative (PMI) at Nicklaus Children’s Hospital (NCH) was created to implement precision medicine targeted to children. Consequently, rapid whole-genome sequencing (rWGS) has been introduced in the hospital’s intensive care units (neonatal, pediatric, and cardiac) for children with a suspected underlying genetic condition. However, clinical integration at NCH and other hospitals remains dependent on the perceived utility and barriers of rWGS by practicing clinicians. Our specific objective is to develop provider outcome measures that evaluate the process of implementation from the provider perspective and the gaps in knowledge and communication that could improve this process.

Methods: We conducted qualitative, open-ended interviews with clinicians (intensivists, geneticists, genetic counselors) at NCH and Children’s Minnesota who have enrolled and/or treated patients undergoing rWGS. The goal was to understand clinician attitudes, perceived clinical utility, knowledge gaps, and practical/ethical challenges experienced when using rWGS in an intensive care setting. Interviews were audiotaped, transcribed, and coded in NVivo software to identify the most frequent and impactful experiences of clinicians using consensus coding.

Results: To date, 24 interviews have been conducted with intensive care physicians (e.g. pediatric cardiologists, neonatologists) and genetics professionals (physician-geneticists, genetic counselors). Themes emerging from the qualitative data suggest that challenges for both geneticists and non-geneticists include: 1) delivering diagnoses of ultra-rare conditions for which limited literature exists, 2) establishing appropriate expectations for rWGS results for parents, and 3) providing clear, informed consent to parents. Both geneticists and intensivists perceived high clinical utility for rWGS, including diagnostic results that guide medical management and avoidance of unnecessary tests and time in the NICU for some patients. A majority of providers emphasized the critical role geneticists and genetic counselors provide in communicating and interpreting the results for the clinical team and families.

Discussion: Intensivists had concerns about their readiness to use rWGS without considerable support from the genetics team. These findings highlight a potential barrier to widespread adoption of rWGS into practice, with many hospitals lacking geneticists and genetic counselors on staff.

A- A+

PgmNr 129: The development of the clinician-reported genetic testing utility index (C-GUIDE): A novel strategy for measuring the clinical utility of genetic testing.

Authors:
R.Z. Hayeems [1,2]; S. Luca [1]; W.J. Ungar [1,2]; A. Bhatt [1]; L. Chad [3,4]; E. Pullenayegum [1,5]; M.S. Meyn [6]

View Session Add to Schedule

Affiliations:
1) Child Health Evaluative Sciences, Hospital for Sick Children, Toronto, Ontario, Canada; 2) Institute of Health Policy Management and Evaluation, University of Toronto; 3) Division of Clinical and Metabolic Genetics, The Hospital for Sick Children; 4) Department of Pediatrics, University of Toronto; 5) Dalla Lana School of Public Health, University of Toronto; 6) Center for Human Genomics and Precision Medicine, University of Wisconsin School of Medicine and Public Health

Purpose: Clinical utility describes a genetic test's value to patients, families, healthcare providers, systems, or society. While the laboratory performance of genetic tests has improved significantly, policymakers are seeking evidence of clinical value. This study aims to define clinical utility from the perspective of clinicians and develop a novel outcome measure that operationalizes this concept.
Methods: Item selection for the Clinician-reported Genetic testing Utility InDEx (C-GUIDE) was informed by a scoping review of the literature. Item reduction and refinement was guided by qualitative and quantitative feedback from semi-structured interviews and a cross-sectional survey of genetics and non-genetics specialists who routinely use genetic testing. Final item selection, index scoring, and structure was guided by feedback from an expert panel of genetics professionals.
Results: A review of 194 publications informed the selection of a preliminary set of 25 items. Iterative rounds of feedback from 35 semi-structured interviews, 113 completed surveys, and 11 expert panelists informed the content and wording of C-GUIDE's final set of 18 items. These items reflect on the ability of a genetic test to contribute to (i) understanding diagnosis and prognosis, (ii) management decision-making related to sub-specialist care, investigations for diagnostic or surveillance purposes, medication use or surgery, (iii) awareness and actionability of current and future reproductive and health risks for index patients and family members, and (iv) psychosocial well-being. C-GUIDE achieves content and face validity for use in a range of diagnostic genetic testing settings.
Conclusion: Work to establish reliability and construct validity is underway. The development of a standardized, clinimetrically-robust tool for assessing the clinical utility of genetic testing provides a practical strategy for adjudicating the value of rapidly evolving genetic testing technologies. Across a range of clinical settings, C-GUIDE can be used in comparative studies to inform decisions related to clinical adoption and reimbursement.

A- A+

PgmNr 130: Group counseling by webinar: An efficient approach to pre-test counseling.

Authors:
B. Swope; E.S. Gordon; C.A. Fine; L. Myers; S.M. Weissman; J. Bailey; A. Fan; E. Jordan; B. LeLuyer; A.F. Rope; H. Shabazz; S.B. Bleyl

View Session Add to Schedule

Affiliation: Genome Medical, South San Francisco, CA.

Given the shortage of genetic counselors in the United States, there is increasing pressure to identify alternative service delivery models. Group counseling has been used successfully in the past in the in-person setting to provide pre-test counseling, but there is limited data on the success and outcomes related to group pre-test counseling using webinars, both recorded and live. Here we present data for group pre-test counseling offered to a total of 589 patients across six programs;1) a genetic screen intended for healthy individuals, 2) an expanded carrier screening panel, and 3) hereditary cancer testing for patients at increased risk. All three programs were made available to employees through an employee wellness program where both testing and counseling was subsidized (to varying degrees). For all patients, the webinar as well as a medical and family history questionnaire was required in order to proceed to testing. All patients had access to 1:1 genetic counseling throughout the process. The number of program registrants was dictated by the employer and all programs filled within 24 hours of the program opening. The webinars covered the difference between screening and diagnostic testing, the type of information that patients could learn through the genetic testing offered in the specific program, the risks, benefits and limitations of testing, implications of positive results on medical management, implications for family members, and privacy considerations. Of 493 participants who completed the medical history questionnaire, 67% were female. 98% of individuals offered healthy genetic screening chose to pursue it (2% chose diagnostic testing after consultation with a genetic counselor). Among those offered cancer or carrier screening, 93% pursued carrier screening and 70% were eligible for and chose to pursue hereditary cancer screening. 91% of participants agreed or strongly agreed that the webinar answered all of their questions, 90% of participants would recommend the program to a friend or family member, and 83% agreed or strongly agreed that the information provided in the webinar was valuable. Our data, across three different genetic testing programs, suggests that group pre-test counseling via webinar is an efficient and accessible approach to delivering pre-test counseling with a high level of patient satisfaction.

A- A+

PgmNr 131: Missed diagnoses: Clinically relevant lessons learned through cases diagnosed by the Undiagnosed Diseases Network.

Authors:
H. Cope [1]; R. Spillmann [1]; J. Sullivan [1]; J.A. Rosenfeld [2]; E. Brokamp [3]; R. Signer [4]; S. Lincoln [5]; J. Martinez-Agosto [4]; K. Dipple [6]; V. Shashi [1]; Undiagnosed Diseases Network

View Session Add to Schedule

Affiliations:
1) Department of Pediatrics, Duke University Medical Center, Durham, NC.; 2) Baylor College of Medicine, Houston, TX; 3) Vanderbilt University Medical Center, Nashville, TN; 4) UCLA Health, Los Angeles, CA; 5) Boston Children's Hospital, Boston, MA; 6) Seattle Children's, Seattle, WA

The Undiagnosed Diseases Network (UDN) is a collaborative nationwide research effort, funded by the National Institutes of Health Common Fund, tasked to solve the most challenging medical mysteries using advanced technologies. Upon referral from a local health care provider, patients with complex medical conditions and extensive non-diagnostic workups can apply for evaluation at one of twelve UDN clinical sites. Since 2014, the UDN has evaluated over 1,100 patients resulting in diagnoses for over 300 (~30% network-wide diagnostic rate). While many UDN patients are ultimately diagnosed utilizing resources that are difficult to access by clinical providers, such as genome sequencing and model organisms, some patients receive diagnoses through thoughtful application of routine, clinically available methods. We present 13 cases that illustrate how basic clinical considerations, testing strategies, and variant interpretation practices resulted in diagnoses through the UDN. Patients include a 3-year-old female with growth failure, global delays and areas of hypo- and hyper-pigmentation whose pre-UDN workup included negative karyotype, microarray and exome sequencing on blood. A diagnosis of mosaic triploidy was made by the UDN upon karyotype of fibroblasts, demonstrating the importance of selecting an appropriate sample type. Other clinical considerations exemplified by these cases include the importance of considering variable expressivity when forming differential diagnoses and the need to recognize all possible inheritance patterns. Testing strategies include evaluating prior testing and utilizing updated test methodologies when appropriate and recognizing limitations of next generation sequencing in diagnosing copy number variants. Variant interpretation considerations include the importance of not relying solely on OMIM to investigate gene-disease associations and the value of periodic variant re-interpretation. Also illustrated is the importance of checking alternate transcript nomenclature when reading the literature and being mindful that two variants may result in a blended phenotype. These cases collectively demonstrate that careful utilization of existing clinical tools and methods can result in increased diagnostic yield, both within specialized undiagnosed disease programs and in general clinical practice. It is evident that a methodical clinical approach utilizing resources available to practicing clinicians is vital to the diagnostic process.

A - A+

PgmNr 132: Genome-wide association meta-analysis of over 237,000 breast, prostate, ovarian, and endometrial cancer cases and 317,000 controls identifies 128 regions containing associations with multiple cancers.

Authors:
S. Kar[1]; S. Lindström[2,3]; J. Dennis[1]; K. Michailidou[1,4]; R. Hung[5]; D.F. Easton[1]; J. Simard[6]; A. Spurdle[7]; T. O'Mara[7]; R. Eeles[8]; B. Pasaniuc[9]; P. Kraft[10]; P. Pharoah[1]; on behalf of the BCAC, OCAC, ECAC, GAME-ON, PRACTICAL, CAPS and PEGASUS consortia

View Session Add to Schedule

Affiliations:
1) University of Cambridge, Cambridge, UK; 2) University of Washington, Seattle, WA, USA; 3) Fred Hutchinson Cancer Research Center, Seattle, WA, USA; 4) The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus; 5) Lunenfeld-Tanenbaum Research Institute, Sinai Health System and University of Toronto, Toronto, ON, Canada; 6) Centre de recherche du CHU de Québec-Université Laval, Laval, QC, Canada; 7) QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia; 8) The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, London, UK; 9) University of California Los Angeles, Los Angeles, CA, USA; 10) Harvard T.H. Chan School of Public Health, Boston, MA, USA

Cancers of the breast, prostate, ovary and endometrium together accounted for ~22% of all new cancer cases diagnosed and ~1.2 million deaths worldwide in 2018. Pleiotropic germline alleles in genes such as *BRCA1/2* that are associated with susceptibility to more than one of these cancer types have yielded fundamental cross-cancer mechanistic and therapeutic insights. They motivate the identification of additional such alleles, particularly among common variants. We meta-analyzed association data from 122,977 breast, 79,194 prostate, 22,406 ovarian, and 12,906 endometrial cancer cases and 317,006 controls of European ancestry for ~9.5 million SNPs (minor allele frequency >1%). We filtered SNPs that achieved genome-wide significance (*P*<5E-8) on meta-analysis using a Bayesian approach that averaged over a range of genetic architectures to identify SNPs where the association met all of three criteria: 1) the strongest association was due to a combination (>=2 of 4) of cancers 2) the posterior probability of this combined association was >80% 3) there was evidence of association with each of the individual cancers in the combination. This rigorous two-step meta-analysis generated the most comprehensive catalog to date of risk loci shared across the four cancers: 229 independent lead SNPs (*P*<5E-8) spanning 128 regions >1 Mb apart were associated with at least two cancers. Five lead SNPs were >1 Mb from any previously reported lead SNP for any of the four cancers and marked completely new risk loci for these cancers, mapping to the proto-oncogene *MYCN* (2p24.3), the tumor suppressor *CCNC* (6q16.2), *ANTXR1* (2p13.3), *EPHB3* (3q27.1) and *MAFB* (20q12). Several other lead SNPs were >1 Mb from known lead SNPs for at least one of the individual cancers contributing to the combination and marked potential new risk loci for these cancers: breast (24 loci), prostate (33 loci), ovarian (34 loci) and endometrial (44 loci). Nearly a third of all lead SNPs were associated with >=3 cancer types, with SNPs mapped to *INCENP*, *RNLS*, *ATM*, *CHEK2*, *CDKN2A* and *TERT* associating with all four cancers. For ~14% of all lead SNPs, particularly SNPs mapped to p53 pathway genes *TP53*, *MDM4*, *ATM*, *CHEK2* and *CASP8*, the allele that conferred

risk for one cancer was protective for another cancer. We followed our locus discovery effort with functional annotation, expression- and network-enrichment analyses, providing a global view of the landscape of susceptibility regions shared across the four cancers.

A- A+

PgmNr 133: Transcriptome-wide association study in African Americans identifies associations with prostate cancer.

Authors:
P.N. Fiorica [1,2]; M. Abdul Sami [2]; J.D. Morris [3]; H.E. Wheeler [2,3,4,5]

View Session Add to Schedule

Affiliations:
1) Department of Chemistry & Biochemistry, Loyola University Chicago, Chicago, IL; 2) Department of Biology, Loyola University Chicago, Chicago, IL; 3) Program in Bioinformatics, Loyola University Chicago, Chicago, IL; 4) Department of Computer Science, Loyola University Chicago, Chicago, IL; 5) Department of Public Health Sciences, Loyola University Chicago, Maywood, IL

Prostate cancer is the most commonly occurring cancer in African American men. The genetic risk for the cancer has been governed by a few rare variants with high penetrance and many commonly occurring variants with lower impact on risk. The incidence and mortality rate of prostate cancer in African Americans is nearly twice that of their counterparts of European descent. Inversely, African Americans are one of the least commonly studied populations in genetics, making up nearly two percent of all genome-wide association study (GWAS) participants. This discrepancy between disease incidence and representation in genetics highlights the need for more studies of the genetic risk for prostate cancer in African Americans; however, the genetics of individuals of recent African Ancestry have been historically difficult to study due to admixture, the presence of both European and African linkage blocks on chromosomes. To better understand the genetic risk for prostate cancer in African Americans, we performed PrediXcan, a transcriptome-wide imputation method that uses reference transcriptome data, in a cohort of 4,769 individuals (2,463 cases and 2,306 controls) from the Multiethnic Genome-wide Scan of Prostate Cancer (phs000306.v4.p1). We used prediction models from 44 tissues in the GTEx project and three models from the Multi-Ethnic Study of Atherosclerosis. We also performed a traditional SNP-level GWAS and MultiXcan, the gene-based test that integrates shared eQTL signals across multiple panels. We predicted 10 gene-tissue pairs to be significantly associated with prostate cancer ($p < 1.82 \times 10^{-5}$). Of these, three genes uniquely associated with prostate cancer: *EBPL* ($p = 2.54 \times 10^{-7}$), *ACTR3B* ($p = 2.38 \times 10^{-6}$), and *TTLL9* ($p = 1.63 \times 10^{-5}$). Increased expression of *EBPL* was predicted to be significantly associated with prostate cancer across 8 tissues. None of these three genes have been reported to be associated with prostate cancer in the NHGRI-EBI GWAS Catalog. At the SNP level, 112 SNPs from a previously identified locus on chromosome 8 met genome-wide significance with rs76595456 being the most significant at $p = 2.08 \times 10^{-15}$. These SNPs confirm findings from previous GWAS of prostate cancer in African Americans, while PrediXcan predicted genes suggest potential directions for prostate cancer research in African Americans.

A- A+

PgmNr 134: Genome-wide germline correlates of the epigenetic landscape of prostate cancer.

Authors:
K.E. Houlahan [1,2,3]; Y. Shiah [1]; A. Gusev [4,5]; B. Pasaniuc [6,7,8]; M.L. Freedman [9,10,11]; H.H. He [2,12]; R.G. Bristow [2,12,13,14,15,16]; P.C. Boutros [1,2,3,8,17,18,19]

View Session Add to Schedule

Affiliations:
1) Ontario Institute for Cancer Research, Toronto, Canada; 2) Department of Medical Biophysics, University of Toronto, Toronto, Canada; 3) Vector Institute, Toronto, Canada; 4) Division of Population Sciences, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA; 5) Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 6) Department of Computational Medicine, University of California, Los Angeles; 7) Department of Pathology and Laboratory Medicine, University of California, Los Angeles; 8) Department of Human Genetics, University of California, Los Angeles; 9) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA; 10) The Eli and Edythe L. Broad Institute, Cambridge, MA; 11) Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA; 12) Princess Margaret Cancer Centre, University Health Network, Toronto, Canada; 13) Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada; 14) Division of Cancer Sciences, Faculty of Biology, Health and Medicine, University of Manchester, Manchester, UK.; 15) The Christie NHS Foundation Trust, Manchester, UK.; 16) CRUK Manchester Institute and Manchester Cancer Research Centre, Manchester, UK.; 17) Department of Urology, David Geffen School of Medicine, University of California, Los Angeles; 18) Jonsson Comprehensive Cancer Center, David Geffen School of Medicine, University of California, Los Angeles; 19) Institute for Precision Health, University of California, Los Angeles

---

Cancer initiation and progression are driven by germline, environmental and stochastic factors. How these interact to produce the molecular phenotypes of primary human tumours remains unknown. To better understand the role germline variation plays, we quantified the influence of germline single nucleotide polymorphisms (SNPs) on the somatic methylome of 589 primary localized prostate tumours with genome-wide DNA and methylation sequencing. We show that known risk loci influence a tumour’s epigenetic landscape, uncovering a mechanism for cancer susceptibility. We then identify and validate 1,178 loci associated with altered methylation levels in tumour tissue but not in non-malignant tissue. These tumour methylation quantitative trait loci (tumour meQTLs) influence chromatin structure, RNA abundance and protein abundance and recapitulate previously reported risk loci. One prominent tumour meQTL is associated with tumour-specific methylation and expression of *TCERG1L,* a transcription elongation factor predictive of rapid biochemical relapse following definitive local management. Another tumour meQTL is associated with expression of the oncogene *AKT1* and is predictive of relapse in both our discovery cohort and an independent 101-patient validation cohort. Taken together, these data reveal a strong interplay between the germline mutational profile and the epigenomic features of primary tumours which can be exploited to understand the role of germline genetics in the heritability of aggressive prostate cancer.

A- A+

PgmNr 135: Integrating polygenic risk scores information with somatic and transcriptome data to unravel the polygenic architecture of prostate cancer.

Authors:
C. Hicks; T.K.K. Mamidi; J. Wu; E.J. Nicklow

View Session Add to Schedule

Affiliation: Department of Genetics, Louisiana State University Health Sciences Center, New Orleans, Louisiana.

Background: Prostate cancer (PCa) development is a complex process involving both germline and somatic variation. Cancer prevention is the holy grail of cancer elimination, but realizing that vision will require a deeper understanding of the link between genetic susceptibility and tumorigenesis. Advances in high-throughput genotyping and the recent surge of next generation sequencing have enabled development of comprehensive catalogues of germline genetic variants and somatic mutations in cancer. The discovery of acquired somatic mutations has been critical to the realization of precision oncology. Likewise, the discovery of germline genetic variants has enabled development of polygenic risk scores (PGRS) critical to the realization of precision prevention. Here, we propose a novel and innovative approach that integrates information of PGRS derived from germline variation with somatic variation using transcriptome data to bridge precision oncology with precision prevention and to establish putative functional bridges between PGRS and oncogenic signaling pathways in PCa.

Methods: We have recently developed and published methods for mapping oncogenic interactions between germline and somatic mutations in PCa. We are now systematically developing novel algorithms to integrate information on PGRS with somatic and gene expression variation to determine how germline, somatic and gene expression interact and converge in pathways to drive PCa, and to develop robust dual-purpose algorithm for risk and outcome prediction. We are addressing this critical unmet need using GWAS, TCGA, GEO and other data resources for both knowledge discovery and algorithm development.

Results: Preliminary results show that germline mutations used in development of PGRS interact with somatic driver mutations involved in PCa. Expression levels of many oncogenes involved in PCa were correlated with PGRS. Crucially, we have discovered multiple molecular networks and signaling pathways enriched for germline mutations used in developing polygenic risk scores and somatic mutations driving PCa.

Conclusions: Preliminary results from this investigation demonstrate that integrating PGRS information with somatic and transcriptome data has the promise to unravel the polygenic architecture of PCa, and provides foundational knowledge for the development of robust risk and outcome prediction models to bridge precision medicine with precision prevention.

A- A+

PgmNr 136: CRISPRi screen of risk-associated cis-regulatory elements reveals 3D genome dependent causal mechanisms in prostate cancer.

Authors:
M. Ahmed [1]; F. Soares [1]; J. Xia [2]; P. Su [1,3]; H. Guo [1]; J.T. Hua [1,3]; M. Wang [1]; S. Chen [1,3]; S. Zhou [1,3]; J. Petricca [1,3]; Y. Zeng [1]; Y. Zhu [4]; T. Severson [4]; Y. Tian [5]; A. Bosch [6,7]; K.E. Houlahan [8]; M. Lupien [1,3,8]; W. Zwart [4]; M.L. Freedman [9,10]; T. Wang [11]; P.C. Boutros [3,8,12,13,14]; M.J. Walsh [6,7]; L. Wang [5]; G.H. Wei [2]; H.H. He [1,3]

View Session Add to Schedule

Affiliations:
1) Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada; 2) Faculty of Biochemistry and Molecular Medicine, Biocenter Oulu, University of Oulu, Oulu, Finland; 3) Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada; 4) The Netherlands Cancer Institute, Oncode Institute, Amsterdam, the Netherlands; 5) Department of Pathology, Medical College of Wisconsin, Milwaukee, WI; 6) Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 7) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 8) Ontario Institute for Cancer Research, Toronto, Ontario, Canada; 9) Department of Medical Oncology, Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA; 10) Eli and Edythe L. Broad Institute, Cambridge, MA; 11) Department of Genetics, Washington University in St. Louis, St. Louis, MO; 12) Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada; 13) Department of Urology, David Geffen School of Medicine, University of California, Los Angeles, CA; Department of Human Genetics, University of California, Los Angeles, CA; 14) Jonsson Comprehensive Cancer Center, David Geffen School of Medicine, University of California, Los Angeles, CA; Institute for Precision Health, University of California, Los Angeles, CA

INTRODUCTION: Prostate cancer is one of the most heritable diseases to date. Hundreds of single nucleotide polymorphisms (SNPs) have been identified by genome-wide association studies (GWAS) to confer risk of prostate cancer in men. Most prostate cancer associated risk SNPs do not directly alter gene codons, rather modulate cis-regulatory elements (CREs) such as enhancers.

OBJECTIVE: The primary objective of this study was to perform a systematic essentiality screening of prostate cancer risk associated CREs.

METHODS: We previously pinpointed 270 CREs that harbour at least one risk SNP in prostate cancer. In this study, we targeted these CREs using dCas9-KRAB complex (CRISPRi) guided by 5,571 sgRNAs in three prostate cancer cell lines - LNCaP, V16A and 22Rv1.

RESULTS: The screen identified 98 CREs essential for growth of at least one cell line. Interestingly, essential CREs are significantly enriched in the gene desert region of 8q24.21. The most essential CRE is an enhancer harbouring the SNP rs11986220, which increases the risk for prostate cancer by up to 1.8 fold. Suppression of this enhancer significantly reduces cell proliferation and tumor growth in LNCaP and V16A models. RNA-seq analysis identifies MYC, an important oncogene, to be it's primary target gene. However, this enhancer neither confers essentiality nor regulates MYC in 22Rv1 cells, despite having almost identical epigenetic profiles as in LNCaP cells. Further investigation reveals that a CTCF binding site unique to 22Rv1 intervenes the MYC promoter-enhancer interaction in this cell line. We performed 3C, HiC and H3K27ac HiChIP assays to establish that the enhancer interacts with MYC promoter only when this CTCF site is deleted in 22Rv1 cells. Intriguingly, this CTCF site is also found variable among primary prostate cancer patients, and especially, the SNP rs11986220 is an eQTL for MYC only in patients with low deposition of CTCF at this locus.

CONCLUSION: Our study reveals that CRISPRi is an efficient technique to perform systematic functional analysis of CREs. We thus discover that the interaction between MYC promoter and rs11986220-containing enhancer is governed by CTCF-mediated 3D genomic structure, and the causal effect of rs11986220 is variable among patients depending on CTCF binding in this locus. This unveils a novel regulatory mechanism in human genome and warrants improvement of current target-gene analysis of GWAS loci by incorporating 3D genome variability.

A- A+

PgmNr 137: Interim results from the IMPACT study: Evidence for PSA screening in *BRCA2* mutation carriers.

Authors:
E.K. Bancroft [1,2]; E.C. Page [2,1]; M.N. Brook [2,1]; M. Assel [3]; J. Offman [4]; Z. Kote-Jarai [2]; A. Vickers [3]; H. Lilja [5, 6, 7]; R.A. Eeles [1,2]; The IMPACT Study Steering Committee and Collaborators

View Session Add to Schedule

Affiliations:
1) Cancer Genetics Unit, Royal Marsden Hosp, Sutton, United Kingdom; 2) Oncogenetics Team, Institute of Cancer Research, London, UK; 3) Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, NY, USA.; 4) School of Cancer and Pharmaceutical Sciences, Faculty of Life Sciences & Medicine, King's College London, Guy's Cancer Centre, Guy's Hospital, London SE1 9RT.; 5) Departments of Laboratory Medicine, Surgery, and Medicine, Memorial Sloan-Kettering Cancer Center, New York, USA.; 6) Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK.; 7) Department of Translational Medicine, Lund University, Malmö, Sweden

Background: Mutations in *BRCA2* cause a higher risk of early-onset aggressive prostate cancer (PrCa). The IMPACT study is evaluating targeted PrCa screening using PSA in men with germline *BRCA1/2* mutations.

Objective: To report the utility of PSA screening, PrCa incidence, positive predictive value of PSA, biopsy and tumour characteristics after three years' screening, by BRCA status.

Design, Setting & Participants: Men aged 40–69 years with germline pathogenic *BRCA1/2* mutation and male controls testing negative for a familial *BRCA1/2* mutation, were recruited. Participants underwent PSA screening for three years, and if PSA >3.0ng/ml, men were offered prostate biopsy.

Outcome Measurements and statistical analysis: PSA levels, PrCa incidence and tumour characteristics, were evaluated. Statistical analyses included Poisson regression offset by person-years follow-up, chi-squared tests for proportions, t-tests for means, and univariate logistic regression was applied to PSA predictors.

Results and limitations: 3,027 subjects (2932 unique individuals) were recruited (919 *BRCA1* carriers, 709 *BRCA1* non-carriers; 902 *BRCA2* carriers; 497 *BRCA2* non-carriers). After 3 years screening 527 men had PSA >3.0ng/ml, 357 biopsies performed, and 112 PrCas diagnosed (31 *BRCA1* carriers, 19 *BRCA1* controls; 47 *BRCA2* carriers, 15 *BRCA2* controls). A higher compliance with biopsy was observed in *BRCA2* carriers compared with controls (82% vs 66%). Cancer incidence rate per 1,000 person years was higher in *BRCA2* carriers than non-carriers (19.4 vs 12.0; p=0.03); *BRCA2* carriers were diagnosed younger (61 vs 64years;p=0.04) and were more likely to have clinically-significant disease than *BRCA2* non-carriers (73% vs 40%;p=0.03). No differences in age or tumour characteristics were detected between *BRCA1* carriers and *BRCA1* non-carriers. The 4 kallikrein marker model discriminated better (AUC=0.73) for clinically-significant cancer at biopsy than PSA alone (AUC=0.65).

Conclusions: After three years' screening, compared with non-carriers, *BRCA2* mutation carriers were associated with higher incidence of PrCa, younger age of diagnosis and clinically-significant tumours. Therefore, systematic PSA screening is indicated for men in this age group. Further follow-up is required to assess the role of screening in *BRCA1* mutation carriers.

A- A+

PgmNr 138: Decreased nuclear PTEN increases microglia activation and synaptic pruning in a murine model with autism-like phenotype.

Authors:
N. Sarn [1, 4]; R. Jaini [1, 3, 5]; S. Thacker [1, 3]; H. Lee [1]; C. Eng [1, 2, 3, 4, 5]

View Session Add to Schedule

Affiliations:
1) Genomic Medicine Institute, Cleveland Clinic, Cleveland, OH.; 2) Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH 44195, USA; 3) Cleveland Clinic Lerner College of Medicine, Cleveland, OH, USA; 4) Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine; Cleveland, OH 44106, USA; 5) Germline High Risk Cancer Focus Group, Comprehensive Cancer Center, Case Western Reserve University School of Medicine; Cleveland, OH 44106, USA

Germline mutations in the gene encoding Phosphatase and Tensin homolog deleted on chromosome TEN (*PTEN*) account for ~10% of all cases of autism spectrum disorder (ASD) with coincident macrocephaly. To explore the importance of nuclear PTEN in the development of ASD and macrocephaly, we generated a mouse model with predominantly cytoplasmic localization of Pten (*Pten*$^{m3m4/m3m4}$model). Cytoplasmic predominant Pten expression leads to a phenotype of extreme macrocephaly and behavior reminiscent of high-functioning ASD. Transcriptomic analysis of the *Pten*$^{m3m4/m3m4}$ cortex revealed upregulated gene pathways related to myeloid cell activation, myeloid cell migration, and phagocytosis. Interestingly, a contrasting downregulation of gene pathways related to synaptic transmission was observed. In vitro follow up of the transcriptomic data revealed a 25-fold increase in C1q expression ($p = 0.0002$) in mutant microglia compared to wild-type. In addition, we assessed phagocytic ability and efficiency of *Pten*$^{m3m4/m3m4}$ microglia. We found 20% increase in the number of phagocytic *Pten*$^{m3m4/m3m4}$ microglia compared to *Pten*$^{WT/WT}$ ($p = .001$). We also found *Pten*$^{m3m4/m3m4}$ phagocytic microglia were ~2 times more efficient in their phagocytic ability compared to *Pten*$^{WT/WT}$ microglia ($p = 0.004$). To assess impact on synaptic pruning, we co-cultured combinations of wildtype and mutant microglia with neurons and found that *Pten*$^{m3m4/m3m4}$ microglia co-cultured with *Pten*$^{m3m4/m3m4}$ neurons resulted in a 2-fold increase in pruning compared to when no microglia were present ($p= 0.0001$). These in-vitro findings on over pruning of neuronal synapses via a potential neuron-microglia cross talk in a *Pten*$^{m3m4/m3m4}$system were consistent with in vivo observations in our murine model. At 3 weeks of age significant increases in microglial cell area were observed in the cortex of mice with *Pten*$^{m3m4/m3m4}$ mutations compared to that of *Pten*$^{WT/WT}$ cortex ($p = <0.0001$). This microgliosis was concurrent with a decrease in *Pten* expression levels to below 50% of wildtype cortex ($p = <0.001$). The decline in Pten expression and concurrent increase in microgliosis was strongly associated with decreased expression of synaptic markers in the cortices of *Pten*$^{m3m4/m3m4}$ mice. Collectively, our data suggest a significant role for nuclear Pten in microglial pathology: decreased nuclear-Pten leads to disruption of normal synaptic architecture, likely contributing to an ASD phenotype.

A- A+

PgmNr 139: Identification of human-specific mRNA targets of fragile X mental retardation protein.

Authors:
Y. Li [1]; Z. Li [2]; Y. Kang [1]; E. Allen [1]; H. Wu [2]; Z. Wen [3]; P. Jin [1]

View Session Add to Schedule

Affiliations:
1) Deptment of Human Genetics, Emory University, Atlanta, GA; 2) Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA Michael St., Atlanta, GA30322; 3) Department of Psychiatrics, Emory University School of Medicine, Atlanta, GA

Fragile X syndrome (FXS) is the most common inherited form of intellectual disability and a leading genetic cause of autism. FXS is caused by the loss of functional fragile X mental retardation protein (FMRP). FMRP is an RNA-binding protein that forms a messenger ribonucleoprotein complex with polyribosomes for the regulation of protein synthesis at synapses. Three-dimensional (3D) aggregate culture of human-induced pluripotent stem cells (iPSCs) has evolved from embryoid body cultures, quite faithfully following human organogenesis, and provides a new platform to investigate human brain development in a dish, otherwise inaccessible to experimentation. We have developed FXS forebrain organoids and observed reduced proliferation of neural progenitor cells, premature neural differentiation and a deficit in the production of GABAergic neurons, findings which were not observed in the FXS mouse model. To identify the mRNAs bound by FMRP, we performed enhanced crosslinking and FMRP immunoprecipitation followed by high-throughput sequencing using human forebrain organoids and mouse embryonic forebrain from similar developmental stages. Our comparative analyses revealed the mRNAs bound by FMRP in both human forebrain organoids and mouse embryonic brains, and they were enriched in mRNAs that are critical for neurodevelopment and axonogenesis. Interestingly, we also identified a large number of mRNAs that were bound by FMRP only in human, and these are enriched in mRNAs that are involved in RNA metabolism and astrocyte differentiation. Furthermore, by overlapping the differentially expressed genes found using RNA-seq in the FXS organoids and human FXS fetal brain tissues, we were able to identify a subset of mRNAs that were bound by FMRP and displayed differential expression in the absence of FMRP specifically in human. Our study has identified human-specific mRNA targets of FMRP, which have the potential to serve as human-specific druggable targets for FXS and autism in general.

A- A+

PgmNr 140: Sex-specific brain transcriptome dysfunction burden in schizophrenia, autism spectrum disorder, and bipolar disorder.

Authors:
Y. Xia [1,2]; C. Chen [1,2]; Y. Chen [1]; Y. Jiang [1]; C. Liu [1,2]; psychENOCODE

View Session Add to Schedule

Affiliations:
1) Psychiatry, SUNY Upstate Medical University, Syracuse, New York.; 2) School of Life Science, Central South University, Changsha, Hunan, China

Sex differences in psychiatric disorders are well-recognized but poorly understood. While males are more prone to neurodevelopmental disorders such as intellectual disability and autism spectrum disorder (ASD), females are more prone to major depressive disorder and anxiety disorders. Understanding the genomic basis for sex differences in these disorders could guide more precise diagnosis and treatment for the individual.

To study sex differences in psychiatric disorders, we used transcriptome data from 2160 postmortem brain samples within the PsychENCODE project. We combined differential expression and gene co-expression network analyses to provide a comprehensive characterization of male and female transcriptional profiles associated with schizophrenia (SCZ), bipolar disorder (BD), and ASD. Since the transcriptome is the product of both genetic and environmental effects, we hypothesized that the burden of transcriptome dysfunction differs between males and females in these disorders, an extension of the well-known genetic liability model in sex bias studies.

When compared with female controls, we identified 2337 differentially expressed genes (DEGs) in females with SCZ and 60 DEGs in females with ASD. When compared with male controls, we found 683 DEGs in males with SCZ and 511 DEGs in males with ASD. However, no DEGs were identified in either females or males with BD. Overlap analysis between males and females within each disorder showed the greatest rearrangement of transcriptional patterns in SCZ and ASD, with only 10% overlap in SCZ and 2% in ASD. A functional enrichment test also showed different pathways predominating in males and females for both SCZ and ASD. Meanwhile, co-expression analysis unveiled sex-specific regulatory networks in SCZ, BD, and ASD. We could then identify key regulators of the sex-specific networks in each disorder. Combining the differential expression and co-expression results, we generated the burden of transcriptome dysfunction for males and females with each disorder.

Our study revealed marked differences per sex in the burden of transcriptome dysfunction in SCZ and ASD. We found female had a higher transcriptome dysfunction burden in SCZ, whereas males had a higher transcriptome dysfunction burden in ASD. These results could lead to more precise sex-specific diagnosis and treatment of patients with these disorders.

A- A+

PgmNr 141: Spatiotemporal gene expression pattern predicts of autism risk genes.

Authors:
S. Chen [1,2]; Y. Shen [1,2,3]

View Session Add to Schedule

Affiliations:
1) Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University Irving Medical Center, New York, NY.; 2) Department of System Biology, Columbia University Irving Medical Center, New York, NY; 3) Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY

Large scale exome sequencing studies have established that autism spectrum disorder (ASD) is a condition with strong but heterogenous genetic causes. Our knowledge of ASD risk genes is far from complete. We hypothesize that ASD risk genes have distinct spatiotemporal expression signatures in developing human brain under normal conditions.In this study, we obtained single-cell RNA-seq data of human fetal brain samples from a range of developmental stages in recent publications to infer brain cell-type specific gene expression. Using these data, we developed a new method, Frisk, to predict plausibility of ASD risk of all genes by Gradient Boosting. We used known ASD risk genes from SFARI Gene database as positives and the genes with *de novo*likely-gene disrupting (LGD) variants in unaffected siblings from the Simons Simplex Collection study as negatives.We assessed the performance by the ability to prioritization of de novo mutations in unknown risk genes, using data of 5964 cases from published ASD studies. Excluding all positive training genes, we selected gene sets based on ranking of Frisk score or other published methods. In each gene set, we calculated the enrichment rate of mutations in cases by comparing observed number with expectation from background rate, and in turn estimated precision and recall in the gene set. Frisk achieves higher enrichment, precision, and recall in prioritizing LGD and deleterious missense (Dmis) variants than other methods. Most of known ASD risk genes are intolerant of loss of function variants, as quantified by ExAC pLI≥0.9. In the genes with pLI<0.9, there is a trend of enrichment of LGD and Dmis variants (p=0.08). But in genes with pLI < 0.9 and Frisk score >0.4, such variants are significant enriched (p=4e-7), implicating ~60 candidate risk genes. These genes likely contribute to ASD risk through previous under-studied biological mechanisms or genetic models. Finally, we observed that high Frisk score is correlated with high expression in GABAergic and dopaminergic neurons from midbrain in late first trimester, GABAergic and excitatory neurons from prefrontal cortex in second trimester. With the unprecedent resolution of single-cell transcriptomics, our method will facilitate systematic discovery of novel risk genes and understanding of pathogenesis of ASDs.

A- A+

PgmNr 142: Insights into the genetic architecture of autism from exome and genome sequencing of over 60,000 individuals.

Authors:
F.K. Satterstrom [1,2,3,4,22]; J. Fu [3,4,5,22]; H. Brand [3,4,5,22]; J.A. Kosmicki [1,2,3,4,6,22]; H. Wang [3,4]; X. Zhao [3,4,5]; R.L. Collins [3,4,6]; M.S. Breen [7,8,9]; S. De Rubeis [7,8,9]; C.E. Carey [1,2,3,4]; C. Stevens [1,3]; C. Cusick [1,3]; E.B. Robinson [1,2,3,4,10]; A.D. Børglum [11,12,13]; D.J. Cutler [14]; J.D. Buxbaum [7,8,9,15,16,17]; K. Roeder [18]; B. Devlin [19,23]; S.J. Sanders [20,23]; M.J. Daly [1,2,3,4,6,21,23]; M.E. Talkowski [1,3,4,5,6,23]; Autism Sequencing Consortium

View Session Add to Schedule

Affiliations:
1) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 2) Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Center for Genomic Medicine, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA; 5) Department of Neurology, Massachusetts General Hospital, Boston, MA, USA; 6) Division of Medical Sciences, Harvard Medical School, Boston, MA, USA; 7) Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 8) Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 9) Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 10) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 11) The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark; 12) iSEQ, Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark; 13) Department of Biomedicine-Human Genetics, Aarhus University, Aarhus, Denmark; 14) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA; 15) Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 16) Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 17) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 18) Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA; 19) Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; 20) Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA; 21) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; 22) These authors contributed equally; 23) These authors also contributed equally

---

The genetic architecture of autism spectrum disorder (ASD) includes a well-established etiological role for rare and *de novo* protein-truncating variants (PTVs) in genes highly intolerant of such variation, as well as an increased burden of large copy number variants (CNVs). Here, we present the largest-ever analysis of rare and *de novo* variation in ASD by combining whole-exome sequencing (WES) from over 60,000 individuals (19,028 ASD cases, 14,031 unaffected siblings and controls, and parents) and whole-genome sequencing (WGS) of 10,049 genomes from 2,669 ASD families, including samples from the Autism Sequencing Consortium, the Simons Simplex Collection, SPARK, and the Danish iPSYCH study. Analyses restricted to *de novo* coding single nucleotide variants (SNVs) and indels conservatively identified 31 genes associated with ASD at a Bonferroni-corrected threshold, while a Bayesian framework that combines *de novo* and rare case-control SNVs and indels discovered 125

genes associated with ASD at a false discovery rate less than 0.1.

We further processed these WES data for rare and *de novo* CNVs using our recently developed GATK-gCNV algorithm, which is well-calibrated to detect CNVs of ≥5 exons when compared against microarray and WGS (sensitivity >99.7%; positive predictive value >90%). As expected, these analyses identified a higher proportion of large *de novo* coding CNVs in probands (5.6%) than in unaffected siblings (2.3%; p<2.2e-16), and more exons were affected by *de novo* CNVs in probands than in siblings (p=1.1e-3). When we considered the 125 genes identified in the Bayesian framework above, 13 were localized to recurrent genomic disorder (GD) segments (e.g. 16p11.2). *De novo* deletions within established GD regions were enriched for paternal origin across all loci but one: 16p11.2, in which *de novo* CNVs displayed a 95% bias for maternal origin (p=4.1e-10). Within the 112 genes not localized to GD regions, we observed strong enrichment of *de novo* CNVs in probands (86 in 11,598 individuals) compared to siblings (1 in 4,547 individuals), supporting the predicted role of these loci in ASD. Finally, WGS analyses of the initial 7,608 genomes revealed an association of ASD with *de novo* variants in conserved promoter regions, with analyses of the full cohort ongoing. These studies suggest that integrated analyses of all classes of genomic variation can provide novel insights into ASD.

A- A+

PgmNr 143: No consistent evidence for effect of rare pathogenic and likely pathogenic exonic variation in bipolar disorder.

Authors:
A.E. Locke [1,2]; X. Jia [3]; F. Goes [4]; W. Wang [5,6]; D.S. Palmer [7,8]; B.M. Neale [7,8,9]; S.M. Purcell [6,10]; N. Risch [11]; C. Schaefer [12]; E.A. Stahl [5,6,9]; L.J. Scott [2]; P. Zandi [13]; Bipolar Sequencing Consortium Case/Control Working Group

View Session Add to Schedule

Affiliations:
1) Division of Genomics & Bioinformatics, Department of Medicine, Washington University School of Medicine, St. Louis, MO; 2) Center for Statistical Genetics and Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI; 3) Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA; 4) Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD; 5) Department of Genetics and Genome Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 6) Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY; 7) Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 8) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA; 9) Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; 10) Department of Psychiatry, Brigham & Women's Hospital, Boston, MA; 11) Institute for Human Genetics, Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA; 12) Division of Research, Kaiser Permanente Northern California, Oakland, CA; 13) Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD

Bipolar disorder (BD) is a serious mental illness that shares clinical features and genetic susceptibility through common variants with schizophrenia (SCZ). Prior studies have shown enrichment of rare, pathogenic protein-coding variants to be enriched in individuals with SCZ, however, the role of rare, pathogenic protein-coding variation in BD has not been examined. We examined the protein-coding (exonic) sequences of 3,987 unrelated individuals with BD and 5,322 controls of predominantly European ancestry across four cohorts from the Bipolar Sequencing Consortium (BSC). We assessed the burden of rare, protein-altering, single nucleotide variants classified as pathogenic or likely pathogenic (P-LP) in several gene sets: (1) in the exome overall; (2) in genes implicated in BD by common variant GWAS; and (3) in genes implicated in SCZ by common and rare variant studies, including neuronal synaptic and loss-of-function intolerant genes. In contrast to SCZ, we observed no overall increased burden of rare coding P-LP variants exome-wide in BD cases (OR=1.00, 95% CI=0.98-1.03, p=0.39). We did observe an increased burden of rare coding P-LP variants within 165 genes implicated by BD GWAS in 3,987 BD cases (meta-analysis OR=1.9, 95% CI=1.3–2.8, p=6.0x10^{-4}). However, this enrichment did not replicate in an additional 9,929 BD cases and 14,018 controls (OR=0.9, p=0.4) from the BipEx collection. Despite shared common genetic effects between BD and SCZ, we observed no significant enrichment of P-LP variants within genes implicated by SCZ GWAS. In further contrast with SCZ, we also did not see enrichment in two broad classes of neuronal synaptic genes (*RBFOX2* and *FMRP*), nor in genes classified as loss-of-function intolerant. In this study, the largest analysis of exonic variation in BD, individuals with BD do not carry a replicable enrichment

of rare P-LP variants across the exome or in any of several groups of genes with biologic plausibility for BD. Moreover, despite a strong shared genetic susceptibility between BD and SCZ estimated by common genetic variation, we do not observe association between BD risk and rare P-LP coding variants in genes known to modulate risk for SCZ.

A- A+

PgmNr 144: Longitudinal genome-wide association study identifies novel loci and functional follow-up implicates putative effector genes for pediatric bone accrual.

Authors:
D.L. Cousminer [1, 2]; Y. Wagley [8]; J.A. Pippin [1]; G.P. Way [1,2]; S.E. McCormack [1,2]; J.A. Mitchell [1,2]; J.M. Kindler [1,2]; H.J. Kalkwarf [3]; J.M. Lappe [4]; M.E. Johnson [1]; H. Hakonarson [1,2]; V. Gilsanz [5]; J.A. Shepherd [6]; S.E. Oberfield [7]; C.S. Greene [2]; B.F. Voight [2]; A.D. Wells [1,2]; B.S. Zemel [1,2]; K.D. Hankenson [8]; S.F.A. Grant [1,2]

View Session Add to Schedule

Affiliations:
1) Children's Hospital of Philadelphia, Philadelphia, PA; 2) University of Pennsylvania, Philadelphia, PA; 3) Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 4) Creighton University, Omaha, NB; 5) Children's Hospital Los Angeles, Los Angeles, LA; 6) University of Hawaii Cancer Center, Honolulu, HI; 7) Columbia University Medical Center, New York, NY; 8) University of Michigan, Ann Arbor, MI

While many genetic loci are associated with adult areal bone mineral density (aBMD), less is known about genetic determinants of pediatric bone accrual. Moving beyond the standard GWAS approach using static phenotypes, we longitudinally modeled pediatric aBMD and bone mineral content (BMC) trajectories to identify novel loci. The ‘Bone Mineral Density in Childhood Study’ is a multi-ethnic cohort of healthy children and adolescents from five US clinical sites with up to seven annual bone scans. We modeled longitudinal bone gain across ~10,000 observations per skeletal site using ‘SuperImposition by Translation and Rotation’ (SITAR). 36 parallel GWAS were performed on SITAR parameters *a-size*, *b-timing* and *c-velocity* using linear mixed models for aBMD and BMC at the distal 1/3 radius, lumbar spine, femoral neck, total hip, total body less head and skull. We observed 27 genome-wide significant signals, plus 13 additional suggestive signals supported by more than one phenotype. 35 of the resulting 40 signals were novel, with only one previously reported in children. 15 signals reside near genes involved in Mendelian disorders of bone density and/or had functional annotations for osteoblast or osteoclast regulation. Since causal effector genes are uncertain at most GWAS loci, we aimed to physically implicate such genes in human mesenchymal stem cell (hMSC)-derived osteoblasts. We extracted proxy SNPs in LD with each sentinel that coincide with open chromatin determined by ATAC-seq. Leveraging high-resolution genome-wide promoter-focused Capture C data, we detected consistent contacts between open-proxy SNPs and candidate effector genes. At three loci, we performed siRNA knockdown for 12 implicated genes (4 at each locus) in hMSCs in six independent, temporally separated experiments using three hMSC donor lines, differentiated the cells into osteoblasts, and then assessed for metabolic and osteoblastic activity. Knockdown of *PRPF38A, KARS* and *TEAD4* (a single gene at each locus) led to an absence of extracellular calcium deposition, providing new candidate genes for further functional follow-up. Given that five of our loci also yield suggestive association when queried against GWAS data for later-life fracture risk, our findings highlight that utilizing a longitudinal approach during the high bone turnover period of bone accrual combined with variant-to-gene mapping can lead to a greater understanding of the pathogenesis of bone loss and osteoporosis.

A- A+

PgmNr 145: Network analysis identifies key genetic drivers of bone mass.

Authors:
B.M. Al-Barghouthi [1,2]; G. Calabrese [1]; L. Mesner [1,3]; C.J. Rosen [4]; M.C. Horowitz [5]; M.L. Bouxsein [6]; D. Brooks [6]; S.M. Tommasini [5]; C.R. Farber [1,2,3]

View Session Add to Schedule

Affiliations:
1) Center for Public Health Genomics, University of Virginia, Charlottesville, VA; 2) Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA; 3) Department of Public Health Sciences, University of Virginia, Charlottesville, VA; 4) Maine Medical Center Research Institute, 81 Research Drive, Scarborough, ME 04074; 5) Department of Orthopaedics and Rehabilitation, Yale School of Medicine, New Haven, CT 06520; 6) Center for Advanced Orthopedic Studies, Beth Israel Deaconess Medical Center, Department of Orthopedic Surgery, Harvard Medical School, Boston, MA 02215

Osteoporosis is a disease characterized by decreased bone mineral density (BMD) and an increased risk of fracture. Genome-wide association studies (GWAS) for BMD have been highly successful, identifying over 1100 associations; however, few of the responsible genes have been identified. Here, we generated directed networks for bone with the goal of identifying causal GWAS genes that are also key drivers in bone networks. In a cohort of Diversity Outbred (DO) mice (N=192; 96/sex), we measured BMD and related traits and generated RNA-seq data on cortical bone from each mouse. A weighted gene co-expression network was constructed consisting of 26 modules. The eigengenes of 15 modules were significantly ($P_{adj}<0.05$) correlated with BMD and related traits. Directed Bayesian networks were generated for each correlated module and key driver analysis identified module genes that were hubs for subnetworks enriched for genes with known roles in the regulation of BMD. We then identified key driver genes that were located in GWAS loci and supported as causal by colocalizing eQTL. As an example, the most significant key driver ($P=3.8 \times 10^{-4}$) was Rhophilin Rho GTPase Binding Protein 2 (*Rhpn2*), a gene not previously associated with BMD. *RHPN2* is located within a BMD GWAS locus and its expression is regulated by a strong eQTL that colocalizes with the BMD association. In summary, we used a directed causal bone network to identify key drivers, including *RHPN2*, that are likely causal regulators of BMD. These data increase our understanding of the genetic basis of BMD and bone strength and highlight genes that impact bone through their role as key network regulators.

A- A+

PgmNr 146: Large-scale global multi-ethnic GWAS doubles the number of osteoarthritis loci and identifies new treatment targets.

Authors:
C. Boer [1]; K. Hatzikotoulas [2]; L. Southam [2]; L. Stefánsdóttir [3]; U. Styrkársdóttir [3]; J.B.J. van Meurs [1]; E. Zeggini [2]; Genetics of Osteoarthritis Consortium

View Session Add to Schedule

Affiliations:
1) Department of Internal Medicine, Erasmus MC, Medical Center, Rotterdam, Netherlands; 2) Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; 3) deCODE genetics, Amgen, Sturlugata 8, IS101 Reykjavik, Iceland

Osteoarthritis is a serious destructive joint disorder. It is one of the most rapidly rising conditions associated with disability and chronic pain. The lifetime risk of developing symptomatic knee and hip osteoarthritis is estimated to be 45% and 25%, respectively. Osteoarthritis risk has a strong genetic component, ranging from 40%-60% depending on the joint affected. Recent genome-wide association studies (GWAS) have been successful, but have mainly focused on individuals of European descent and involve only a few large cohorts. Here we present results from a global multi-ethnic GWAS meta-analysis of >20 cohorts, including >179,000 osteoarthritis cases and >650,000 controls, among the largest for any complex disease to date.

We identify >132 loci genome-wide significantly associated with osteoarthritis (hip, knee, hand, finger, thumb and/or spine), of which 84 are novel, thereby doubling the number of osteoarthritis-associated loci. Eight of these variants are low-frequency or rare (Minor Allele Frequency≤0.001%) and have medium to large effect sizes (OR=1.8-9.0). We identify strong evidence for genetic correlation between osteoarthritis strata (rg 0.5-0.8), indicating shared genetic aetiology across different joints affected. We find significant enrichment (P-value<0.05) of osteoarthritis-associated variants in active enhancer locations, not only in bone and cartilage-derived tissues but also in muscle and tendon, showing osteoarthritis to be a disease of the whole joint.

We find strong signal enrichment (False Discovery Rate, FDR<0.05) in bone, cartilage and skeletal developmental pathways, and pathways associated with nerve growth, development and psychiatric traits. We also find highly significant genetic correlations (FDR<0.05) with pain, which is the main disease symptom. By using Mendelian randomization framework, we disentangle correlation from causation, we disentangle correlation from causation, identify novel disease associated genes using tissue specific eQTL analysis and identify novel targets for the treatment of all forms of osteoarthritis and associated clinical pain.

A- A+

PgmNr 147: From GWAS to causal variants: Separate regulatory base pairs at *GDF5-UQCC1* underlie common knee osteoarthritis risk and developmental dysplasia of the hip.

Authors:
T.D. Capellini [1,2]; P. Muthuirulan [1]; Z. Liu [1]; A.M. Kiapour [3]; J. Cao [1]; J. Sieker [4]; S. Yarlagadda [1]; D.E. Maridas [5]; V. Rosen [5]; M. Young [1]

View Session Add to Schedule

Affiliations:
1) Human Evolutionary Biology, Harvard University, Cambridge, MA.; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Department of Orthopaedic Surgery, Boston Children’s Hospital, Harvard Medical School, Boston, MA; 4) Department of Pathology and Laboratory Medicine, Tufts Medical Center, Boston, MA; 5) Department of Developmental Biology, Harvard School of Dental Medicine, Boston, MA

In Europeans and Asians, a 130kb haplotype spanning the *GDF5-UQCC1* locus has risen to very high frequency, likely due to past selection on a variant influencing human height. Today, hitchhiking variants on this haplotype are associated with a number of prevalent skeletal diseases, including developmental dysplasia of the hip (DDH) and common knee osteoarthritis (OA). However, as for most GWAS associations, the causal variants are still unknown. We used a combination of genomic and functional studies to identify separate single base pair regulatory changes that lead to hip and knee changes when recapitulated in mice. The key steps were: (1) Intersection of GWAS intervals with ATAC-seq studies of human and mouse hip and knee chondrocytes. This approach highlighted a putative DDH SNP (rs4911178 G to A) in a growth/hip enhancer (*GROW1*), and an OA SNP (rs6060369 C to T) in a knee enhancer (*R4*). (2) Deletion of the corresponding enhancers in human chondrocytes and mouse models, showing that the enhancers are required for normal expression of *GDF5* expression (but not nearby genes). *GROW1*$^{-/-}$ mice had hip shape defects in the direction of human DDH, while *R4*$^{-/-}$ mice had knee defects and developed OA, with shape correlating with OA severity. (3) Testing of single base pair GWAS variants in reporter assays with human chondrocytes. Compared to non-risk alleles, each risk allele (DDH “A”, knee OA “T”) reduced *GDF5* expression. (4) Knocking each disease-associated risk allele into the endogenous enhancers of mouse *Gdf5*. *GROW1*$^{A/+}$ or *R4*$^{T/+}$ mice show lower *Gdf5* expression in chondrocytes from hips and knees, respectively. In addition, allelic replacement mice had several hip (*GROW1*$^{A/A}$) or knee (*R4*$^{T/T}$) morphological alterations in the directions of effect of enhancer loss and disease progression. In each humanized enhancer line, other hindlimb joints were largely unaffected, highlighting how enhancer specificity links to disease etiology. (5) ChIP studies showing that PITX1, a transcription factor involved in hip/knee development and knee OA, was bound at each enhancer in humans *in vitro* and mouse *in vivo,* with risk variants reducing PITX1 binding. These studies identify separate regulatory variants on the same haplotype, but in distinct enhancers, that underlie common human joint diseases. Similar methods may help identify causal GWAS variants for other diseases, and produce useful humanized mouse models for mechanistic and treatment studies. (Supported by NIHAR070139).

A- A+

PgmNr 148: Role of MMP2 in early craniofacial development in zebrafish.

Authors:
B. Tandon [1]; Q. Yuan [1]; L. Maili [1,2]; S.H. Blanton [3]; G.T. Eisenhoffer [4]; A. Letra [5]; J.T. Hecht [1,2,3,5]

View Session Add to Schedule

Affiliations:
1) Department of Pediatrics, UT Health Science Center, Houston, TX.; 2) Graduate School of Biomedical Sciences, University of Texas Health Science Center, Houston TX; 3) Center for Craniofacial Research, University of Texas Health School of Dentistry, Houston TX; 4) Department of Pediatrics, University of Texas MD Anderson Cancer Center, Houston TX; 5) University of Texas Health School of Dentistry, Houston TX

Vertebrate craniofacial development is a critical and finely tuned process that starts during early development. Defects in this process can lead to many different pathologies, the most common of which is non-syndromic cleft lip and palate (NSCLP). NSCLP occurs in 1/700 births annually affecting 4,000 newborns in the US and 135,000 worldwide. Candidate gene analysis and whole exome sequencing of pedigrees of affected and unaffected individuals reveal both an environmental and genetic contribution to NSCLP. Previous work done in our lab using zebrafish identified and characterized *crispld2* as a biologically relevant NSCLP gene. Perturbation of *crispld2* resulted in altered migration and proliferation of neural crest cells (NCCs) defining the underlying mechanism by which it affects craniofacial development. To further elucidate the role of *crispld2* in NSCLP, we used RNA-seq and *in silico* network approaches to identify genes that were differentially expressed between wild type and *crispld2* morphant zebrafish, and previously known to play a role in craniofacial development. One of the genes identified in this process is matrix metallopeptidase 2 (*mmp2*). We assessed association of variants in *MMP2* with human NSCLP in well-characterized families and found significant association between NSCLP and *MMP2*/rs243836 ($p=0.002$) in our Hispanic families. Morpholino knock down of *mmp2* in zebrafish embryos results in embryos with smaller head, eyes and body axis and abnormal mandibular arch skeleton at low concentrations and severe cardiac edema, curved body axis and death at higher morpholino concentrations. The viscerocranium is derived from cranial NCCs that populate the seven pharyngeal arches of the zebrafish embryo. With evidence of *crispld2* in affecting NCC migration, these preliminary experiments point to a CRISPLD2-mediated role of MMP2 in the development of craniofacial skeleton. Experiments using mutant and transgenic lines with high-throughput morphometric analysis are further defining the biological role of MMP2 in NSCLP. Our study provides a model to test putative genes associated with NSCLP *in vivo*, explore their molecular mechanism in early orofacial development and identify new targets that can be tested in our human NSCLP families.

A- A+

PgmNr 149: Feline precision medicine implicates *UGDH* as a novel gene for disproportionate dwarfism.

Authors:
L. Lyons [1]; R. Buckley [1]; R. Grahn [2]; 99 Lives Consortium

View Session Add to Schedule

Affiliations:
1) Veterinary Medicine & Surgery/College of Veterinary Medicine, Univ Missouri, Columbia, Columbia, Missouri.; 2) University of California, Davis, Davis, California

Despite the contribution of a few major genes for disproportionate dwarfism in humans, many dwarf patients are yet genetically undiagnosed. In domestic cats, disproportionate dwarfism is autosomal dominant and has led to the development of a defined breed, the Munchkin or Minuet. This study examined the genetic aspects of feline dwarfism to consider cats as a new biomedical model. Dwarf cats were phenotyped by radiography and deviations from normal forelimb development were quantified. Dwarf cats was genetically analyzed using; parentage testing to develop a pedigree, familial linkage analyses using STRs, genome-wide association studies of case-controls on a 63K DNA array and whole genome sequencing of three dwarf cats. Each genetic approach localized the dwarfism phenotype to a region on cat chromosome B1. No coding variants were identified but a critical region of ~5.7 Mb from cat chromosome B1:170,278,183-175,975,857 (human chromosome 4) was defined, which then implicates a novel gene controlling disproportionate dwarfism and excludes *FGFR4*. Sixty-five candidate variants were identified after considering the dwarf cats to be heterozygous and the variant absent in the 192 additional cats in the 99 Lives cat genome consortium dataset. Overall, 48 genes and transcripts are defined within this critical region. However, no variants were coding nor protein altering, including 33 intergenic and 32 intronic variants found within the eight genes: *LIMCH1*, *APBB2*, *RBM47*, *CHRNA9*, *UGDH*, *RFC1*, *WDR19*, and *TMEM156*. High priority candidate structural variants (SV) were characterized manually in affected individuals using the integrated genomics viewer (IGV) STIX (structural variant index) was used to validate candidate SVs by searching BAM files for discordant read-pairs that overlapped candidate SV regions. SV breakpoints were validated with Sanger sequencing. A complex translocation SV shared in the three dwarfism cats but absent in the 192 normal cats was identified in the *UDP-Glucose 6-Dehydrogenase (UGDH)*, which produces a protein involved in the biosynthesis of glycosaminoglycans. Histologic samples of the distal radius epiphyseal cartilage plate from a neonatal dwarf kitten showed the columnar arrangement of chondrocytes was disorganized with proteoglycan depletion as determined by weak metachromasia with toluidine blue stain. This variant is heterozygous in all dwarf cats genotyped to date, supporting *UGDH* as a novel candidate gene for disproportionate dwarfism.

A- A+

PgmNr 150: Comprehensive genetic analysis yields new insights into the etiologies of right-sided and left-sided colorectal cancer.

Authors:
J.R. Huyghe [1]; T.A. Harrison [1]; S.A. Bien [1]; H. Hampel [2]; J.C. Figueiredo [3,4]; S.L. Schmit [5]; C. Qu [1]; Y. Lin [1]; J.A. Baron [6]; A.J. Cross [7]; B. Diergaarde [8]; D. Duggan [9]; S. Harlid [10]; D.M. Levine [11]; L.C. Sakoda [12,1]; M.L. Slattery [13]; F.J.B. van Duijnhoven [14]; B. Van Guelpen [10,15]; A.T. Chan [16,17,18,19,20,21]; M. Hoffmeister [22]; M.A. Jenkins [23]; R.E. Schoen [24]; P. Vodicka [25]; E. White [1,26]; G. Casey [27]; R.B. Hayes [28]; P.A. Newcomb [1,26]; S.B. Gruber [4]; L. Hsu [1,11]; U. Peters [1,26]; on behalf of CCFR, CORECT and GECCO

View Session Add to Schedule

Affiliations:
1) Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.; 2) Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, USA.; 3) Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA.; 4) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA.; 5) Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA.; 6) Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA.; 7) Department of Epidemiology and Biostatistics, Imperial College London, London, UK.; 8) Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, and UPMC Hillman Cancer Center, Pittsburgh, PA.; 9) Translational Genomics Research Institute - An Affiliate of City of Hope, Phoenix, Arizona, USA.; 10) Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå, Sweden.; 11) Department of Biostatistics, University of Washington, Seattle, Washington, USA.; 12) Division of Research, Kaiser Permanente Northern California, Oakland, California, USA.; 13) Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA.; 14) Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands.; 15) Wallenberg Centrum for Molecular Medicine, Umeå University, Umeå, Sweden.; 16) Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.; 17) Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.; 18) Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.; 19) Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.; 20) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA.; 21) Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA.; 22) Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.; 23) Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia.; 24) Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA.; 25) Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic.; 26) Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA.; 27) Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA.; 28) Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, New York, USA.

Sporadic colorectal cancer (CRC) is a heterogeneous disease, consisting of etiologically and clinically distinct case subgroups with different prognosis and drug responses. The anatomic sublocation of the primary CRC tumor has long been recognized as an important clinical factor. Known differences in molecular characteristics among tumors arising in different locations of the colorectum suggest differences in underlying mechanisms of carcinogenesis. Tumors originating in the proximal, right-sided colon more frequently display microsatellite instability, the CpG island methylator phenotype, and $\mathrm{BRAF}^{\mathrm{V600E}}$ mutations, while tumors originating in the distal, left-sided colon or rectum, display higher frequencies of mutations for many driver genes and more chromosomal instability. Epidemiological studies also suggest differences in mechanisms and pathways involved. Established risk factors including physical activity, anthropometric traits, and smoking were shown to be differentially associated by sublocation. In this study, we examined whether genetic risk factors for CRC differ by anatomic sublocation. To identify new risk loci specific to certain sublocations, we performed GWAS meta-analyses leveraging data of 48,214 CRC cases and 64,159 controls of European ancestry. We characterized effect heterogeneity at CRC risk loci using multinomial modeling. We identified 14 loci that reached genome-wide significance ($P<5\times10^{-8}$) and that were not reported by previous GWAS for overall CRC risk. Multiple lines of evidence support candidate genes at many of these loci, including *MUTYH*, *PTGER3*, *LCT*, *MLH1*, *CDX1*, *KLF14*, *BMP7*, *PYGL*, and *BCL11B*. We detected substantial heterogeneity between right- and left-sided CRC. Just over half (61) of 110 known and new risk variants showed no evidence for heterogeneity. In contrast, 22 variants showed association with left-sided CRC (distal colon and rectal cancer), but no evidence for association or an attenuated association with right-sided CRC (proximal colon cancer). These included loci implicating Wnt/β-catenin and Hedgehog signaling genes, suggesting differential involvement of these pathways. There was strong evidence for right-sided colon-specific effects for only two loci. Our results demonstrate that left-sided and right-sided CRC have, to a sizeable extent, different genetic etiologies. Future studies of risk factors and mechanisms of carcinogenesis should take into consideration the anatomic sublocation of the tumor.

A- A+

PgmNr 151: GWAS for non-melanoma skin cancer in the UK Biobank cohort: Novel loci include the checkpoint inhibitor CTLA4, where common variants protect against skin cancer and increase risk for auto-immune traits.

Authors:
V. Agarwala [1,2]; Y. Tanigawa [1]; G. Venkataraman [1]; M. Aguirre [1]; M. Rivas [1]

View Session Add to Schedule

Affiliations:
1) Biomedical Informatics, Stanford University, Palo Alto, CA.; 2) Stanford School of Medicine, Stanford, Palo Alto, CA.

Non-melanoma skin cancers (mainly basal cell carcinoma, squamous cell carcinoma) represent the most common cancer type (nearly 3m cases per year in the US). Most such cancers are curable, but screening and early detection are important, especially among those with increased risk. One of the strongest risk factors is prior skin cancer; a third to a half of individuals who have skin cancer will develop a second skin cancer within five years, likely due to a combination of environmental and genetic risk factors. Identifying genetic factors may elucidate the underlying pathophysiology of skin cancer, and inform targeted population screening programs.

Here we report the largest genome-wide association study of non-melanoma skin cancer conducted to date, in the UK Biobank cohort of 337,119 white British individuals (16,789 cases; 320,330 controls). We identify 53 independent genome-wide significant signals across 44 unique loci; 34 of these confirm previously reported associations, and 19 represent novel signals, 5 of which have MAF <5%. Novel loci potentially implicated include CTLA4 (immune checkpoint inhibitor), JDP2 (cell cycle regulator implicated in cancer types), and PTPN22 (lymphoid immune cell phosphatase).

On further interrogation of variants at the CTLA4 locus in a PheWAS, we find that several common variants (one of which is the missense SNP rs231775 at chr2:204732714; MAF = 0.39) are significantly associated with multiple phenotypes in the UKBB. The minor allele at the CTLA4 locus is protective against non-melanoma skin cancer (OR = 0.93; p=1.1e-09), and increases risk for auto-immune conditions including hypothyroidism (OR = 1.19; p=4.8e-61). Given that CTLA4 inhibitors are widely used clinically to treat melanoma, but cause an auto-immune side effect profile that can include thyroiditis, these results suggest that GWAS/PheWAS may proactively inform target selection for drug development in cases where the biology may be uncharacterized.

Finally, we computed polygenic risk scores (PRS) for non-melanoma skin cancer. A model including 3,842 variants and covariates performed with AUC 0.704. Individuals in the top 1% of the PRS distribution had a relative skin cancer risk of 6.04 relative to those in the 40th-60th percentile; individuals in the bottom 1% had a relative risk of 0.165. These data collectively update our understanding of the genetic basis of non-melanoma skin cancer, and suggest potential utility of this information in risk stratification.

A- A+

PgmNr 152: Discovering susceptibility genes shared by neural crest derived tumors neuroblastoma and melanoma.

Authors:
M. Avitabile [1,2]; M. Succoio [2]; A. Testori [1]; A. Cardinale [1,2]; Z. Vaksman [3,4]; V.A. Lasorsa [1,2]; S. Cantalupo [5]; M. Esposito [1]; F. Cimmino [2]; A. Montella [2]; D. Formicola [5]; J. Koster [6]; V. Andreotti [7]; P. Ghiorzo [7]; M.F. Romano [1]; S. Staibano [8]; M. Scalvenzi [9]; F. Ayala [10]; H. Hakonarson [4,11,12]; V.M. Corrias [13]; M. Devoto [4,11,14]; M.H. Law [15]; M.M. Iles [16]; K. Brown [17]; S. Diskin [3,4]; N. Zambrano [1,2]; A. Iolascon [1,2]; M. Capasso [1,2,5]

View Session Add to Schedule

Affiliations:
1) Department of Molecular Medicine and Medical Biotechnology University of Naples Federico II, Naples, 80136, Italy; 2) CEINGE Biotecnologie Avanzate, Naples, 80145, Italy; 3) Division of Oncology and Center for Childhood Cancer Research, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA.; 4) Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.; 5) IRCCS SDN, Naples, 80133, Italy; 6) Department of Oncogenomics, Academic Medical Center, University of Amsterdam, Meibergdreef, Amsterdam, 1011, The Netherlands.; 7) Università degli Studi di Genova; 8) Department of Advanced Biomedical Sciences, University Federico II of Naples; 9) Dipartimento di Medicina clinica e Chirurgia, Università degli Studi di Napoli Federico II, Naples, 80136, Italy; 10) National Cancer Institute, "Fondazione G. Pascale"-IRCCS, Naples, Italy; 11) Division of Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA.; 12) The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA.; 13) Experimental therapy in Oncology, Istituto Giannina Gaslini, Genoa, Italy.; 14) Department of Translational and Precision Medicine, University of Rome Sapienza, Rome, Italy; 15) Statistical Genetics, QIMR Berghofer Medical Research Institute Brisbane, Queensland, 4006, Australia; 16) University of Leeds, Leeds, UK; 17) Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Cells that arise from the neural crest (NC) cells have stunning features of motility, proliferation, and pluripotency, so they can participate to both normal embryogenesis and development of pathologies such as malignancy and tumor metastasis. Neuroblastoma (NB) and malignant cutaneous melanoma (CMM) are NC cells-derived tumors and recent genome-wide association studies (GWAS) have demonstrated that common DNA variants are risk factors of these two diseases. To identify shared loci between these two tumors, we took a three-staged approach to conduct cross-disease meta-analysis of GWAS for NB and CMM (2101 NB cases and 4202 controls; 12874 CMM cases and 23203 controls). Findings were replicated in 1403 NB cases and 1403 controls of European ancestry and in 636 NB, 508 CMM cases and 2066 controls of Italian origin. We found a cross association at locus 1p13.2 (rs2153977, OR=0.91, $P=5.36x10^{-8}$). We also detected a suggestive ($P<10^{-7}$) NB-CMM cross association at 2q37.1 with opposite effect on cancer risk. Pathway analysis of 110 NB-CMM risk loci with $P<10^{-4}$ demonstrated enrichment of biological processes such as cell migration, cell cycle, metabolism and immune response, that are essential of human NC cells development, underlying both tumors. *In vitro* and *in silico* analyses indicated that the rs2153977-T protective allele, located in a NB and CMM enhancer, decreased expression of *SLC16A1* via long-range loop formation and altered a T-box protein binding site. Upon depletion of *SLC16A1*, we observed a decrease of cellular proliferation and invasion in both NB and CMM cell lines, suggesting its role as oncogene. These findings demonstrate that high expression of *SLC16A1, due to a risk variant,* might play a role in malignant neuroblastic and melanocyte transformation and disease progression, co-opting molecular features used by developing NC cells.

A- A+

PgmNr 153: Discovery of novel rare and common genetic variants for colorectal cancer using TOPMed whole genome sequencing data.

Authors:
Wang [1,2]; T.A. Harrison [1]; J. Huyghe [1]; C. Qu [1]; S. Chen [3]; S. Bien [1]; D.V. Conti [4]; C. Kooperberg [1]; M.J. Gunter [5]; V. Moreno [6,7,8]; P.D.P Pharoah [9]; K.R. Curtis [1]; T.O. Keku [10]; S.L. Schmit [11]; P.C. Scacheri [12]; A. Kundaje [13]; S.J. Gallinger [14]; M.A. Jenkins [15]; G.R. Abecasis [3]; D.A. Nickerson [16]; H.M. Kang [3]; G. Casey [17]; S.B. Gruber [4]; L. Hsu [1]; U. Peters [1,2]

View Session Add to Schedule

Affiliations:
1) Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA; 2) Department of Epidemiology, University of Washington, Seattle, WA; 3) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI; 4) Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA; 5) Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon, France; 6) Cancer Prevention and Control Program, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain; 7) ER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain; 8) Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain; 9) Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 10) Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, NC; 11) Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL; 12) Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH; 13) Department of Genetics, Stanford University, Stanford, CA; 14) Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada; 15) Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia; 16) Department of Genome Sciences, University of Washington, Seattle, WA; 17) Center for Public Health Genomics, University of Virginia, Charlottesville, VA

The NHLBI Trans-Omics for Precision Medicine (TOPMed) Project generated deep-coverage whole genome sequencing on >230 million genetic variants (Freeze 5), which provides better imputation quality and allows analysis of rare and less frequent variants. Using TOPMed imputed genotyping data on 59,062 colorectal cancer (CRC) cases and 71,723 controls of European ancestry, we performed genome-wide association testing on 64.8 million variants, adjusting for age, gender, study, genotyping platform, and principal components. A total of 4,999 variants were statistically significantly associated with CRC risk at a genome-wide significance level ($p<5x10^{-8}$). Using clumping methods, we identified 132 independent region that were not in linkage disequilibrium ($R^2>0.1$) with previously reported GWAS variants. The majority of these 132 are within ±1Mb distance from known loci (range of MAF= 0.5 to 0.005; range of $P=4.9x10^{-8}$ to $1.4x10^{-71}$), indicating many secondary signals. In addition, we identified 15 novel independent regions, among which eight were rare variants (range of MAF=0.01-0.003; $P_{min}=9.3x10^{-14}$). The most significant variant (*rs200451572*; MAF=0.006), located in a novel independent region, is a rare missense variant in the β-glucuronidase (*GUSB*) gene, which coverts conjugated bilirubin for re-absorption in the gut. Other new variants are

related to chronic inflammation, obesity, cell recognition and adhesion, and ubiquitin signaling, suggesting involvement of these pathways in colorectal tumorigenesis. By leveraging deep-coverage TOPMed reference panel on large sample size, we identified novel genetic variants, both common and rare, associated with CRC risk. We are currently conducting functional analysis on novel variants, and conditional analysis on secondary signals.

A- A+

PgmNr 154: Polygenic risk scores improve prediction and risk stratification: Results from a pan-cancer analysis in the UK Biobank.

Authors:
L. Kachuri [1]; R.E. Graff [1]; K. Smith-Byrne [2]; S.R. Rashkin [1]; E. Ziv [3]; J.S. Witte [1]; M. Johansson [2]

View Session Add to Schedule

Affiliations:
1) Department of Epidemiology & Biostatistics, University of California, San Francisco, San Francisco, CA, USA; 2) International Agency for Research on Cancer, World Health Organization, Lyon, France; 3) Division of General Internal Medicine, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA

The contribution of germline variants to cancer risk at the population level, and in relation to modifiable risk factors, has not been well quantified. To evaluate this, we conducted a comprehensive analysis of 16 cancers in European ancestry UK Biobank participants (n=413810, 52 to 4170 incident cases).

Polygenic risk scores (PRS) were derived by abstracting variants with minor allele frequency ≥0.01 and $P<5x10^{-8}$ from the literature and selecting independent ($r^2<0.3$) SNPs with the smallest *P*. Cancer-specific PRS, modifiable risk factors, and family history of cancer were modelled using Cox regression accounting for competing risks.
All models were well-calibrated, and each PRS was associated with the relevant cancer ($P<0.01$). Compared to standard weighting of each SNP in the PRS by its log odds ratio (β), we observed that weights incorporating the β standard error, β/SE(β), yielded the highest hazard ratios (HR) and best risk discrimination, as measured by the C index. This result suggests that accounting for uncertainty in SNP effect sizes improves power and PRS predictive performance in independent populations.

The relative influence of genetic and modifiable risk factors varied across cancers. High genetic risk (PRS≥80thpercentile) accounted for 4.0% (lung) to 30.3% (testicular) of new cases and was the dominant risk factor for some cancers, such as prostate (attributable fraction (AF_{PRS})=0.23, $P=2.9x10^{-219}$;$AF_{modifiable}$=0). Heritable and modifiable risk factors had more even contributions for other cancers, such as breast (AF_{PRS}: 0.17, $P=1.5x10^{-112}$; $AF_{modifiable}$=0.10) and colorectal (AF_{PRS}: 0.17, $P=1.0x10^{-65}$; $AF_{modifiable}$=0.14). For all cancers, 5-year absolute risk trajectories significantly diverged ($P<10^{-10}$) between individuals at high, average (20th<PRS<80thpercentiles), and low genetic risk. A healthier lifestyle achieved a 22.9-43.9% 5-year risk reduction for those with high genetic risk. Incorporating the PRS into a prediction model improved risk discrimination compared to conventional risk factors alone (testicular: ΔC=0.14, thyroid: 0.10, leukemia: 0.06, breast: 0.06, prostate: 0.05, melanoma: 0.04), and compared to family history of cancer (prostate: C=0.76 vs. 0.72, breast: C=0.62 vs. 0.56, colorectal: C=0.71 vs. 0.68).

In summary, we developed a novel approach for combining SNPs in the PRS, and produced quantitative data supporting the clinical utility of PRS in complementing conventional cancer risk factors for improving risk stratification.

A- A+

PgmNr 155: Hundreds of GWAS on a deeply-imputed cohort of 117,242 Ashkenazi Jews identify novel associations with uncommon and rare coding variants across a wide range of diseases.

Authors:
A. Kleinman [1]; V. Vacic [1]; S. Pitts [2]; R. Gentleman [2]; 23andMe Research Team

View Session Add to Schedule

Affiliations:
1) Research, 23andMe, Mountain View, CA; 2) Therapeutics, 23andMe, South San Francisco, CA

Population isolates have long proved useful in studying the effects of the genetics of human disease. The Ashkenazi Jews, who underwent a bottleneck event in the 9th-11th century, are one of the largest and best-characterized such populations, and are a good population for mapping diseases to founder variants that might otherwise be difficult to find. We constructed a jointly-called dataset of 890 Ashkenazi Jewish (AJ) whole genomes sequenced to a mean depth of 31x, and verified that it gives superior imputation performance into Ashkenazi genotypes as compared with the Haplotype Reference Consortium. We imputed the dataset into a deeply-phenotyped cohort of 117,242 genotyped individuals of Ashkenazi descent who consented to participate in 23andMe's research program, and ran GWAS on 347 phenotypes, including autoimmune, skin, blood, cancer, metabolic, musculoskeletal, neurological, renal and infection phenotypes. We found an uncommon missense variant in the DNA repair gene ATM (p.Leu2307Phe, AJ MAF=3.1%) that associated with several cancer-related phenotypes, including breast cancer (P = 1.74 × 10-9, OR = 1.7; 95% CI = 1.46, 2.01), thyroid cancer (P = 1.97 × 10-10, OR = 2.53; CI = 1.95, 3.28), non-Hodgkins lymphoma (P = 1.97 × 10-12, OR = 2.20; CI = 1.80, 2.69), leukemia (P = 2.02 × 10-9, OR = 2.03; CI = 1.63, 2.53), myeloproliferative neoplasms (P = 1.25 × 10-21, OR = 3.45; CI = 2.77, 4.30), as well as two potential precancerous phenotypes, moles (P = 2.72 × 10-12, OR = 1.62; CI = 1.40, 1.85) and uterine fibroids (P = 3,14 × 10-16, OR = 1.73; CI = 1.53, 1.95), and we replicated these associations in a separate cohort of 107,333 individuals of whole and partial Ashkenazi ancestry. We also found other biologically plausible novel associations with low-frequency and rare coding variants that are heavily enriched in AJs, including p.Thr273Met ALPL with osteoporosis (MAF=3.4%; P = 4.54 × 10-9, OR = 1.33; CI = 1.21, 1.46), p.Ser578Thr TGFBR2 with height (MAF=4.1%; P = 1.55 × 10-16, β = -0.37 inches; CI = -0.46, -0.28), and p.Cys398Gly TSHR with hypothyroidism (MAF=0.2%; P = 2.00 × 10-16, OR = 4.63 ; CI = 3.37, 6.35). This work is the first to our knowledge to run GWAS at scale on a cohort of over a hundred thousand Ashkenazi Jews, and helps illuminate the contribution of AJ-enriched functional variants to risk across a variety of diseases.

A- A+

PgmNr 156: Characteristics of genes that underlie both recessive and dominant Mendelian conditions.

Authors:
J.X. Chong [1,3]; M.J. Bamshad [1,2,3]; University of Washington Center for Mendelian Genomics

View Session Add to Schedule

Affiliations:
1) Pediatrics, University of Washington, Seattle, Washington.; 2) Genome Sciences, University of Washington, Seattle, Washington.; 3) Brotman Baty Institute for Precision Medicine

---

Matchmaking has become an key part of gene discovery, syndrome delineation, and establishing genotype-phenotype relationships. The vast majority of matches are gene-only, forcing investigators to prioritize cases on other types of data, such as mode of inheritance. Matches with a different mode of inheritance are increasingly common, but are often ignored because it is assumed it is unlikely that a single gene underlies both autosomal dominant and autosomal recessive MCs (D-R genes). However D-R genes are surprisingly common, so we sought to identify characteristics that would enable researchers to correctly prioritize matches.

An analysis of OMIM identifies: 981 autosomal genes exclusively underlying dominant MCs (AD), of which 350 are “de novo” (DN; pathogenic variants are typically de novo, rather than inherited); 1942 genes exclusively underlying recessive MCs (AR); 211 X-linked genes, and 373 D-R genes. D-R genes constitute ~10% of all known genes for MCs, slightly greater than the ~9.5% that are DN. As measured by recent estimates of gene-based selection coefficients (reduction in fitness due to heterozygous loss of function, Weghorn et al. 2019), D-R genes are seemingly subject to somewhat stronger selection (median shet_drift=0.019) than AR genes (median 0.012, $p=3.x10^{-6}$), but far less than AD genes (median 0.069, $p<2x10^{-16}$). D-R and AR genes have median pLI (probability of loss of function intolerance) scores ~0, while the median for AD genes is 0.82. The gnomAD LOEUF score, which detects genes depleted of heterozygous loss of function variants, is able to more finely discriminate between AD, D-R, and AR genes (0.40, 0.68, and 0.88, respectively). D-R genes are expressed in fewer tissues (~24% of tissues in GTEx) than both AD and AR genes (~50%), and are enriched for proteins with structural, transmembrane, and channel activity GO annotations. Finally, D-R genes more often lead to dermatologic, hematologic, audiologic, and ophthalmologic manifestations than AD and AR genes (FDR≤0.1), as well as fewer manifestations than AD, but not AR, genes (AD p=0.002; AR p=0.2). These data indicate that D-R genes are less constrained and under weaker selection than AD genes despite underlying a dominant phenotype. More importantly, low pLI scores for D-R genes may be used incorrectly to rule out matches with AD/DN MCs. D-R genes also underlie MCs that are less severe, more limited in scope. We use these metrics to construct a predictor for D-R genes.

A- A+

PgmNr 157: Cohort analysis/reanalysis identifies *CDH11* a novel gene for Teebi hypertelorism syndrome.

Authors:
D. Li [1]; E.J. Bhoj [1]; C. Seiler [1]; M.E. March [1]; M.R. Battig [1]; P. Sleiman [1]; E. Bedoukian [1]; L.C. Pyle [1]; M. Wilson [2]; C. Patel [3]; J. Christodoulou [4]; E.H. Zackai [1]; H. Hakonarson [1]

View Session Add to Schedule

Affiliations:
1) The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania.; 2) Children's Hospital at Westmead, Australia; 3) Genetic Health Queensland, Australia; 4) The University of Melbourne, Australia

Teebi syndrome (MIM #145420) is a rare craniofacial disorder characterized by striking hypertelorism, prominent forehead, broad nasal tip, broad eyebrows, and thin upper lip with everted lower lip. Psychomotor development is typically normal, although developmental delay has been reported in some patients. In search of genetic causes, we previously described three patients in two families with mutations in *SPECC1L*, a gene encoding a cytoskeletal protein required for cell-cell adhesion and migration (Bhoj et al., 2015).

In a cohort of 17 families with a clinical Teebi diagnosis or clinical features resembling *SPECC1L* hypertelorism syndrome, we identified two pathogenic variants in *SPECC1L* (Bhoj et al., 2018) and pathogenic variants in *MED12/HYAL2* in two other unrelated patients. Interestingly, we also identified six novel heterozygous missense mutations in *CDH11* (cadherin 11) in eleven patients from six unrelated families (3 *de novo* cases, 2 dominant cases, and one adopted child), resulting in an overall likely molecular diagnostic yield of 58.8%.

Cadherins are single chain transmembrane glycoproteins that mediate calcium-dependent cell-cell adhesion, which is essential for tissue development. Biallelic loss-of-function mutations in *CDH11* were recently identified in four families with Elsahy-Waters syndrome, a rare disorder characterized by distinctive facial features, premature loss of teeth, vertebral and genital anomalies, and intellectual disability. The characteristic faces resemble Teebi syndrome, however our eleven patients with clinical diagnosis of Teebi syndrome have none of other dental, vertebral or genital anomalies, and not all have intellectual disability, suggesting *CDH11* mutations can cause two similar but still different phenotypes. All the six variants reside in the extracellular cadherin domain, and they may disrupt the homodimerization and/or interaction with cadherins on the opposing cells to form trans-dimers, which disrupts cell-cell adhesion. We hypothesize that mechanistically these mutations induce a signal transduction defect through a dominant negative effect. Interestingly, two genes identified (*SPECC1L* and *CDH11*) are responsible for cell-cell adhesion and migration. We are currently conducting follow-up functional studies in a cellular model to investigate the b-catenin pathway and in a zebrafish model to examine whether these variants recapitulate striking hypertelorism seen in our patients.

A- A+

PgmNr 158: Known disease variants in a population-wide analysis of 135,638 Finns.

Authors:
H.O.H Heyne [1,2,3,4]; S.M.L Lemmelae [1]; J.K. Karjalainen [1,2,3,4]; A.S.H. Havulinna [1]; A.P. Palotie [1,2,3,4,5]; M.J.D. Daly [1,2,3,4]

View Session Add to Schedule

Affiliations:
1) Finnish Institute for Molecular Medicine, University of Helsinki, Helsinki, Finland; 2) Program for Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 3) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 5) Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA

Rare genetic variants influencing common or rare diseases often have large effect sizes with potential direct treatment implications; studying their effects comprehensively requires however large cohort sizes. The population-wide cohort of FinnGen offers an unprecedented opportunity to systematically investigate phenotype associations of rare variants because the founding bottleneck of Finland concentrates them to higher population frequencies.

Here, we studied genome-wide associations of 1,801 disease endpoints using nation wide electronic health record data of 135,638 Finnish individuals. These individuals carry a total of 1582 coding variants that are annotated as (likely) pathogenic (224 variants) or “conflicting interpretations of pathogenicity” (1358 variants) in ClinVar with phenotype associations in FinnGen (p-value < 0.01). We systematically compare those variants’ direction of effect and associated phenotypes between FinnGen and ClinVar. We validate known/disputed variant effects and find previously unknown phenotype associations. Of particular interest, the Finnish population bottleneck concentrates generally diverse recessive mutational spectra onto one or a small number of higher elevated frequency variants. This ‘Finnish disease heritage’ allows us to use FinnGen to examine the phenotypic consequences and departure from Hardy-Weinberg equilibrium for 777 variants (107 (likely) pathogenic) reported to be causes of severe recessive disease - lending empirical data to longstanding questions regarding selection in these scenarios. In heterozygous form, the 777 variants have more significant associations with phenotypes than random (p-value 10^{-9}). We identify cases where heterozygous variant carriers display different or opposite-effect phenotypic associations from homozygous variant carriers – including a variant in *SCN5A* that has previously been found in recessive/compound heterozygous state in multiple individuals with severe cardiac arrhythmia. By contrast, the variant has a protective effect for cardiac arrhythmia/atrial fibrillation phenotypes (ICD codes I47/I48/I49, $p=6x10^{-6}$, beta -0.63) in heterozygous carriers, which could, despite much lower frequency, be replicated in the UK Biobank (p=0.04, beta -0.39). By systematically investigating known disease variants in population wide electronic health record data we validate and refine disease associations while providing insights into pleiotropic variant effects.

A- A+

PgmNr 159: Analysis of phenotype penetrance in heterozygous carriers of autosomal recessive disorders using whole genome sequencing in a precision medicine clinic.

Authors:
Y.-C.C. Hou [1]; H.-C. Yu [1]; R. Martin [1]; E.T. Cirulli [1]; N.M. Schenker-Ahmed [1,2]; M. Hicks [1]; I.V. Cohen [1,3]; T.J. Jönsson [4]; R. Heister [1]; L. Napier [1]; C.L. Swisher [1]; S. Dominguez [1]; H. Tang [1]; W. Li [5]; B.A. Perkins [1]; J. Barea [1]; C. Rybak [1]; E. Smith [1]; K. Duchicela [1]; M. Doney [1]; P. Brar [5]; N. Hernandez [1]; E.F. Kirkness [5]; A.M. Kahn [6]; J.C. Venter [5]; D.S. Karow [1,2]; C.T. Caskey [1,7]

View Session Add to Schedule

Affiliations:
1) Human Longevity, Inc., San Diego, CA, 92121; 2) Department of Radiology, UC San Diego School of Medicine, San Diego, CA, 92093; 3) Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, CA, 92093; 4) Metabolon Inc., Morrisville, NC, 27713; 5) J. Craig Venter Institute, La Jolla, CA, 92037; 6) Division Cardiovascular Medicine, UC San Diego School of Medicine, San Diego, CA, 92037; 7) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030

Purpose: Increasing number of individuals have undergone elective genome sequencing (EGS) which provides an opportunity to evaluate risk, assess genotype-phenotype correlation, and evaluate population prevalence. A comprehensive study surveying genome-wide disease-associated genes in adults with deep phenotype data has not been reported. Here we report results of a three-year precision medicine study with a focus on the analysis of phenotype penetrance (i.e., detectable phenotype changes) in heterozygous carriers of autosomal recessive (AR) disorders. **Methods:** We enrolled 1190 adult participants (402 women [33.8%]; mean age, 54 years [range: 20-89+]; 70.6% European), for metabolomics, advanced imaging, and clinical laboratory tests. We obtained the medical history and a 3-generation pedigree from each participant. HLI Search, a previously described software program which rapidly identifies disease-associated variants, was employed for the systematic analysis of EGS. **Results:** Using the genome-wide approach, 86.3% (1027/1190) of individuals had at least one pathogenic/likely pathogenic (P/LP) variant in 680 AR genes. When systematically analyzed phenotype and AR genetic findings, 10% (104/1027) of heterozygous carriers of AR disorders had phenotype penetrance. Notably, we identified 8 heterozygous carriers of the *PKDH1* P/LP variants (mean age 53.1 years) with innumerable hepatic and/or renal cysts. The rest 5 heterozygous carriers of *PKDH1* (mean age 35.6 years) had no cysts. For hemochromatosis, 18% (12/68) of the *HFE* p.Cys282Tyr heterozygotes had elevated liver R2*, a marker of liver iron content, compared to 8% (87/1034) in individuals without any *HFE* pathogenic variants (p=0.0156, Fisher Exact), suggesting iron regulation is compromised. Genotype-phenotype correlation of *ALPL* (hypophosphatasia) and *LMBRD1 (*methylmalonic aciduria and homocystinuria) were also identified. Genomic and metabolomic analysis revealed 6% (62/1027) of heterozygous carriers of AR genes enriched in amino acid, lipid, nucleotide, and cofactor, and vitamin pathways with metabolic penetrance, affecting serum metabolites levels. In particular, 30% (10/33) of the *PAH* heterozygous carriers had elevated phenylalanine. **Conclusions:** Overall 10% (104/1027) of participants with a P/LP variant in AR genes had detectable mild phenotype changes, suggesting some of the “outside of

normal range” physiological measurements may be partially explained by the variants detected in AR genes using EGS.

PgmNr 160: Defining the phenotypic signature of *CFTR* mutation carriers in the UK Biobank.

Authors:
Y. Lin [1]; N. Panjwani [2]; F. Mohsin [1]; M. Sutton [1]; J. Dennis [5,6]; J.M. Rommens [2,3]; L. Sun [1,4]; L.J. Strug [1,2,4]

View Session Add to Schedule

Affiliations:
1) Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON; 2) Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON; 3) Department of Molecular Genetics, University of Toronto, ON; 4) Department of Statistical Sciences, University of Toronto, ON; 5) Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; 6) Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN

Cystic Fibrosis (CF) is a common recessive, genetic disorder affecting several organs including those of the digestive and respiratory systems. CF is caused by mutations in the CF Transmembrane Conductance Regulator (*CFTR*). Other genes, referred to as modifier genes, contribute to disease severity and co-morbidities and have been identified in CF genome-wide association studies (GWAS). The frequency of *CFTR* mutation carriers is 1 in 25 among Whites, and carriers are assumed not to exhibit clinical features. With the availability of the UK Biobank resource, we can now determine whether there is a carrier phenotype and if CF modifier genes have an impact on the *CFTR* mutation-carrier background. Given early presentation of gastro-intestinal and pancreatic co-morbidities in CF, we hypothesized that carriers would display greater susceptibility to digestive system complications that are exacerbated by CF modifier genes.

With permission, data were obtained from the UK Biobank, a cohort with genetic and phenotypic data on ~500,000 individuals from the UK. The first phase of our study employed diagnosis codes from the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10), which were mapped to 1,472 phecodes (Wu et al 2018). Those with one copy of the most common CF-causing variant, F508del, were defined as carriers. Over 263,000 unrelated Whites (3.3% carriers) with at least 1 ICD-10 code were included in the initial study. All analyses used logistic regression in phenome-wide association studies (PheWAS) for F508del carriers and SNPs from *SLC26A9, SLC9A3* and *SLC6A14* modifier genes previously identified by GWAS. Results showed association between carrier status and several digestive phenotypes, including obstruction of bile duct, where *CFTR* carriers showed a two-fold elevation in odds (OR=2.09, p-value=7.3E-05). PheWAS of *SLC9A3*, a CF modifier of intestinal obstruction, showed association with esophagitis, GERD and related conditions (p-value=1.8E-06), suggesting CF modifier genes also contribute to digestive tract phenotypes among individuals without CF.

Current work includes evaluating the impact of modifiers on the carrier background and defining carrier signatures using phenotypes beyond ICD-10 codes. Improved understanding of elevated burden of disease could be integrated with genetic testing to achieve better health for the large numbers of CF carriers in populations.

A- A+

PgmNr 161: Identification of asymptomatic individuals with homozygous deletions of *SMN1* via routine carrier screening.

Authors:
J. Chaperon; E. Hendricks; M. Westmeyer; S. Leonard

View Session Add to Schedule

Affiliation: Natera, Inc., San Carlos, California.

Background: Case reports of healthy individuals with a homozygous exon 7 deletion of *SMN1* have been previously reported in the literature. Significant intrafamilial variability has been reported for individuals with the same genotype, including between affected dizygotic twins. *SMN2* copy number is a known phenotypic modifier and other potential protective modifier genes have been reported, such as expression of the *PLS3* gene. The incidence of healthy individuals with a homozygous deletion of *SMN1* is unknown.

Objective: To present data on individuals identified with a homozygous deletion of *SMN1* via routine carrier screening and to determine the proportion who were asymptomatic at the time of testing (>18 years of age).

Results: Between April 2015 and December 2018, 344,407 individuals were screened at a reference laboratory for *SMN1* deletions to determine carrier status for Spinal Muscular Atrophy (SMA). *SMN2* copy number was also determined. Seven individuals were identified with a homozygous deletion of *SMN1* and therefore, suspected to be affected with SMA. Clinical follow-up revealed two individuals were known to be affected at the time of testing. Patient age in the cohort ranged from 19-41 years with clinically affected individuals ages 21 and 32. Three individuals were confirmed as clinically asymptomatic and two individuals were presumed unaffected. The rate of unaffected individuals with a homozygous deletion in *SMN1* in the population screened was ~0.0015% (1/68,881). The affected individuals had 2 and 3 copies of *SMN2*; 2 clinically asymptomatic individuals had 3 and >3 copies of *SMN2*; and 3 clinically asymptomatic individuals had 4 copies of *SMN2*.

Discussion: Our data suggest that the incidence of asymptomatic individuals of reproductive age with homozygous *SMN1* deletions is low and supports that *SMN2* copy number is a predictor of phenotype. While 5 individuals with >3 copies of *SMN2* were not reported to be symptomatic at the time of testing, a clinical diagnosis cannot be ruled out. Incidence of clinically asymptomatic individuals with a homozygous deletion in *SMN1* is important information for risk assessment and genetic counseling. Additionally, *SMN2* copy number should be considered during medical work-up because individuals with SMA types III and IV may be ambulatory with symptoms later in life.

A- A+

PgmNr 162: Cell type specificity of intralocus interactions reveals novel disease mechanisms.

Authors:
O. Corradin [1]; D.C. Factor [2]; M. Madhavan [2]; S. Nisraiyya [1]; A. Barbeau [1]; P. Hall [1]; F.J. Najm [2]; T.E. Miller [2,3,4]; Z.S. Nevin [2]; R.T. Karl [2]; K.C. Allan [2]; M.S. Ellit [2]; B. Lima [1]; K. Hazel [1]; A. Hoang [1]; G.K. Dhillon [2]; C. Volsko [5]; C.F. Bartels [2]; E. Shick [2]; D.J. Adams [2]; R. Dutta [5]; A. Kozlenkov [7,8]; S. Dracheva [7,8]; P.C. Scacheri [2,4]; P.J. Tesar [2,6]

View Session Add to Schedule

Affiliations:
1) Whitehead Inst Biomedical Research, Cambridge, Massachusetts.; 2) Department of Genetics, Case Western Reserve University, Cleveland, OH; 3) Department of Stem Cell Biology and Regenerative Medicine, Lerner Research Institute, Cleveland Clinic, Cleveland, OH; 4) Department of Pathology, Case Western Reserve University School of Medicine, Cleveland, OH; 5) Department of Neurosciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH; 6) Department of Neurosciences, Case Western Reserve University School of Medicine, Cleveland, OH; 7) James J. Peters VA Medical Center, Bronx, NY; 8) Friedman Brain Institute and Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY

For the study of complex traits that involve multiple tissues, delineating the cell type impacted by the risk allele is an essential step for moving from disease association to new disease insights. Several approaches integrate GWAS results with cell type specific chromatin activity in order to identify disease causal cell types. However, the results stem from enrichment analysis and thus cannot identify risk loci that act outside of the disease causal cell type.

Here, we present a new strategy to identify cell types most likely affected by the presence of disease alleles in a locus-by-locus manner. We previously showed that SNPs that physically interact with the same target gene as a given GWAS SNP can profoundly impact the risk for disease. We termed these SNPs ‘outside variants.’ Outside variants provide novel information about the intralocus genomic regions that contribute to disease risk and enable us to distinguish regulatory regions that alter risk from those that do not. We compare this risk information to active chromatin regulatory regions in order to distinguish likely pathogenic cell types.

We applied this approach to multiple sclerosis (MS) and identified pathogenic cell types for 113/163 MS risk loci. We predict many loci to act primarily in T cells, but also identify unexpected myeloid and central nervous system (CNS) specific risk loci. Using CRISPR we targeted mutations to outside variants and demonstrate that these sites are not only functional but also specifically impact the identified cell type. We further validated our approach by demonstrating that loci identified to act in T cells explain the majority of genetic correlation with other autoimmune disorders, while CNS specific predictions explain little heritability of these traits. Likewise, our CNS predictions are enriched for heritability of CNS traits such as depression.

Excitingly, when we applied this approach to oligodendrocytes, we identified two MS risk loci that dysregulate transcriptional pause release. Using chemical genetics, we find that inhibition of transcriptional pausing is a dominant pathway blocking generation of new myelin. Furthermore, we

find dysregulation of these genes in MS patient brain tissue. These data implicate cell intrinsic aberrations in MS risk outside of the immune system and demonstrate the utility of our approach to reveal novel disease insights.

A- A+

PgmNr 163: Mechanistic dissection of chromatin topology disruption as an indirect, strong effect driver of neurodevelopmental disorders.

Authors:
K. Mohajeri [1,2,3]; E. D'haene [4,5]; R. Yadav [1,3]; H. Gu [6,7]; B. Menten [4,5]; A. Presser Aiden [6,7]; C. Lowther [1,3]; S. Erdin [1,3]; M. Moyses Oliveria [1,3]; P. Boone [1,3]; E. Lieberman-Aiden [6,7]; J. Gusella [1,3]; S. Vergult [4,5]; M. Talkowski [1,3]

View Session Add to Schedule

Affiliations:
1) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 2) Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA; 3) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA; 4) Center for Medical Genetics, Ghent University, Ghent, Belgium; 5) Dept. of Biomolecular Medicine, Ghent University, Ghent, Belgium; 6) The Center for Genome Architecture, Baylor College of Medicine, Houston, TX; 7) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

Structural variants have the potential to create long-range positional effects, uncouple genes from regulatory elements, and facilitate aberrant 3D chromatin folding. In an independent study, we analyzed the breakpoints of balanced chromosomal abnormalities (BCAs) from 387 congenital anomaly cases and 247 BCA-harboring controls, revealing genome-wide significant enrichments of intergenic BCA breakpoints predicted to disrupt topologically associating domains (TADs) at multiple loci, with the most significant enrichment at chromosome 5q14.3. Among the 11 5q14.3 BCA carriers, all were cases with neurodevelopmental disorders (NDD) with breakpoints localized to a single TAD housing a known NDD driver, *MEF2C*. Our previous targeted expression studies revealed BCA breakpoints disrupting a distal loop boundary but not *MEF2C* directly, resulted in decreased *MEF2C* expression, while HiC analyses of 6 5q14.3 BCA cases found alterations to canonical regulatory contacts. Given these results, we performed a mechanistic dissection of the regulatory network associated with the 5q14.3 locus and its constituent 3D functional elements using Cas9-based genome editing. We generated >180 cell lines in an isogenic background, representing deletions of *MEF2C* alongside 4 TAD and loop boundary targets within the 5q14.3 region in iPS-derived neural stem cells (NSCs) and cortical induced neurons (iNs). Using Nanostring-based targeted expression profiling, we surveyed expression of all 9 protein-coding genes within a 6Mbp window of 5q14.3 in each line. In NSCs, deletion of the case-disrupted distal loop boundary resulted in a statistically significant 20% increase in *MEF2C* expression, with *MEF2C* as the only differentially expressed gene within the TAD. In contrast, *MEF2C* was not differentially expressed in matched iNs. Probing the underlying contact patterns revealed evidence of loop maintenance via CTCF buffering demonstrated by UMI-4C in both iNs and NSCs when we deleted the distal loop boundary, while deletion of the proximal boundary partner displayed increased contacts with predicted enhancers in the adjacent TAD. Our findings highlight compensatory mechanisms against 3D chromatin disruption while underscoring their associated complexity on a cell type and variant class basis. These results suggest potentially novel regulatory mechanisms driving phenotypic outcomes for this genomic disorder region, with significant implications for interpretation of pathogenic structural variation.

A- A+

PgmNr 164: Targeted delivery of chimeric pioneer factors reveals differential activity and chromatin accessibility dependent upon genetic context.

Authors:
Y. Tan; J.H. Goell; I.B. Hilton

View Session Add to Schedule

Affiliation: Department of Bioengineering, Rice University, Houston, TX

Pioneer factors open heterochromatic regions and displace nucleosomes, establishing regions of high chromatin accessibility in which other transcription factors can bind. Their function sets in motion divergent genetic regulatory programs based on their activity and broad consensus sequences. Thus, their role in decompacting repressive chromatin environments has implications in cell reprogramming, cancer progression, and development. We sought to query their role at individual loci and to develop a tunable system through which we could alter chromatin architecture by displacing nucleosomes and opening heterochromatin. To accomplish this goal, we engineered chimeric proteins utilizing the catalytically dead Cas9 (dCas9) protein fused with pioneer factor domains to direct programmable, site-specific pioneer factor activity. Synthetic pioneer factor activity at targeted loci was measured using both gene expression and chromatin accessibility assays across several well-validated test loci. Our data show that pioneer factors, as well as several other epigenetic effector domains, when fused to dCas9, exhibit locus-position specific activities and that these synthetic biomolecules produce footprints of chromatin accessibility that exponentially decay from the target site. Our data also suggests that cofactor dependent activities lead to activation or repression based upon the presence or absence of cell type-specific transcription factors. Together, our new tools and results establish an assay pipeline to query pioneer factor activity using synthetic dCas9-based transcription factors across a spectrum of different endogenous genomic contexts.

PgmNr 165: Incorporation of spatial mapping and confirmation of gene signatures by a multiplex *in situ* hybridization technology into single cell RNA sequencing workflows.

Authors:
J. Phatak; H. Lu; L. Wang; H. Zong; M. Rouault; C. May; X. Ma; C. Anderson

View Session Add to Schedule

Affiliation: Advanced Cell Diagnostics, 7707 Gateway Blvd, Newark, CA.

Complex and highly heterogenous tissues such as the brain are comprised of multiple cell types and states with exquisite spatial organization. Single-cell RNA sequencing (scRNA-seq) is now being widely used as a universal tool for classifying and characterizing known and novel cell populations within these heterogenous tissues, ushering in a new era of single cell biology. However, the use of scRNA-seq presents some limitations due to the use of dissociated cells which results in the loss of spatial context of the cell populations being analyzed. Incorporating a multiplexed spatial approach that can interrogate gene expression with single cell resolution in the tissue context is a powerful addition to the scRNA-seq workflow. In this study, we used the RNAscope Multiplex Fluorescent and RNAscope HiPlex *in situ* hybridization (ISH) assays to confirm and spatially map the diverse striatal neurons that have been previously identified by scRNA-seq in the mouse brain (Gokce *et al*, *Cell Rep*, 16(4):1126-1137, 2016). We confirmed the gene signatures of two discrete D1 and D2 subtypes of medium spiny neurons (MSN): *Drd1a/Foxp1, Drd1a/Pcdh8, Drd2/Htr7*, and *Drd2/Synpr*. The heterogenous MSN subpopulations were marked by a transcriptional gradient, which we could spatially resolve with RNA ISH. Numerous striatal non-neuronal cell populations identified by scRNA-seq, including vascular cells, immune cells, and oligodendrocytes, were also confirmed with the multiplex ISH assay. Finally, the spatial relationship between the D1 and D2 MSN subtypes identified by Gokce *et al.* was visualized using the RNAscope HiPlex assay, which allows for detection of up to 12 RNA targets simultaneously in intact tissues. In conclusion, we have demonstrated the utility of two multiplexed RNAscope ISH assays for the confirmation and spatial mapping of scRNA-seq transcriptomic results in the highly complex and heterogenous mouse striatum at the single cell level. Incorporating spatial mapping by the RNAscope technology into single cell transcriptomic workflows complements scRNA-seq results and provides additional biological insights into the cellular organization and functional states of diverse cell types in healthy and disease tissues

A- A+

PgmNr 166: Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases.

Authors:
B. Soskic [1,2]; E. Cano-Gamez [1,2]; D.J. Smyth [1,2]; W.C. Rowan [3]; N. Nakic [4]; J. Esparza-Gordillo [3]; L. Bossini-Castillo [1]; D.F. Tough [5]; C. Larminie [3]; P.G. Bronson [6]; D. Wille [7]; G. Trynka [1,2]

View Session Add to Schedule

Affiliations:
1) Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK; 2) Open Targets, Wellcome Genome Campus, Cambridge CB10 1SA, UK; 3) Human Genetics, GlaxoSmithKline R&D, Stevenage SG1 2NY, United Kingdom; 4) Functional Genomics, Molecular Science and Technology R&D, GSK Medicines Research Centre, Stevenage, SG1 2NY, UK; 5) Epigenetics DPU, Immuno-Inflammation and Oncology Therapy Area, GlaxoSmithKline, Medicines Research Centre, Stevenage SG1 2NY, UK; 6) Statistical Genetics & Genetic Epidemiology, Biogen, Cambridge, MA, USA; 7) Biostatistics, GSK Medicines Research Centre, Stevenage, UK, SG1 2NY, UK

Complex immune disease variants are enriched in active chromatin regions of T cells and macrophages. However, whether these variants function in specific cell states or stages of cell activation is unknown. We stimulated T cells and macrophages in the presence of thirteen different cytokine cocktails linked to immune diseases and profiled active enhancers and promoters together with regions of open chromatin. We observed that T cell activation induced major chromatin remodeling, while additional exposure to cytokines fine-tuned the magnitude of these changes. Therefore, we developed a new statistical method (Chromatin Element Enrichment Ranking by Specificity, *CHEERS*) that accounts for subtle changes in chromatin landscape to identify SNP enrichment across cell states. Our results point towards the role of immune disease variants in early rather than late activation of memory CD4+ T cells, and with limited differences across polarizing cytokines. Furthermore, we demonstrate that variants associated with inflammatory bowel disease, Crohn’s disease, ulcerative colitis and celiac disease are enriched in chromatin regions specifically active in Th1 cells. These variants intersected the peaks in proximity of genes known to play a key role in the biology of Th1 cells such as IL-12 signaling, IL-23 signaling and regulation of IFN-γ signaling. We also show that Alzheimer’s disease variants are enriched in different macrophage cell states. Our results represent the first in-depth analysis of immune disease variants across a comprehensive panel of activation states of T cells and macrophages, and demonstrate that the early stages of T cell activation are dysregulated in common immune diseases.

A- A+

PgmNr 167: Single cell chromatin accessibility in human pancreatic islets reveals cell type- and state-specific regulatory programs of diabetes risk.

Authors:
D. Gorkin [1,3]; J. Chiou [2]; C. Zeng [3,4]; Z. Cheng [1,3]; J. Han [1,3]; M. Schlichting [3,4]; S. Huang [3]; J. Wang [3,4]; Y. Sui [3,4]; A. Deogaygay [3]; M. Okino [3]; Y. Qiu [4]; Y. Sun [3]; P. Kudtarkar [3]; R. Fang [4]; S. Preissl [1,3]; M. Sander [3,4,5]; K.J. Gaulton [3,5]

View Session Add to Schedule

Affiliations:
1) Center for Epigenomics, University of California San Diego, La Jolla CA; 2) Biomedical Graduate Studies Program, University of California San Diego, La Jolla CA; 3) Department of Pediatrics, Pediatric Diabetes Research Center, University of California San Diego, La Jolla CA; 4) Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla CA; 5) Institute for Genomic Medicine, University of California San Diego, La Jolla CA

Genetic risk variants for complex, multifactorial diseases are enriched in cis-regulatory elements. Single cell epigenomic technologies create new opportunities to dissect cell type-specific mechanisms of risk variants, but data from disease-relevant human tissues has been lacking. Given the central role of pancreatic islets in type 2 diabetes (T2D) pathophysiology, we generated single-cell ATAC-seq profiles from 14.2k islet cells across 3 donors and identified 13 cell clusters including multiple alpha, beta and delta cell clusters which represented hormone-producing and signal-responsive cell states. We cataloged 244,236 islet cell type accessible chromatin sites and identified transcription factors (TFs) underlying both lineage- and state-specific regulation. To integrate these data with GWAS, we developed a framework to measure the enrichment of genetic association within accessible chromatin profiles of single cells, which revealed heterogeneity in the effects of beta cell states and TFs on fasting glucose levels and T2D risk. We further used machine learning to predict the cell type-specific regulatory function of genetic variants, and single cell co-accessibility to link distal sites to putative cell type-specific target genes. Through integrative analyses we localized 239 fine-mapped T2D risk signals to islet accessible chromatin, and further prioritized variants at these signals with predicted regulatory function and co-accessibility with target genes. At the KCNQ1 locus, the causal T2D variant rs231361 had predicted effects on an enhancer with beta cell-specific, long-range co-accessibility to the insulin promoter, and deletion of this enhancer reduced insulin gene and protein expression in human embryonic stem cell-derived beta cells. Our findings provide a cell type- and state-resolved map of gene regulation in human islets, illuminate likely mechanisms of T2D risk at hundreds of loci, and demonstrate the power of single cell epigenomics for interpreting complex disease genetics.

A- A+

PgmNr 168: Applying confidence intervals to clinical polygenic risk scores in 60,000 exome+ sequenced individuals.

Authors:
K. Dunaway; A. Bolze; J. Rizko; S. Luo; S. White; L. Sharma; N. Washington; W. Lee; J. Lu

View Session Add to Schedule

Affiliation: Helix, San Carlos, California

The Polygenic Risk Score (PRS) is an emerging tool for the clinician to understand an individual's genetic risk of disease by stratifying their score into categories based on many common genetic variants and has been applied to common diseases such as Type 2 Diabetes (T2D), Coronary Artery Disease (CAD), and Prostate Cancer (PCa). The majority of these PRS have been developed thus far using microarray technology but recently this research is shifting towards utilizing sequencing. Applying previously developed PRS to sequenced samples presents a technical challenge of how to leverage low confidence sequencing calls. Conservative PRS methods would no-call these sites, substituting population allele frequency for the low confidence call. However, while these methods maintain high confidence results by restricting the reportable range, ignoring sequencing information potentially leads to erroneous categorization. Utilizing lower confidence calls may improve the accuracy of correctly categorizing an individual’s disease risk, but requires capturing potential error to approximate the likelihood of misclassifying an individual’s result. The confidence in risk estimation should be utilized to determine when a PRS test is inconclusive.

In this study, we developed a method for reporting PRS confidence in clinical tests by utilizing known sequencing and imputation error rates to provide confidence intervals together with disease risk estimates. We applied this method to T2D, CAD, and PCa tests for 60,000 individuals sequenced using Helix’s Exome+. Utilizing this dataset, which targets exons and common variants, we show that an individual’s genotype can often score in close proximity to a category boundary with confidence intervals that span different categories, leading to a higher likelihood of an inconclusive result. We also report how altering acceptable risk thresholds affects the tradeoff between sensitivity and accuracy, such that a more restrictive cutoff produces more inconclusive results. Finally, we applied this method to standard coverage whole genome sequencing (WGS) and low coverage WGS, showing that it can be applied across multiple sequencing data types.

A- A+

PgmNr 169: Sibling difference analyses reveal polygenic risk score confounding.

Authors:
P.K. Joshi [1]; P.R.H.J Timmers [1]; D.W. Clark [1]; J.F. Wilson [1,2]

View Session Add to Schedule

Affiliations:
1) Centre for Population Health Sciences, Univ Edinburgh, Edinburgh, United Kingdom; 2) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom

---

Motivation: Polygenic risk scores are increasingly discussed for predicting disease risk and phenotypes of interest (1). However, it is currently unknown how much of their predictive ability is due to causal mechanisms, and how much is due to confounding by socioeconomic status, ancestry, assortative mating and genetic nurture (2). Analysis of the effects of polygenic risk scores within sibling pairs (PRSsib) cannot plausibly be confounded by any postulated confounder we are aware of. At the same time, commercial ventures have started looking into in vitro genetic embryo selection amongst sibling embryos (IGES (3)). PRSsib analysis can thus disentangle causality, confounding and genetic nurture, and reveal the magnitude of plausible effects of IGES.

Method: Published independent summary statistics were used to analyse the effect of PRSsib in 20,000 siblings from UK Biobank and compared with conventional population PRS analyses for height, educational attainment, low density lipoprotein and cardiovascular disease using the best p-value threshold method under PRSice (4).

Findings: We find the relative effect(ratio/se) of PRSsib as against PRS on the selected outcome is 95%/5% for height and 103%/4% for LDL , but 47%/4% for educational attainment and 147%/58% (underpowered) for prevalent heart disease, implying that ~50% of PRS effects for education may be due to confounding or genetic nurture.

Conclusions: For some traits, polygenic risk scores appear to track associations that are due to both genetic causes and confounding (most obviously genetic nurture). For these traits, IGES, even if it were granted ethical approval (which we oppose for complex traits), is likely to be less successful in predicting differences than in the population.

1. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018 Sep;50(9):1219–24.

2. Kong A, Thorleifsson G, Frigge ML, Vilhjalmsson BJ, Young AI, Thorgeirsson TE, et al. The nature of nurture: Effects of parental genotypes. Science. 2018 Jan 26;359(6374):424–8.

3. Polygenic Risk Scores and Genomic Prediction: Q&A with Stephen Hsu [Internet]. GEN - Genetic Engineering and Biotechnology News. 2019

4. Choi SW, O’Reilly P. PRSice 2: POLYGENIC RISK SCORE SOFTWARE (UPDATED) AND ITS APPLICATION TO CROSS-TRAIT ANALYSES [Internet]. Vol. 29, European Neuropsychopharmacology. 2019. p. S832.

A- A+

PgmNr 170: Pervasive hidden founder effects in large population scale biobanks impact polygenic load and risk estimation.

Authors:
G.M. Belbin [1,2]; A. Moscati [2]; I. Overcast [3]; M. Daya [4]; G. Wojcik [5]; N. Zaitlen [6]; I. Mathieson [7]; C.R. Gignoux [4]; E.E. Kenny [1,2]

View Session Add to Schedule

Affiliations:
1) Center for Genomic Health, Icahn School of Medicine at Mount Sinai, NY; 2) Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, NY; 3) Biology Department, Graduate Center of the City University of New York, New York, NY; 4) Department of Medicine, University of Colorado Denver, Aurora, CO; 5) Stanford University, Stanford CA; 6) Department of Neurology, UCLA, Los Angeles CA; 7) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia

Recent work has highlighted limitations in the transferability of polygenic risk scores (PRS) across different populations, even at an ultra-fine scale (i.e. within Finland). Here we investigate the impact of founder effects on polygenic risk. By examining patterns of cryptic relatedness we have previously identified known and hidden founder effects both in a diverse biobank from New York City (BioMe; N~50,000), with >25% of participants belonging to one of 7 founder populations, and in the UK Biobank (UKBB; N~500,000), where we observe the same for >7% of participants falling into 18 founder populations. We note that a small number of founder populations are self-identified canonical founder populations in these biobanks (e.g. Ashkenazi Jews (AJ), Puerto Rican, South Asian), but many others can only be detected via wholescale analysis of patterns of cryptic relatedness (e.g. Garifuna in NYC and assimilated Jewish populations in the UK).

To explore the impact of founder effects on PRS we employ a framework that utilizes both simulations and empirical data. First we leverage genomic data from BioMe for the canonical AJ founder population (N=3560), along with non-Ashkenazi Europeans (on-AJ; N=4137). To test for differences in polygenic load between groups, we simulated effect sizes onto genomic data across a range of heritabilities (h^2 from 0.3 to 0.9) and genetic architectures (from 50 to 10000 causal sites) and noted significant differences in polygenic load for 29/49 scenarios (study-wise significance threshold p<0.001, two-sided T-test)). We also used real world summary statistics to estimate PRS in AJ versus Non-AJ, and note significant differences in PRS distribution for 15/16 traits we tested (e.g. anorexia, two-sided T-test p<2.2e-16). Further, we could extract trait information from electronic health records and demonstrate attenuated risk prediction in AJ vs Non-AJ populations (e.g. serum trigylcerides;pearson’s r Non-AJ=0.22; AJ=0.17). We are currently extending this framework to other the founder populations we have detected in BioMe and UKBB, and preliminary data supports the general phenomenon we observed in the AJ population across many other founder populations.

Together this suggests that pervasive, and often hidden, founder effects can impact both polygenic load and risk estimation, with important implications for the generation and implementation of PRS within large-scale, population-based biobanking initiatives.

A- A+

PgmNr 171: Machine-learning based deconvolution of biobank-driven GWAS data with 170,000 individuals enlightens the finest-scale genetic, evolutional, and polygenic risk score divergence within Japanese population.

Authors:
J. Hirata [1,2]; S. Sakaue [1,3,4]; K. Suzuki [1]; M. Akiyama [3,5]; M. Kanai [1,3,6]; M. Hirata [7]; K. Matsuda [8]; Y. Murakami [9]; Y. Kamatani [3,10]; Y. Okada [1,11]

View Session Add to Schedule

Affiliations:
1) Department of Statistical Genetics, Osaka University Graduate School of Medicine, Osaka, Japan; 2) Pharmaceutical Discovery Research Laboratories, TEIJIN PHARMA LIMITED, Hino, Japan; 3) Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; 4) Department of Allergy and Rheumatology, Graduate School of Medicine, the University of Tokyo, Japan; 5) Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan; 6) Department of Biomedical Informatics, Harvard Medical School, Boston, MA; 7) Laboratory of Genome Technology, Institute of Medical Science, the University of Tokyo, Tokyo, Japan; 8) Department of Computational Biology and Medical Sciences, Graduate school of Frontier Sciences, The University of Tokyo, Tokyo, Japan; 9) Division of Molecular Pathology, Institute of Medical Science, the University of Tokyo, Tokyo, Japan; 10) Kyoto-McGill International Collaborative School in Genomic Medicine, Sakyo-ku, Kyoto, Japan; 11) Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan

---

We humans are unprecedentedly thriving mammals covering surprisingly large areas on earth. The key aspect of this success is the diversity encoded in the genome, which makes us adapt to local environment. While the large-scale landscape of positive selection and human adaptation has been decoded, we have yet to know whether the subtle genetic differences within a single population can be disentangled, and whether they have impact on complex traits. To address this, we applied a series of machine-learning methods (PCA, t-SNE, UMAP and PCA-UMAP) to large-scale biobank-driven genomic data in the Japanese population (n = 169,719), which offers a unique opportunity to study the genetic structure within the understudied non-European population consisting of thousands of islands. A novel ML method, PCA-UMAP, conspicuously disentangled the finest-scale genetic structures of islandic subpopulations located only several dozens of kilometers apart. ADMIXTURE analysis revealed that those subpopulations consisted of quite different ancestral components. Phylogenetic analysis together with genetic data from worldwide populations further elucidated the demographic history of the subpopulations within Japanese islands. A genome-wide selection scan (XP-EHH) showed that the islandic subpopulations were under distinct and unique selection pressure, with the identification of in total 36 loci under selective sweeps with genome-wide significance (P< 5×10^{-8}). Importantly, stratified linkage disequilibrium score regression (S-LDSC) analysis revealed that these loci were enriched in heritability of complex human traits, suggesting such traits as driving forces of natural selection in the Japanese population. Finally, we performed phenome-wide polygenic risk score (PRS) analyses from the genome-wide association studies on 67 complex traits. We observed the PRS divergence between the deconvoluted subpopulations, which could be both a result of polygenic adaptation and a result of biases from the uncorrected structure, in a trait-dependent

manner. Such uncorrected structure could be a critical pitfall in clinical application of PRSs. Our study is instrumental for the upcoming clinical application of PRSs to be truly beneficial and individualized.

A- A+

PgmNr 172: Optimisation of polygenic risk scores across 16 common diseases using cross-trait and cross-ethnic information.

Authors:
V. Plagnol; E. Kraphol; P. Sorensen; C.C. Spencer; A. Heger; M. Sivley; R. Moore; G. Lunter; M. Weale; P. Donnelly

View Session Add to Schedule

Affiliation: Genomics plc, United Kingdom

Genetic predictors that aggregate the effect of common variants into polygenic risk scores (PRS) are attracting the attention of the clinical community, with potential for improved patient management and disease prevention. Whether these approaches effectively translate into effective clinical practice will largely depend on the predictive accuracy of these scores.

Current strategies for PRS construction rely on a well powered genome-wide association study (GWAS) for the target trait. However, much information relevant to prediction is encoded in association signals that are below genome-wide significance threshold. For such variants, independent information provided by related but distinct traits, as well as the functional annotation of each variant, can impact the variant selection through improved fine-mapping. This in turn impacts the PRS weight associated with each variant. In addition, the covariance structure of effect sizes across studies can also refine effect size estimates and PRS weights. Lastly, the inclusion of non-European data can point to disease associated variants that are rare in European descent individuals, and that would have been otherwise missed.

Motivated by these observations, we have defined a statistical framework that leverages cross-trait and cross-ethnic information to improve PRS accuracy. Using summary statistics from more than 10,000 traits that have been assembled and curated by Genomics plc, we jointly fine-mapped causal variants. We maximised ethnic diversity in the underlying GWAS to improve prediction accuracy across non-european ethnic groups. We then integrated that information to optimise a set of PRS computed for 16 common diseases.

We evaluated each of these PRS in the UK Biobank, providing a systematic assessment of the ability to predict common diseases based on currently best available GWAS data. We found a general trend where the PRS show higher odds ratio in earlier onset cases, reflecting a stronger genetic load in thes cases. The cross-trait information provided a measurable improvement upon the single study approach, with higher benefit observed in more polygenic traits such a coronary artery disease. For well studied traits such as breast cancer, the inclusion of non-European data recovered half of the gap in variance explained between European and East Asian/South Asian populations, with smaller gains observed for African-American/African samples.

A- A+

PgmNr 173: Screening human embryos for polygenic traits has limited utility.

Authors:
S. Carmi [1]; E. Karavani [1]; O. Zuk [2]; D. Zeevi [3]; G. Atzmon [4,5,6]; N. Barzilai [4,5]; N.C. Stefanis [7,8,9]; A. Hatzimanolis [7,8,9]; N. Smyrnis [7,8]; D. Avramopoulos [10,11]; L. Kruglyak [3,12,13]; M. Lam [14,15,16]; T. Lencz [14,15,17]

View Session Add to Schedule

Affiliations:
1) Public Health, The Hebrew University of Jerusalem, Jerusalem, Israel; 2) Statistics, The Hebrew University of Jerusalem, Jerusalem, Israel; 3) Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA; 4) Medicine, Albert Einstein College of Medicine, Bronx, NY, USA; 5) Genetics, Institute for Aging Research, Albert Einstein College of Medicine, Bronx, NY, USA; 6) Biology, Faculty of Natural Science, University of Haifa, Haifa, Israel; 7) Psychiatry, National and Kapodistrian University of Athens Medical School, Eginition Hospital, Athens, Greece; 8) University Mental Health Research Institute, Athens, Greece; 9) Neurobiology Research Institute, Theodor-Theohari Cozzika Foundation, Athens, Greece; 10) Psychiatry, Johns Hopkins University School of Medicine, MD, Baltimore, USA; 11) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 12) Biological Chemistry, University of California, Los Angeles, Los Angeles, CA, USA; 13) Howard Hughes Medical Institute, University of California, Los Angeles, Los Angeles, CA, USA; 14) Psychiatry Research, Zucker Hillside Hospital, Glen Oaks, NY, USA; 15) Psychiatric Neuroscience, Feinstein Institute for Medical Research, Manhasset, NY, USA; 16) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 17) Psychiatry, Hofstra Northwell School of Medicine, Hempstead, NY, USA

Background: Genome-wide association studies have led to the development of polygenic score (PS) predictors that explain increasing proportions of the variance in human complex traits. In parallel, progress in preimplantation genetic testing now allows genome-wide genotyping of embryos generated via *in vitro* fertilization (IVF). Jointly, these developments suggest the possibility of screening embryos for polygenic traits such as height or cognitive function. However, no published study has evaluated the expected outcomes of embryo screening, which hampers attempts for a well-informed discussion of the ethical, legal, and societal issues associated with such a procedure.

Methods: We used simulations, theory, and real data to evaluate the potential gain of PS-based embryo selection, defined as the expected difference in trait value between the top-scoring embryo and an average, unselected embryo.

Results: The gain increases very slowly with the number of embryos, but more rapidly with increased variance explained by the PS. Given currently available polygenic predictors and typical IVF yields, the average gain due to selection would be ≈2.5cm if selecting for height, and ≈2.5 IQ (intelligence quotient) points if selecting for cognitive function. These mean values are accompanied by wide confidence intervals due to random assortment, unaccounted-for genetic factors, and environmental factors. In real data drawn from 28 nuclear families with up to 20 offspring each, the offspring with the highest PS for height was the tallest only in 7 families, was shorter than the family average in 5 families, and was on average 3cm shorter than the tallest child.

Discussion: With current predictive accuracy of complex quantitative traits, PS-based embryo selection has limited utility; however, gains will become larger with increasing GWAS sample sizes. In practice, utility will be limited by assortative mating, aneuploidy, implantation failure, lower PS accuracy in non-European populations, and selection for multiple traits.

A- A+

PgmNr 174: Rapid whole genome sequencing (rWGS) impacts resource utilization and improves management of critically ill infants with congenital heart disease.

Authors:
N. Sweeney [1,2,3]; N. Nahas [1]; S. Chowdhury [1]; S. Caylor [1]; J. Caciki [1]; S. Batalov [1]; Y. Ding [1]; N. Veeraraghavan [1]; D. Dimmock [1]; S. Kingsmore [1]; Rady Children’s Institute for Genomic Medicine

View Session Add to Schedule

Affiliations:
1) Rady Children's Institute for Genomics Medicine, San Diego, California.; 2) Rady Children's Hospital, San Diego, California.; 3) University of California San Diego, La Jolla, California.

Introduction: Congenital heart disease (CHD) is the most common congenital anomaly and a major cause of infant morbidity and mortality. Mortality is highest in infants with underlying genetic conditions but the molecular basis of only ~20% of CHD is ascertained using current standard of care genetic methodology. The hospital cost for the management of children and adults with CHD has progressively increased. Recent studies have shown that whole genome sequencing (WGS) leads to a 4-fold increase in the diagnostic rate over CMA alone and 2-fold increase over CMA plus targeted gene sequencing. **Methods**: Twenty-four critically ill infants underwent rapid WGS (rWGS) to ~45X coverage per IRB protocol. Variants were curated by referencing databases of genetic variation, gene and disease information, and a software tool was used for pathogenicity and disease causality assessment. Confirmation of diagnostic genotype(s) was by Sanger sequencing and/or array CGH. **Results**: Forty-six percent (11/24) of these children received actionable rWGS results while rate of diagnosis via oligo-SNP array was only 5% (p=0.0033, 95% CI 14.64-60.52). Rate of diagnosis by oligo-SNP array plus targeted gene panels was 9.5% (p=0.0077, 95% CI 10.07-56.77). Eighty-two percent (9/11) obtained a diagnosis that explained their cardiac and associated phenotype. The diagnoses ranged from disease process limited to the cardiovascular system like Left Ventricular Noncompaction to syndromes affecting multiple organ systems, like Coffin Siris, Kabuki and CHARGE syndrome. Change in management occurred in 100% (11/11) of patients and included listing for cardiac transplantation, transfer to a pediatric lung transplant center, avoidance of intraoperative cholangiogram, medication changes, enlistment of additional subspecialists and initiation of palliative care. There was a statistically significant decrease in average daily hospital costs comparing the time periods prior to rWGS results to the time period post rWGS results in the whole cohort of patients ($(F_{1,19})$=6.333, p=0.021). **Conclusions:** rWGS proved superior to current conventional genetic testing in the rate of diagnosis in critically ill children with CHD and provided timely actionable information that improved medical management. There is a strong signal that rWGS leads to decrease hospital spending in critically ill children with CHD.

A- A+

PgmNr 175: Polygenic architecture of computationally derived aortic diameter from 20,939 British adults predicts the risk for aortic aneurysm and dissection.

Authors:
C. Tcheandjieu [1,2]; K. Xiao [1,2]; H. Tejeda [1,2]; E. Ingelsson [1]; J. Fries [3,4]; J. Priest [1,2]

View Session Add to Schedule

Affiliations:
1) Cardiovascular Institute, Stanford School of Medicine, Stanford, CA.; 2) Department of pediatric Cardiology, Stanford School of Medicine, Stanford, CA.; 3) Department of Computer Science, Stanford University, Stanford, CA.; 4) USACenter for Biomedical Informatics Research, Stanford University, Stanford, CA

Enlargement of the aorta and aortic aneurysms are important risk factors for aortic dissection, a leading cause of death in the developed world. While Mendelian genetics are known to play an important role in disorders of the aorta, the contribution of common variation to clinical disease is not known.

Using Hough Transform and Watershed image segmentation, we performed automated extraction of Ascending Aortic Diameter (AAoD) at the level of the pulmonary artery bifurcation from cardiac MRI of 29,783 individuals from the UK Biobank. We studied the relation between the standardized derived-AAoD and 49 million genotyped and imputed SNPs using linear regression with adjustment for age, sex, body surface area, and 10 genomic principal components in 20,939 unrelated white British; and applied a Bonferroni-threshold for multiple testing correction. We developed a polygenic risk score for AAoD (PRS(AAoD)) including 111,288 independent SNPs (correlation <0.8) to predict the risk of aortic aneurysm and dissection and validated in an independent set of 314,614 unrelated white British in the UK Biobank.

We identified 62 new independent SNPs across 26 loci, and 33 protein-coding genes strongly associated with AAoD. The top associated SNP, rs6974735 (p=2x10^{-32}), is located in the promoter of Elastin (ELN); a Mendelian-disease gene playing an important structural role in the architecture of large vessels. Several other genes including PRDM9, FLNB1, EDN1, ABCC9, GATA2 involved in vascular smooth muscle formation were also associated with AAoD. Moreover, eQTLs in associated loci were significantly enriched in the aorta (p=1x10^{-05}) and blood vessels (p=2x10^{-04}). Finally, a higher PRS(AAoD) was associated with an increased risk of aortic aneurysm (OR=1.38 [1.13-1.69] per SD-increase in PRS, p=0.001) and aortic dissection (OR=1.10 [1.03-1.17] per SD-increase in PRS, p=0.005).

Using automated techniques in image processing, we measured AAoD from cardiac MRI in individuals from the general population identifying several susceptibility loci for AAoD, including genes known to play major roles in vascular smooth muscle formation and structure. Our findings suggest a strong polygenic component to morphology of the aorta, and may contribute to the early detection of patients at risk of aortic aneurysm and dissection.

A- A+

PgmNr 176: Large-scale GWAS and multi-omic follow-up reveal new mechanisms for heart failure.

Authors:
M. Arvanitis[1, 2]; Y. Zhang[3]; B. Ren[3]; W.S. Post[2]; A. Battle[1]

View Session Add to Schedule

Affiliations:
1) Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA; 2) Department of Medicine, Division of Cardiology, Johns Hopkins University, Baltimore, MD, USA; 3) Ludwig Institute for Cancer Research, San Diego, CA, USA

Heart failure is a complex disease whose clinical manifestations depend on an incompletely understood interplay between genetic predisposition and environmental insults. The largest genome wide association study to-date has only revealed 1 locus, with little understanding of mechanisms underlying genetic risk.

We conducted a genome wide association meta-analysis of over 10,000 Heart Failure cases and 400,000 controls spanning seven different cohorts of European ancestry and identified one known (*PITX2*) and two novel loci (*ACTN2* and *ABO*) associated with Heart Failure. Genetic correlation revealed a shared heritable component between Heart Failure and other cardiovascular diseases as well as an association between Heart Failure and immune-related traits (Asthma, Sarcoidosis, Rheumatoid Arthritis).

One of our identified genome wide significant loci (*ACTN2*) showed broad associations with both ischemic and non-ischemic Heart Failure, a predominant influence in Heart Failure with reduced over preserved ejection fraction and no association with Heart Failure risk factors (Hypertension, Coronary Disease, Atrial Fibrillation). These findings point to a putative role for this locus at influencing myocardial function reserves in response to a variety of cardiac muscle insults. Using chromatin state data from Roadmap Epigenomics we show that two variants within the *ACTN2* locus overlap a muscle specific active enhancer region. We validate this finding with a human embryonic stem cell to cardiomyocyte differentiation model and show that a candidate casual variant overlaps an ATAC-seq peak that emerges during cardiomyocyte differentiation. Using Hi-C on the same cardiomyocyte model we show that the ATAC-seq peak binds to the promoter of *ACTN2*, a gene that encodes for an actin binding protein, known to lead to cardiomyopathy when affected by loss of function mutations. The other novel locus (*ABO*) has an immune specific epigenetic profile with H3K4me1 and H3K27ac peaks, in primary hematopoietic stem cells. The GWAS signal reveals strong evidence of colocalization with *ABO*expression quantitative trait loci (eQTL) in eQTLGen whole blood.

Using an integrative approach we provide evidence for new heritable mechanisms linked to Heart Failure pathogenesis. We reveal a novel locus that predisposes to both ischemic and non-ischemic Heart Failure via cardiac muscle specific effects and provide evidence for broadly shared heritability between Heart Failure and immune mechanisms.

A- A+

PgmNr 177: Overlap of fetal-specific cardiac regulatory variants and GWAS lead variants supports fetal origins of cardiovascular disease.

Authors:
K.A. Frazer; M.K.R. Donovan; W.W. Young Greenwald; D. Jakubosky; P. Benaglio; E.N. Smith; A. D’Antonio-Chronowska; D. D’Antonio

View Session Add to Schedule

Affiliation: UC San Diego, SAN DIEGO, California.

We sought to examine if genetic factors in the developing fetus affect cardiovascular disease later in life by identifying fetal-specific expression quantitative trait loci (eQTLs) and determining if they overlap GWAS variants for adult cardiac diseases. We recently established the utility of iPSC-derived cardiovascular progenitor cells (CVPCs) for identifying cardiac regulatory variants*. Here, we derived iPSC-CVPCs from 180 individuals and showed that their transcriptomes are more similar to fetal heart than to adult cardiac tissues. We examined the cellular composition of 8 iPSC-CVPCs via scRNA-seq and found they were comprised of two cardiac cell types: cardiomyocytes (CMs) and epicardium derived cells (EPDCs). We established and used CM-specific and EPDC-specific expression signatures to deconvolute the bulk RNA-seq data for all 180 iPSC-CVPCs, determining the relative ratios of both cell types in all samples. We leveraged these data in combination with WGS to perform an eQTL analysis, resulting in the discovery of eQTLs mapping to 13,449 eGenes. Considering cell type compositions across all 180 iPSC-CVPCs using an interaction test, we observed that 2,051 (15%) eGenes were associated with one cell type, indicating eQTLs function differently between CMs and EPDCs. We next investigated if fetal eQTLs have the same associations with gene expression as adult cardiac tissue eQTLs. We performed a colocalization analysis on all eGenes with eQTLs in GTEx adult cardiac tissues, and found that 4,387 (33%) fetal eGenes were not shared. To assess whether fetal-specific eQTLs are associated with complex adult cardiac traits, we colocalized eQTLs with summary statistics from GWAS (pulse rate and myocardial infarction) and found 10 fetal-specific eGenes, including *CLPTM1*, which is associated with congenital malformations. These results show that analysis of the eQTLs in iPSC-CVPCs identifies cardiac disease GWAS variants that are active in the fetal but not adult heart, indicating that they play a role in development. Our findings provide genetic evidence supporting the fetal origins of the cardiovascular disease hypothesis and highlight the importance of investigating genetic associations across stages of development (i.e. fetal and adult tissues) to fully understand the genetic underpinnings of complex traits and disease.

* Benaglio P. et al. Allele-Specific NKX2-5 Binding Underlies Multiple Genetic Associations with Human EKG Traits. 2019 Nat. Gen. *In Press.*

A- A+

PgmNr 178: Comprehensive epigenomic and transcriptomic profiling of human embryonic heart reveals the regulatory landscape of heart development and implicates noncoding sequences in congenital heart defects.

Authors:
J. VanOudenhove [1]; A. Wilderman [1, 2]; T. Yankee [1, 2]; J. Cotney [1, 3]

View Session Add to Schedule

Affiliations:
1) Genetics and Genome Sciences, UConn Health, Farmington, CT; 2) Graduate Program in Genetics and Developmental Biology, UConn Health, Farmington, CT; 3) Institute for System Genomics, UConn, Storrs, CT

Spatiotemporal regulation of gene expression during development can occur through tissue-specific gene regulatory sequences known as enhancers. Genome-wide association studies (GWAS) show that a majority of disease-associated variants are enriched in enhancers and there is growing evidence that enhancer alterations can result in birth defects or predisposition to disease later in life. Congenital heart defects (CHDs) are the most common form of birth defect, effecting 1 in 100 live births. Of the 80% of cases of CHDs without a pathogenic copy number variation, less than 10% have identifiable loss of function mutations in genes. Therefore, noncoding mutations could be a substantial contributor to the remaining unknown cases. However, our limited understanding of the language of the noncoding genome and lack of functional annotations from early developing heart prevent causative assignment of noncoding variation in CHDs. To address this, we created a comprehensive catalog of chromatin state annotations during critical stages of human heart development (4 to 8 post conception weeks). We generated genome-wide profiles of seven post-translational histone modifications (H3K4me1-3, H3K27ac, H3K27me3, H3K9me3, and H3K36me3) for two human embryonic hearts from each of nine distinct Carnegie stages (CS13-14, CS16-21, and CS23) for a total of 144 primary ChIP-seq datasets. Using imputation followed by segmentation with a 25 state chromatin model developed by Roadmap Epigenome we identified 177,412 heart enhancers. Of these 34,034 had not been previously annotated in Roadmap. We identified 92% of all validated heart positive enhancers from the Vista Enhancer Browser (n=281), a 7.5-fold enrichment versus active enhancers lacking activity in the heart ($p=2.2x10^{-16}$). To explore the impact these chromatin states have on gene expression, we generated bulk strand-specific RNA-seq data at comparable time points for three embryonic hearts. We find enhancers are enriched near genes expressed more strongly in the heart than other tissues. Finally, we evaluated the enrichment of heart trait-associations from the GWAS Catalog in enhancers from our data and found significant enrichment of SNPs associated with CHDs, electrocardiogram measures, aortic root size, and atrial fibrillation. Our functional annotations will allow for better interpretation of whole genome sequencing data of patients with heart related conditions and advance the field of personalized genomic medicine.

A- A+

PgmNr 179: Dynamic genetic regulation of gene expression during cellular differentiation.

Authors:
K. Rhodes [1]; R. Elorbany [2,3]; B. Strober [4]; N. Krishnan [5]; K. Tayeb [6]; A. Battle [4,5]; Y. Gilad [1,7]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics, University of Chicago, Chicago, Illinois.; 2) Interdisciplinary Scientist Training Program, University of Chicago, Chicago, Illinois.; 3) Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, Illinois.; 4) Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland.; 5) Department of Computer Science, Johns Hopkins University, Baltimore, Maryland.; 6) Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, Maryland.; 7) Department of Medicine, University of Chicago, Chicago, Illinois.

Genetic regulation of gene expression is dynamic, changing through time during important processes like differentiation and development. Yet, almost all studies of the genetics of gene regulation involve data collected at a single time point, usually from adult individuals. We hypothesized that some unexplained disease-associated loci may act by influencing the dynamics of gene expression during cellular differentiation. To address this hypothesis, we differentiated induced pluripotent stem cells (iPSCs) from 19 human individuals to cardiomyocytes and collected RNA-seq data every 24 hours during the 16 day differentiation. We then combine this high-resolution timecourse dataset with existing genotype information to explore temporal patterns of gene regulation. We find that differentiation time is the main driver of variation within the dataset and that this variation is genetically controlled. We also identify trajectories of gene clusters and cell lines throughout the differentiation. We then identified dynamic eQTLs, or eQTLs variants whose effect size changes throughout the differentiation time course. We identified 550 significant dynamic eQTLs (eFDR <=.05) whose effects change linearly through time. We find that these linear dynamic eQTLs are enriched at annotated enhancer elements in iPSCs and primary heart tissue. Furthermore, they are enriched for genes related to dilated cardiomyopathy (p=.001, Fisher’s exact) and include variants associated with cardiac electrophysiology. We also find 693 significant dynamic eQTLs (eFDR<=.05) whose effects change nonlinearly through time. Of these nonlinear dynamic eQTLs, 28 have their strongest effect in the middle of the timecourse and would not have been identified using data from mature cell types. One such variant has previously been associated with BMI, highlighting that transient regulatory effects of dynamic eQTLs may have phenotypic consequences. The data generated in this study represent a novel resource for investigating temporal mechanisms underlying disease associations, and our results demonstrate that dynamic genetic effects on gene regulation may play a role in human disease.

PgmNr 180: Population-specific causal disease effect sizes at loci impacted by negative selection.

Authors:
H. Shi [1,3]; S. Gazal [1,3]; M. Kanai [3,5,6,7,8]; A.P. Schoech [1,2,3]; Y. Okada [8,9]; S.S. Kim [1,3,4]; A.L. Price [1,2,3]

View Session Add to Schedule

Affiliations:
1) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 2) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA; 5) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 6) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 7) Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; 8) Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan; 9) Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan

Trans-ethnic genetic correlations are significantly less than 1 for many diseases and complex traits, complicating precision medicine for diverse populations and motivating efforts to understand the evolutionary processes that contribute to population-specific causal disease effect sizes (Martin et al. 2019 Nat Genet). We developed a new method, trans-ethnic stratified LD score regression (X-LDSC), to stratify squared trans-ethnic genetic correlation by functional annotation using genome-wide summary association statistics. X-LDSC regresses products of z-scores from each population on trans-ethnic stratified LD scores, generalizing stratified LD score regression; we stratify the *squared* trans-ethnic genetic correlation by functional annotation to robustly handle noisy heritability estimates. We confirmed via extensive simulations that X-LDSC yields approximately unbiased estimates of squared trans-ethnic genetic correlation and its enrichment/depletion in functional annotations.

We applied X-LDSC to 29 diseases and complex traits with summary association statistics in East Asians (average N=91K) and Europeans (average N=164K) available from Biobank Japan, UK Biobank, and other sources (average trans-ethnic genetic correlation $r_g = 0.83$ (s.e. 0.01) and $r^2_g = 0.69$ (0.02)), meta-analyzing enrichments across traits. We determined that squared trans-ethnic genetic correlation was 0.79x (s.e. 0.02) smaller than the genome-wide average at SNPs in the top quintile of background selection statistic (McVicker et al. 2009 PLoS Genet), implying more population-specific causal effect sizes at regions impacted by negative selection; results were similar for several other selection-related annotations. Accordingly, squared trans-ethnic genetic correlation was 0.42x (s.e. 0.10) smaller than the genome-wide average in coding regions, with analogous depletions in other functional regions of the genome that are impacted by negative selection.

We investigated the evolutionary processes driving population-specific disease effect sizes by extending the Eyre-Walker model of coupling between fitness and trait effects (Eyre-Walker 2010 PNAS) to two populations. Simulations under this model demonstrated that population-specific causal effect sizes at loci impacted by negative selection are expected under models with differential selection coefficients in the two populations.

A- A+

PgmNr 181: Strong selection for cold and carbohydrate diets in ancient Eurasian.

Authors:
Y. Souilmi [1]; R. Tobler [1]; A. Johar [1]; M. Williams [1]; S. Grey [2,3]; J. Teixeira [1]; A. Rohrlach [4,5]; G. Gower [1]; C. Turney [6]; M. Cox [7]; W. Haak [5]; C.D. Huber [1]; A. Cooper [1]

View Session Add to Schedule

Affiliations:
1) Australian Centre for Ancient DNA, The University of Adelaide, Adelaide, South Australia, Australia; 2) Transplantation Immunology Group, Immunology Division, Garvan Institute of Medical Research, Australia; 3) St Vincent's Clinical School, Faculty of Medicine, UNSW, Darlinghurst, New South Wales, Australia; 4) ARC Centre of Excellence for Mathematical and Statistical Frontiers, The University of Adelaide, Adelaide, South Australia 5005, Australia; 5) Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany; 6) Chronos Radiocarbon Facility and Palaeontology, Geobiology and Earth Archives Research Centre (PANGEA), University of New South Wales, Sydney, NSW 2052, Australia; 7) Statistics and Bioinformatics Group, School of Fundamental Sciences, Massey University, Palmerston North 4410, New Zealand

The last decade saw dramatic growth in modern and ancient human genomic datasets particularly from western Eurasia reshaping our understanding of human demographic history, migrations, and introgression. Considering the major socio-cultural and environmental changes anatomically modern humans experienced since the “Out of Africa”, only a few studies have used ancient genomes to explore human adaptation. Consequently, our present understanding is predominantly based on genomes from modern populations, where the apparent lack of strong genomic signatures of selection has led to views that recent phases of adaptation involved selection on polygenic traits or from standing genetic variation.

To investigate signatures of selection in ancient human populations, we examined more than 1000 ancient western Eurasian genome-wide datasets. Our results suggest that adaptation via hard selective sweeps has played a more substantial role in recent human history than previously appreciated. In addition, we show that large-scale genetic mixing during and after the Bronze Age has masked genomic signatures of selection in modern populations. Furthermore, we find that selective sweeps aggregate in interacting gene cohorts directly involved in responding to pathogenic pressure increased carbohydrate metabolism and cold adaptation. This represents a new model of hard sweep-driven polygenic selection that challenges conventional models of adaptation that propose either rare hard selective sweeps or subtle and widespread polygenic selection.

We further determined the onset of the selective pressure, thus, creating a detailed reconstruction of adaptation during the past 50,000 years. We found that cold adaptation played an important role after the out-of-Africa expansion, as well as oxidative stress response following the transition from hunter-gatherer to farming lifestyles and carbohydrate-based diets in the past 10,000 years. This study highlights the unique potential of ancient genomes to unmask previously hidden evolutionary histories, reveals relevant medical genomic loci, and offers a tool to reduce the number of causal genes in genome-wide association studies.

A- A+

PgmNr 182: The effects of mutation subtype on the allele frequency spectrum and population genetics inference.

Authors:
K. Liao [1]; J. Carlson [2,4]; S. Zöllner [1,3]; The BRIDGES Consortium

View Session Add to Schedule

Affiliations:
1) Biostatistics, University of Michigan, Ann Arbor, Michigan.; 2) Bioinformatics, University of Michigan, Ann Arbor, Michigan.; 3) Psychiatry, University of Michigan, Ann Arbor, Michigan.; 4) Genome Sciences, University of Washington, Seattle, Washington.

The allele frequency spectrum (AFS) is a summary of genetic variation in a population that is commonly used for population genetics inference such as testing for selection and inferring demographic history. Current AFS-based methods consider all sites equally and thus use a single AFS to perform inference. However, mutational mechanisms such as biased gene conversion and mutation rate heterogeneity are known to operate on specific types of sites. As a result, different sites have distinct allele frequency spectrums and it is unclear how much population genetics inference is biased by failing to account for these differences.

Using the BRIDGES dataset containing whole genome sequencing data for 3556 unrelated Europeans at ~10x coverage, we first classified each point mutation into one of 96 trinucleotide mutation subtypes (MSTs) from their adjacent nucleotides. For each MST, we then constructed an AFS and summarized it using the proportion of singletons to doubletons and Tajima's D. This analysis reaffirmed previous findings that MSTs with higher mutation rates have a lower proportion of singletons. In addition, systematically higher Tajima's D values for A>G MSTs (enriched for high-frequency variants) and lower values for C>A MSTs (enriched for low-frequency variants) are consistent with the effects of biased gene conversion.

We then assessed how population genetics inference is affected by MSTs having distinct allele frequency spectrums. Using DaDi, we inferred demographic history assuming an exponential growth model for each MST-specific AFS. The estimated effective population size and time since the population started growing varied drastically across MSTs by almost two orders of magnitude. In addition, using only intergenic sites, we split the genome into non-overlapping 100kb windows and tested for selection in every window. We then determined the empirical p-value for every window and assessed the relationship between the local site composition and empirical p-value. This analysis showed that regions abundant in certain MSTs have an inflated rate of false positive signals of selection up to two times larger than the traditional 0.05 level. Overall, our results show significant impacts of mutation rate heterogeneity and biased gene conversion on common methods of population genetics inference. As a result, these methods should be modified to account for mutation subtypes in order to avoid biased results.

A- A+

PgmNr 183: Identifying and characterising germline hypermutators.

Authors:
J. Kaplanis [1]; E. Prigmore [1]; P. Short [1]; J. Korbel [2]; M. Hurles [1]

View Session Add to Schedule

Affiliations:
1) Wellcome Trust Sanger Institute, Hinxon, United Kingdom; 2) EMBL, Heidelberg, Germany

The human germline mutation rate is known to vary between individuals. Most of the variation in the population is explained by parental age, but little is known about rare outliers with extreme mutation rates. We have tackled this problem from three distinct approaches. First, we identified germline hypermutators and sought genetic causes for this phenotype. Germline hypermutators are individuals who are born with an unusually large number of *de novo* mutations (DNMs). A large number of DNMs increases the chance of having a damaging mutation in a developmentally important gene. We selected ten individuals in the Deciphering Developmental Disorders study with extreme number of DNMs in their exome and whole genome sequenced both them and their parents. One of these individuals has 277 DNMs genome-wide, four times more DNMs than expected. The excess mutations all phased paternally, are distributed across the genome and have a distinctive mutational signature that does not correspond to any known mutational signatures in cancer. To investigate possible causes, we looked in the father's genome for rare damaging variants in DNA repair genes and found a protein truncating variant (PTV) in the gene *NTHL1*and a missense variant in *BRCA2*. *BRCA2*is a well-known cancer gene that is involved in the double strand break repair pathway. *NTHL1*is a DNA glycosylase involved in the base-excision repair pathway and has been associated with elevated mutation rates in multiple cancers.

Second, we focused on a specific mutator gene, *MBD4,*in which PTVs have been shown to be associated with a four-fold increase in C>T mutations at CpG dinucleotides in tumours. This same CpG mutation signature accounts for ~16% of de novo mutations in the germline which raises the question of whether *MBD4*PTV germline carriers also show an increased number of C>T germline mutations in their offspring. We identified 14 paternal carriers of *MBD4* PTVs within the DDD study and whole genome sequenced the trios. We will present an analysis of the properties of the DNMs in these and determine whether they are indeed hypermutated.

Finally, we will present results on a test across genes known to be associated with DNA repair to identify genetic variants that affect the mutation rate on individuals in the DDD. Our analyses provide new insights into the role of genetic variation on the human germline mutation rate and uncover possible genetic causes of germline hypermutation.

A- A+

PgmNr 184: The evolutionary impact of an ancient deletion polymorphism in the human growth hormone receptor gene.

Authors:
O. Gokcumen; S. Resendez; M. Saitou; K. Dean; F. Wu; X. Mu

View Session Add to Schedule

Affiliation: University at Buffalo, Buffalo, NY

The growth hormone receptor (GHR) gene codes the receptor protein for the growth hormone. GHR is highly conserved among mammals. However, its third exon is polymorphically deleted with 30% allele frequency in the human population. This deletion has previously been associated with human height, longevity, the age to menarche, and response to growth hormone treatments, among other developmental and metabolic phenotypes. However, the evolutionary history and the mechanism through which the polymorphism affect phenotype remains unknown.

Our lab has previously reported that Neanderthal and Denisovan genomes also carry the deletion allele (GHRd3), suggesting that the GHRd3 and ancestral alleles were segregating in the human-Neanderthal-Denisovan ancestral population. Using population genetics analysis, we resolved the haplotype architecture of the locus harboring GHRd3. This analyses suggested that non-neutral, population-specific evolutionary forces shaping the variation in this locus. Further, we identified single nucleotide variants that tag the deletion allele (R2 with GHRd3 > 0.96). We showed that the haplotype carrying the GHRd3 allele is significantly associated ($p < 3e-09$) with 'Standing height' in a cohort of 450,000 British individuals available through UK Biobank dataset.

To investigate the phenotypic impact of GHRd3 at the developmental and molecular level more thoroughly, we constructed a CRISPR-Cas9 based mouse where we deleted exon 3, modeling the human polymorphism. Using this model, we showed that there is a differential rate of growth between GHRd3 and wildtype mice. Moreover, comparative RNA-sequencing analyses from liver tissues showed that GHRd3 affects the expression of genes enriched for metabolic processes.

Taken together, our study suggests the non-neutral evolution of GHRd3 in humans and verified previous associations with developmental phenotypes. Furthermore, we identified novel biological targets of GHRd3, affecting metabolic pathways in the liver. Our integrative approach sheds new light on the evolutionary impact of ancient exonic structural variants and highlights GHRd3 as a potential candidate for metabolic disease susceptibility.

A- A+

PgmNr 185: Reconstructing Denisovan anatomy using DNA methylation maps.

Authors:
D. Gokhman [1]; N. Mishol [1]; M. de Manuel [2]; D. de Juan [2]; J. Shuqrun [1,3]; T. Marques-Bonet [2,4]; Y. Rak [5]; L. Carmel [1]

View Session Add to Schedule

Affiliations:
1) Department of Genetics, The Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem 91904, Israel.; 2) Institute of Evolutionary Biology (UPF-CSIC), 08003 Barcelona, Spain.; 3) Alpha matriculation program.; 4) Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain.; 5) Department of Anatomy and Anthropology, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, 6997801, Israel.

Denisovans are an extinct group of humans whose morphology remains unknown. Here, we present a method for reconstruction of anatomical profiles using ancient DNA methylation patterns. Our method is based on linking hypermethylation events (which we parallel with reduced activity) to known loss-of-function phenotypes in modern humans. We then apply a unidirectionality approach to determine whether all regulatory changes associated with the activity level of a gene and with a phenotypic alteration point to the same direction of change. We tested the performance of this method by reconstructing Neanderthal and chimpanzee skeletal profiles and matching them against their known morphology. We obtained >85% precision in identifying divergent traits. We further confirm this approach by testing it on chimpanzee gene expression and histone modification data, resulting in similar prediction accuracy (>80%). We then apply this method to the Denisovan and offer a putative morphological profile. We suggest that Denisovans likely shared with Neanderthals traits such as a projecting face, robust jaws, a sloping forehead and a wide pelvis. We also identify Denisovan-derived changes, such as an increased dental arch, and lateral cranial expansion. Our predictions match the only confirmed and morphologically informative Denisovan bone to date, as well as the Xuchang skull, which was suggested by some to be a Denisovan. We conclude that DNA methylation can be used as a tool to reconstruct anatomical features, including some that do not survive in the fossil record.

A- A+

PgmNr 186: Combining ATAC-seq and high-throughput reporter assays to identify sequences, factors, and genetic variants that regulate gene expression in different environmental contexts.

Authors:
F. Luca [1]; C. Kalita [1]; A. Alazizi [1]; A. Findley [1]; X. Wen [2]; R. Pique-Regi [1]

View Session Add to Schedule

Affiliations:
1) Center for Molecular Medicine and Genetics, Wayne State University, Detroit, Michigan.; 2) Department of Biostatistics, University of Michigan, Ann Arbor, Michigan.

A large fraction of loci important in determining human traits and disease conditions are located in non-coding regions of the genome. These regions likely contain specific regulatory sequences that control gene transcription and can also interact with changes in the cellular environment (e.g. drug treatment). However, the extent to which the environment can modulate genetic effects on quantitative phenotypes is still to be defined. Here we combined ATAC-seq and a high throughput allele-specific reporter assay we recently developed (BiT-STARR-seq) to identify regulatory sequences, transcription factors and genetic variants that regulate gene expression in different environmental context. We characterized 9,263 (dexamethasone), 2,615 (copper), and 2,115 (selenium) regions with differentially accessible chromatin (FDR<10%) in response to treatment. Using BiT-STARR-seq to target 26,068 putative regulatory regions with genetic variants that are likely to impact gene regulation, we identified differential gene expression enhancer activity (FDR 10%) in LCLs treated with retinoic acid (2,173), dexamethasone (3,859), selenium (274), and caffeine (6,743). Differential enhancer regions for dexamethasone were significantly enriched in differentially accessible regions from ATAC-seq (OR=1.33, p value=0.0002). We identified thousands of regions with allele-specific enhancer activity (FDR<10%) with retinoic acid (3,756), dexamethasone (2,098), selenium (6,973), and caffeine (5,734). These regions are enriched in footprints for transcription factors identified in the matched cellular environment. For example, when cells are treated with dexamethasone, we find enrichment for ASE in CNOT3 footprints, a transcription factor involved in early B-cell development and downregulated in response to dexamethasone in LCLs. Our results demonstrate that ATAC-seq, together with an improved footprint model, and the targeted approach of BiT-STARR-seq are excellent tools for rapid profiling of transcription factor binding activity to study cellular regulatory response to the environment and molecular mechanisms underlying GxE.

A- A+

PgmNr 187: Exon-skipping regulation in complex disease.

Authors:
R. Liu [1,2]; A. Byrnes [1,2]; M. Daly [1,2]; H. Huang [1,2]

View Session Add to Schedule

Affiliations:
1) Stanley Center, Broad Institute of MIT and Harvard, Cambridge, MA.; 2) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA

mRNA isoforms can be generated from a single gene locus through alternative splicing. Abnormality in alternative splicing has been linked to many human disorders. To date, the extent to which the genetic regulation of splicing contributes to complex diseases/traits has not been systematically evaluated. Here, using RNA-seq data and summary statistics from GWAS of complex diseases/traits, we present a study to identify genomic variants regulating exon-skipping with the goal to understand their contribution to complex diseases/traits.

We designed a novel approach to identify variants regulating exon-skipping events. We applied this method on the imputed genotypes and the RNA-seq data across 34 tissues, available from the GTEx v7 release. For each tissue, we found between 50 and 592 exon-skipping events regulated by genomic variants. We performed fine-mapping on these associations and mapped them to 95% credible sets. Per tissue, between 14 and 278 exon-skipping regulations were mapped to a set with≤5 variants.

We found that these regulatory variants extensively shared across tissues and those with high causal probability are located close to the exons being regulated. To understand the contribution of these regulatory variants to human complex traits/diseases, we fine-mapped genome-wide significant loci associated with 23 complex traits/diseases. On the genome-wide scale, we noted a clear disease-tissue specificity. For example, the causal variants for autoimmune disorders (IBD, CD, and RA) are significantly enriched as the splicing regulatory variants in immune-relevant tissues (whole blood), and variants causal to type-2-diabetes significantly enriched as the splicing regulatory variants in the pancreas. Specifically, we found 11 variants have non-trivial posterior probability(>10%) for both the regulation of exon-skipping and complex traits/disorders. For example, rs11589479 regulates the skipping of the 19th exon in *ADAM15*, a susceptible gene for Crohn's Disease, in all tissues tested.

In summary, we designed a novel approach to identify variants regulating exon-skipping events and demonstrated these variants contribute to the human complex disorders in a disease-tissue specific manner. Results from this approach provided critical insights into the functional mechanism of the disease genetic associations and contributed to our understanding of the genetic architecture of human complex disorders.

A- A+

PgmNr 188: Genetic basis of alternative polyadenylation is an emerging molecular phenotype for human traits and diseases.

Authors:
L. Li [1,2]; Y.P. Gao [3]; F.L. Peng [2]; EricJ. Wagner [4]; W. Li [1,2]

View Session Add to Schedule

Affiliations:
1) Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX USA; 2) Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX USA; 3) Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX USA; 4) Department of Biochemistry & Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA

Genome-wide association studies have identified thousands of non-coding variants that are statistically associated with human traits and diseases. However, the vast majority of variants occur in non-coding regions, thus posing a significant challenge for elucidating the molecular mechanisms by which these variants contribute to diseases and phenotypes. Alternative Polyadenylation (APA) occurs in approximately 70% of human genes and substantively impacts cellular proliferation, differentiation and tumorigenesis. But the roles of genetic determinants of APA in various human tissues and their association with phenotypic traits and diseases have not been systematically examined. To obtain insights into the genetic basis of APA regulation in human tissues, we used our DaPars (Dynamic analyses of alternative polyadenylation from RNA-seq) algorithm to construct a landscape of tissue-specific human APA events. We describe the first atlas of human 3’-UTR alternative polyadenylation Quantitative Trait Loci (3’aQTLs), i.e. ~0.4 million genetic variants associated with APA of target genes across 46 Genotype-Tissue Expression (GTEx) tissues from 467 individuals. 3’aQTLs are significantly enriched in 3’UTRs region and largely distinct from other QTLs such as eQTLs and splicing-QTLs. Using multivariate adaptive shrinkage (mash) method, we estimate the effect size of 3’aQTLs shared across 46 tissues and found that although 85.9% of tissues had 3’aQTLs with the same sign, only 15.7% were shared with 3’aQTLs of a similar magnitude. Mechanistically, 3’aQTLs could alter polyA motifs and RNA-binding protein binding sites, leading to thousands of APA changes. Importantly, co-localization analyses found that 16.1% of trait-associated loci co-localize with one or more 3’aQTLs in human tissues. Furthermore, very few of the 3’aQTL–co-localizing trait-associated loci overlapped with eQTLs, indicating that 3’aQTLs and eQTLs are largely mutually independent. We also mapped 3’aQTLs in 13 major immune cell types, providing additional resolution for interpretation of autoimmune disease-associated loci. Collectively, the genetic basis of APA (3’aQTLs) thus represent a novel molecular phenotype to explain a large fraction of non-coding variants and to provide new insights into complex traits and disease etiologies.

A- A+

PgmNr 189: *N*^{6}-methyladenosine (m^{6}A) methylation of mRNAs makes a large contribution to genetics of common diseases.

Authors:
K. Luo [1]; Z. Zhang [2,3]; M. Qiu [2,3]; H. Shi [2,3]; Y. Zou [4]; G. Wang [1]; A. Zhu [2,3]; M. Qiao [5]; Z. Li [1,6]; M. Stephens [1,4]; X. He [1]; C. He [2,3]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA; 2) Department of Chemistry, Department of Biochemistry and Molecular Biology, and Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL 60637, USA; 3) Howard Hughes Medical Institute, The University of Chicago, Chicago, IL 60637, USA; 4) Department of Statistics, The University of Chicago, Chicago, IL, 60637, USA; 5) Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; 6) Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, Zhejiang 325000, China

N^6-methyladenosine (m^{6}A) plays a critical role in regulating various aspects of mRNA metabolism and translation in eukaryotes. Despite rapid progress in the field, we know very little about what may affect m6A deposition specificity and level, and connections between m^{6}A and common human diseases. In this work, we mapped tens of thousands of genetic variants associated with m^{6}A variation (m^{6}A-QTLs) in 60 human lymphoblastoid cell lines. This large list of QTLs enabled us to better understand the mechanisms regulating m^{6}A by analyzing sequence features near these QTLs. In addition, these QTLs served as “natural perturbations” of m^{6}A and allowed us to gain understanding of the downstream effects of m^{6}A, at both molecular and phenotypic levels. Our comprehensive analysis with this unique approach revealed a number of important insights into m^{6}A biology and its contribution to complex diseases:
1) We found m^{6}A consensus motif (RRACH), RNA secondary structure, RNA binding proteins (RBP) and transcriptional processes all play important roles in regulating m^{6}A installation and levels. We highlight several RBPs, all implicated in pre-mRNA splicing, as novel regulators of m^{6}A.
2) It is commonly believed that the m^{6}A modification promotes translation and mRNA decay via recognition by reader proteins. Our joint analysis of QTL data of m^{6}A and related molecular traits suggests that the downstream effects of m^{6}A are likely heterogeneous, and depend on the specific contexts of m^{6}A sites, e.g. binding of RBPs with regulatory functions at nearby regions.
3) We found that m^{6}A-QTLs are highly enriched in GWAS signals of complex traits. And the signal remains even if we include expression and splicing QTLs in the joint analysis. We were able to leverage this enrichment to discover putative causal variants and genes of immune-related diseases. Our results thus uncover variants perturbing mRNA modification as a new class of genetic variants of complex human diseases.

A- A+

PgmNr 190: The contribution of miRNA regulation to inter-individual and inter-population variability in immune responses.

Authors:
M. Rotival [1]; M. Silvert [1,2]; K.J. Siddle [3,4]; J. Pothlichet [1,5]; H. Quach [1]; L. Quintana-Murci [1]

View Session Add to Schedule

Affiliations:
1) Institut Pasteur, CNRS UMR 2000, 75015 Paris, France; 2) Sorbonne Universités, École Doctorale Complexité du Vivant, 75005 Paris, France; 3) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA; 5) Present address: DIACCURATE, Institut Pasteur, 75015 Paris, France

---

While the regulatory role of micro-RNAs (miRNAs) in the immune response is increasingly recognized, their contribution to the intra- and inter-population differences of immune responses is poorly characterized. Furthermore, the contribution of miRNA modifications, i.e. shift of miRNA start/end site and terminal adenylation or uridylation, to the regulation of immune response remains unclear. To understand how population variability in miRNAs expression contributes to shape immune responses, we stimulated monocytes from 200 individuals of African and European ancestry with 3 different TLR ligands (LPS, PAM3CSK4 and R848 activating TLR4, TLR1/2 and TLR7/8 respectively) and Influenza A virus (IAV). Through miRNA sequencing of a total of 977 samples of resting and activated monocytes at 6h of stimulation, we show that monocytes display a strong shift in their miRNA profiles, with 80 miRNAs (12% of all miRNAs) being up-regulated upon stimulation (1% FDR, 28 with $\log_2$FC > 1), as well as a strong reduction in 3' uridylation ($p<4.6\times10^{-7}$) in response to viral stimuli, leading to shorter miRNA isoforms (isomiRs). Interestingly, the intensity of the miRNA response to immune stimulation is markedly different between populations, with 95 miRNAs showing population differences ($p^{bonf}<0.01$) in expression (N=70), isomiR levels (N=18) or both (N=7). Among these miRNAs, we find key modulators of the immune response, such as mir-155, mir-146a or mir-222. Integrating miRNA levels with whole genome genotyping and exome sequencing data, we identify 101 miR- and 28 isomiR-QTLs, including rs2910164 responsible for a shift in start site of mir-146-3p that leads to the loss of 73% of its targets. Overall, we find that miR-QTLs are largely shared across conditions and account for up to 60% of population-differences in expression of the miRNAs they regulate. Finally, integrating miRNA data with RNA-seq from the same individuals, we have quantified the relative impact of transcription and miRNA regulation on the variability of immune responses. In doing so, we show that miRNAs account for ~16% of the expression variability of their target genes (and up to 60%), with 89% of miRNA-target interactions being stimulation condition-specific. Overall, these results highlight the importance of miRNAs and their isomiRs in driving the diversity of immune responses, both between individuals and populations.

PgmNr 191: The genetic architecture of DNA replication timing in human pluripotent stem cells.

Authors:
Q. Ding [1]; A. Bracci [1]; M. Edwards [1]; M. Hulke [1]; Y. Hu [1]; Y. Tong [1]; X. Zhu [2]; J. Hsiao [2]; C. Charvet [1]; S. Ghosh [3,4,5]; R. Handsaker [3,4]; M. Stephens [2]; S. McCarroll [3,4]; K. Eggan [3,5,6]; Y. Gilad [2]; F. Merkle [7]; J. Gerhardt [8,9]; D. Egli [10]; A. Clark [1]; A. Koren [1]

View Session Add to Schedule

Affiliations:
1) Department of Molecular Biology and Genetics, Cornell University, Ithaca NY 14853; 2) Department of Human Genetics, University of Chicago, Chicago IL 60637; 3) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge MA 02142; 4) Department of Genetics, Harvard Medical School, Boston MA 02115; 5) Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge MA 02138; 6) Howard Hughes Medical Institute, Harvard University, Cambridge MA 02138; 7) Wellcome Trust - Medical Research Council Institute of Metabolic Science, University of Cambridge, Cambridge, United Kingdom; 8) Ronald O. Perelman and Claudia Cohen Center for Reproductive Medicine, Weill Cornell Medicine, New York NY 10065; 9) Department of Obstetrics and Gynecology, Weill Cornell Medicine, New York NY 10065; 10) Department of Pediatrics, Columbia University, New York NY 10032

In eukaryotic cells, DNA is replicated according to a strict spatiotemporal program that intersects with chromatin structure and gene regulation. This program is fundamentally mediated by the differential activation of DNA replication origins along S phase. The mechanisms controlling human replication origin specification and activation timing are poorly understood. To elucidate genetic mechanisms regulating DNA replication timing, we inferred replication timing for 108 human embryonic stem cell lines (ESCs) and 194 induced pluripotent stem cell lines (iPSCs) from deep whole-genome sequencing. We then associated inter-individual variation in replication timing with genetic polymorphisms to identify more than 1,500 *cis*-acting replication timing quantitative trait loci (rtQTLs) – base-pair-resolution determinants of replication origin activities. We reveal a “histone code” specifying replication origin locations, composed of H3K4me3, H3K9me3, H3K36me3, H3K56ac, and one or more variable acetylations, which characterizes and can even predict replication origins across human cell types (ESC, iPSC, and lymphoblastoid cell lines). These histone marks likely specify origin locations via recruitment of their respective demethylases and acetyltransferases. Active histone marks and pluripotency-related transcription factors, including OCT4, NANOG and EP300, promote early replication. Conversely, CTCF, RAD21, and YY1 repress origin activity when bound to DNA. Furthermore, up to six *cis*-rtQTLs additively cooperate to determine local replication timing. In summary, we reveal that human replication timing is controlled by a multi-layer mechanism that operates on target DNA sequences, composed of dozens of effectors working combinatorically, and follows principles analogous to transcriptional regulation: a histone code, activators and repressors, and a promoter-enhancer logic.

A- A+

PgmNr 192: A multi-model approach for deep functional variomic profiling of ASD-associated *PTEN* missense mutations identifies multiple molecular mechanisms underlying protein dysfunction.

Authors:
F. Meili [1,2,3]; K. Post [1,2,3]; M. Belmadani [1,4]; P. Ganguly [1,2,3]; R. Dingwall [1,2,3]; T. McDiarmid [1,2,5]; C. Harrington [1,2,3]; M. Edwards [1,2,3]; B. Young [1,3]; A. Niciforovic [1,2,3]; B. Callahan [1,4]; S. Rogic [1,4]; W. Meyers [1,2,3]; A. Cau [1,2,3]; T. O'Connor [1,2,3]; C. Rankin [1,2,5]; S.X. Bamji [1,2,3]; D.W. Allen [1,2,3]; C. Loewen [1,3]; P. Pavlidis [1,4]; K. Haas [1,2,3]

View Session Add to Schedule

Affiliations:
1) University of British Columbia; 2) Djavad Mowafaghian Centre for Brain Health; 3) Department of Cellular and Physiological Sciences; 4) Department of Psychiatry; 5) Department of Psychology

Functional variomics promises to provide the foundation for personalized medicine by linking genetic variation to disease expression, outcome and treatment, yet its utility is dependent on the assays employed to adequately evaluate mutation impact on all aspects of a protein’s function. In order to fully assess the impact of 104 missense (MS) and nonsense (NS) variants of PTEN (phosphatase and tensin homologue deleted on chromosome 10) associated with autism spectrum disorder (ASD), somatic cancer and PTEN hamartoma tumor syndrome (PHTS), we take a deep phenotypic profiling approach using 18 assays in 5 model systems spanning phylogeny, including yeast, fly, worm, rat primary neuronal culture, and a human cell line. Our approach allows correlation of phenotypes in assays ranging from molecular function to neuronal morphogenesis and behavior. Most assays strongly correlate with measures of lipid phosphatase function, while others implicate additional PTEN functions. We find that protein instability is a major mechanism of dysfunction for variants altering amino acids across the entire protein structure. Other variants, located in discrete functional domains including the N-terminus substrate binding and the catalytic domains exhibited impacts ranging from loss of function to antimorphic/dominant negative phenotypes independent of effects on stability. Results indicate that 31 of the 48 variants tested from ASD individuals are likely pathogenic and 2 are likely benign and not causal.

A- A+

PgmNr 193: Integrating molecular and clinical phenotypes towards clinical insight on genotype-phenotype relationships for missense germline *PTEN* variation.

Authors:
S. Thacker [1, 2, 3]; T. Mighell [1, 4]; I.N. Smith [2]; M. Seyfi [2]; B.J. O'Roak [5, 9]; C. Eng [2, 3, 6, 7, 8, 9]

View Session Add to Schedule

Affiliations:
1) These authors contributed equally; 2) Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA; 3) Cleveland Clinic Lerner College of Medicine, Cleveland, OH 44195, USA; 4) Neuroscience Graduate Program, Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR 97239; 5) Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR 97239; 6) Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH 44195, USA; 7) Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine; Cleveland, OH 44106, USA; 8) Germline High Risk Cancer Focus Group, Comprehensive Cancer Center, Case Western Reserve University School of Medicine; Cleveland, OH 44106, USA; 9) Joint senior authors

PTEN is a nexus between cancer and autism spectrum disorder (ASD), where germline mutations lead to either condition. The single gene-disparate phenotype observation is a clinical challenge that begs resolution. In this study, we sought to integrate molecular phenotype data (i.e. downstream information from all possible variants) from a clinically, rigorously annotated cohort of *PTEN* Hamartoma Tumor Syndrome (PHTS) individuals (N = 256) in order to differentiate between clinical outcomes (i.e. cancer or ASD). We found that the effect a *PTEN* missense variant has on the lipid phosphatase activity (i.e. fitness score) of PTEN explains 40% of the variation in disease burden (i.e. Cleveland Clinic Score) and 22% of the variation in head circumference in adult PHTS patients. Furthermore, we found that abundance score (i.e. the steady-state expression of a PTEN mutant) explains 9% of both the variation in disease burden and head circumference in adult PHTS patients. These findings lend critical insight into how variants modulating the lipid phosphatase activity of PTEN explain the penetrance and expressivity of PHTS overgrowth phenotypes. Notably, the above relationships only hold for missense, not nonsense, variation, suggesting that missense variants can participate in different PTEN biology. Despite the clinically informative nature of molecular phenotypes, we found that in isolation, they could not predict clinical outcome; there were no differences in their distributions between PTEN-ASD and PTEN-cancer groups ($P>0.05$). Subsequently, modeling aggregated molecular phenotype data for pathogenicity and clinical outcomes for missense variants showed high (AUC = 0.92) and moderate accuracy (AUC = 0.77), respectively. The accuracy of these models confirms molecular phenotypes inform clinical outcomes. Finally, we applied unsupervised k-means clustering to identify six distinct groups of PHTS individuals: two ASD clusters, two cancer clusters, and two mixed clusters (between sum-of-squares variance = 50.6%). Together, our data illustrate genotype-specific effects influence clinical outcomes, highlighting the deeply shared biology of ASD and cancer and indicating the existence of divergent points. Obtaining comprehensive molecular phenotype data will inform precision care for PHTS individuals (versus a cohort).

A- A+

PgmNr 194: A humanized yeast assay to define dominant-negative properties of pathogenic alleles in highly conserved, essential genes.

Authors:
R. Meyer-Schuman [1]; A. Antonellis [1,2]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA; 2) Department of Neurology, University of Michigan, Ann Arbor, MI 48109, USA

Aminoacyl-tRNA synthetases (ARSs) are ubiquitously expressed, essential enzymes that charge tRNA molecules with amino acids. Heterozygosity for mutations in five ARS genes causes dominant axonal peripheral neuropathy, which is characterized by impaired motor and sensory function. These mutations decrease ARS function *in vitro*, reduce viability in yeast and worm complementation assays, and are dominantly toxic to mouse and worm neurons. The allelic spectrum mainly comprises missense mutations; null alleles have not been observed in patients, but are observed in unaffected populations, indicating that haploinsufficiency is not the disease mechanism. All five ARS genes implicated in dominant neuropathy encode enzymes that function as homodimers. This observation raises the possibility of a dominant-negative effect, in which inactive mutant subunits dimerize with wild-type subunits in a heterozygous cell and reduce overall ARS activity below the level required for peripheral nerve function. To test the dominant-negative properties of pathogenic ARS alleles, we developed a humanized yeast assay. Here, we co-express pathogenic and wild-type ARS alleles and measure yeast growth as a proxy for ARS function; we began by studying alanyl-tRNA synthetase (*AARS*). We first confirmed that the wild-type human *AARS* open reading frame rescues deletion of the yeast ortholog *ALA1*, and that the pathogenic mutation R329H *AARS* does not, supporting previous data that R329H is a loss-of-function allele. We then compared the growth of yeast co-expressing wild-type and R329H AARS (WT/R329H) to yeast co-expressing wild-type and a null AARS allele (WT/-). WT/R329H yeast formed fewer colonies than WT/-, indicating a dominantly toxic growth defect associated with R329H. We then investigated remaining WT/R329H colonies for acquired genetic traits that conferred a growth advantage, such as a relative increase in WT copy number, reduced R329H copy number, or mutations that reduce R329H transcript abundance. To further define this dominant toxicity as a dominant-negative dimerization with WT AARS, we placed R329H in *cis* with an engineered mutation that reduces AARS dimerization, and evaluated subsequent yeast growth. Here, we present our unpublished data that establishes a humanized yeast assay for defining the dominant-negative properties of ARS alleles. This work provides a framework for using yeast as a model to identify human dominant-negative alleles in conserved, essential genes.

A- A+

PgmNr 195: Modeling VUSes in a clinically relevant time frame.

Authors:
M. Bainbridge [1]; J. Friedman [1]; C. Hopkins [2]; K. McCormick [2]; T. Brock [2]; D. Dimmock [2]; S. Kingsmore [1]; C. Hobbs [1]

View Session Add to Schedule

Affiliations:
1) Rady Children's Institute for Genomic Medicine, Rady Children's Hospital, San Diego, California.; 2) Nemametrix, Eugene, OR

Variants of uncertain significance (VUSes) can impede patients getting a diagnosis and are often interpreted by clinicians as benign. Testing VUSes in model systems can take months or years – which is clinically irrelevant for acutely ill patients. Here we present a platform to rapidly assess the pathogenicity of VUSes in C. elegans in <14 days and provide a model system for assessing drug response on patient specific variants.

In this system the human ortholog replaces the worm gene in its native locus ensuring the gene is expressed at similar levels, in similar tissues and at similar times to the native gene. Variants of interest can then be rapidly installed in the humanized worm using CRISPR/Cas9. Multiple phenotypic read outs including electrophysiology, peristalsis, and locomotion can then be used to assess pathogenicity.

We assessed *CACNB4*, the cause of episodic ataxia and epilepsy, in this model system. We first demonstrate that a knockout (KO) of the worm ortholog *ccb-1* is lethal which can be rescued by human-*CACNB4* but fails to rescue when the gene contains a pathogenic frameshift mutation (p.Q204fs). Next, we show that a known disease causing mutation (p.C104F) increases the rate of the worm's feeding by ~15%, ($p<0.05$) whereas a benign mutation (p.M219V) has no significant effect on feeding speed. We then assessed a complex patient derived VUS (p.HYP484R) and found feeding speed and patterns were nearly identical to the pathogenic allele ($p<0.05$). Lastly, we assessed whether Diamox, a drug used to treat patients with the p.C104F mutation can be efficacious in treating our patient with the p.HYP484R mutation.

Rapid *in vivo* modeling of patient derived variants compliments ultra-rapid clinical sequencing in acutely ill patients. In addition to assessing pathogenicity it can also be used to assess drug response and potentially help guide care.

A- A+

PgmNr 196: Phenotype-driven blood biomarker and muscle functional genomics significantly increase neuromuscular disease diagnosis of >400 patients by resolving VUSs and multi-gene inheritance.

Authors:
S. Chakravorty [1]; K. Berger [2]; S.P.V. Shenoy [1]; B.R.R. Nallamilli [3]; L. Rufibach [4]; S. Shira [4]; M. Wicklund [5]; M. Harms [6]; T. Mozaffar [7]; D. Arafat [2]; G. Gibson [2]; M. Hegde [1,3]; Jain COS Consortium

View Session Add to Schedule

Affiliations:
1) Human Genetics and Pediatrics, Emory University, Atlanta, GA 30322, USA; 2) Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA; 3) PerkinElmer Genomics, Global Laboratory Services, Waltham, Massachusetts, USA; 4) Jain Foundation Inc., Seattle, WA, USA; 5) University of Colorado Denver, Neurology, CO, USA; 6) Columbia University Neurology, Institute for Genomic Medicine, NY, USA; 7) University of California Irvine Neurology, CA, USA

50-70% of inherited neuromuscular disease (NMD) patients remain undiagnosed even after DNA testing, a barrier for disease management and trial enrolment. Recently, using a 35 gene NGS panel on 4656 limb-girdle muscular dystrophy (LGMD)-suspected patients, we identified the major hurdles were: a) lack of genotype-phenotype correlation knowledge in heterogeneous NMDs such as in LGMDs with >30 monogenic subtypes, b) high prevalence (72%) of variants of uncertain significance (VUSs), c) >30% of all patients had pathogenic variant(s) or VUSs in ≥2 genes (multi-genic), and d) the lack of less-invasive biomarker-driven approaches. Our objective was to functionally resolve VUSs and multi-genic cases, by combining clinical-genetic data with an array of functional assays using minimally-invasive biomarker approach or target muscle biopsies to understand geno-phenotype correlations. For example, Dysferlinopathy caused by variants in the *DYSF* gene is the second-most prevalent LGMD-subtype. We show in a cohort study of 394 Dysferlinopathy-suspected NMD patients, a significant increase in diagnostic yield from 35% to 85% by using a combinatorial blood biomarker-driven CD14+ monocyte assay and whole blood targeted RNA-sequencing along with clinical correlation. This significant increase in diagnostic yield was achieved using a tiered analytical approach through reclassification of VUSs, identification of novel causal variants and pathomechanisms including abnormal splicing or allele or gene expression. Importantly, this further facilitated patient stratification, and mapping *DYSF* variant landscape. Moreover, in clinically-suspected Pompe cases (considered a monogenic disorder), we identified a high prevalence (72 cases) of one pathogenic variant or VUS each in *GAA* and another LGMD recessive gene, which are being functionally resolved using blood or muscle-based enzyme assays, immunoblotting, and RNA-Seq. We are also performing clinical correlation analysis in the multi-genic cases and identifying mostly unusual phenotypic expressions with either very-slow or -fast disease progression. Using clinically driven-functional omics, we are resolving the nature of the defect in different pathways that lead to multi-genic contribution in NMDs. Our results show the importance of using a multi-tiered diagnostic approach that includes biomarkers, omics platforms and geno-phenotype correlations not only for precision medicine diagnostics but also for testing clinical trial efficacy.

A- A+

PgmNr 2418: A high-throughput assay to test functional significance of *BRCA1* RING domain missense substitutions: Methodology, calibration, and results.

Authors:
K.A. Clark [1]; A.M. Paquette [1]; K. Tao [1]; J.S. Rosenthal [1]; A.K. Snow [3]; J. Unger [1]; K.M. Boucher [2]; J. Gertz [1]; A. Thomas [2]; K.E. Varley [1]; S.V. Tavtigian [1]

View Session Add to Schedule

Affiliations:
1) Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT; 2) Department of Internal Medicine, Division of Epidemiology, University of Utah, Salt Lake City, UT; 3) Department of Population Sciences, University of Utah, Salt Lake City, UT

The increase in targeted sequencing of disease-predisposition genes has led to a concomitant increase in variants of uncertain significance (VUS). This is especially true for *BRCA1*, the first defined breast cancer susceptibility gene. Development and calibration of functional tests to evaluate the repertoire of VUS for *BRCA1* is invaluable to provide information on clinically actionable variants. Several functional assays for BRCA1 have been developed that either measure a direct activity of BRCA1 (e.g. binding its partner, BARD1), or evaluate downstream effects of BRCA1 function, such as cell survival or DNA repair. A cursory analysis of these current assays raises the concern that the high throughput assays provide an indirect measure of function, and the more direct assays are less compatible with a high throughput approach. To address these shortcomings, we developed a mammalian two hybrid assay (M2H) to evaluate VUS in the RING domain of BRCA1. Our M2H assay has several features that improve upon current assays: 1) Oligo array libraries containing single nucleotide changes to produce every possible amino acid change that can arise from a single base pair substitution in the BRCA1 RING domain; 2) Four different silent barcodes for each variant, enabling a requirement for 3-4 replicates of each VUS; 3) BRCA1 mutants are expressed individually in the reporter line, and then pooled for sorting, allowing for massively parallel analysis of VUS; 4) Development of a dual red/green fluorescent M2H reporter cell line, enabling cell sorting into 6 bins on increasing red and green fluorescence; and 5) Variants within each red/green bin are identified by RNA-seq to associate individual sequence variants with their M2H activity. The subsequent key step is to calibrate the assay to generate values that can be incorporated into the American College of Medical Genetics (ACMG) variant classification framework. We are currently testing calibration models, using a set of classified variants. Using the Expectation-Maximum Likelihood (EM) algorithm, we fit a linear model to the estimated mean of the distribution of the variant sequences across the 6 red/green bins. This allows us to estimate the odds of pathogenicity for each variant, which is the key component of Bayesian “Integrated Evaluation.” We hope that our well calibrated high-throughput assay will inform the classification of *BRCA1* VUS and lay the ground work for the evaluation of VUS in additional genes.

A- A+

PgmNr 198: Utilizing RNA sequencing for the diagnosis of unsolved cases of Cornelia de Lange syndrome and related neurodevelopmental disorders.

Authors:
S. Rentas [1]; K. Rathi [2]; M. Kaur [3]; P. Raman [2]; I.D. Krantz [3,4]; M. Sarmady [1]; A. Abou Tayoun [5]

View Session Add to Schedule

Affiliations:
1) Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA; 2) Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA; 3) Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA; 4) Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 5) Al Jalila Children’s Specialty Hospital, Dubai, UAE

Testing for suspected Mendelian disorders includes genomic analysis by targeted gene panels and exome sequencing. Limitations of these methods lead to 60% of cases remaining without a molecular diagnosis. Additional new testing modalities are needed for non-diagnostic cases. RNA sequencing (RNA-seq) has recently been shown to boost diagnostic yield in neuromuscular diseases through identification of abnormal mRNA splicing and expression. However, utility of RNA-seq in neurodevelopmental disorders has not been fully investigated, likely due to the concern of sourcing relevant tissue for RNA analysis. Here we establish a protocol to test neurodevelopmental disorders by RNA-seq and address the lack of accessibility of appropriate tissue by showing patient derived B lymphoblastoid cell lines (LCLs) perform better than whole blood by highly expressing hundreds of genes implicated in Mendelian neurodevelopmental disorders.

To demonstrate the clinical utility of this approach, we focused on patients without a molecular diagnosis but clinically presenting with Cornelia de Lange Syndrome (CdLS) or related cohesinopathy. This group of disorders causes multiple developmental abnormalities due to mutations in genes involved in cohesin function and transcription. Comparison of GTEx expression data from LCLs and whole blood for 14 genes involved in these disorders showed LCLs have significantly higher expression of this gene set and they expressed the most abundant isoform found in brain for all 14 genes. LCLs further expressed 35/40 brain-expressed isoforms for these genes, indicating biologically relevant transcript diversity was captured. An RNA-seq pipeline was designed to detect exonic and splice site variants and abnormal mRNA splicing. Validation was performed on 10 CdLS LCL samples encoding a variety of pathogenic variant types. Our pipeline found all abnormally spliced transcripts and called pathogenic splice site or coding variants in all but two cases. The missed calls were due to inadequate coverage over these two regions indicating the likely added value of performing higher-depth of sequencing for this assay. Finally, we investigated five test cases with a suspected cohesinopathy but lacking molecular diagnosis and discovered de novo, abnormal splicing of *NIPBL* transcript in two cases. Altogether, our work establishes RNA-seq as a viable frontline diagnostic tool in neurodevelopmental disorders and as a vital reflexive test when DNA testing is negative.

PgmNr 199: Retrospective whole exome sequencing analysis reveals variants in BAF complex genes associated with global developmental delay, epilepsy, and dysmorphism.

Authors:
J. Lattier[1]; L. Meng[1,2]; F. Xia[1,2]; P. Liu[1,2]; W. Bi[1,2]; B. Yuan[1,2]; V. Patel[1]; D. Scott[3,4]; R. Marom[2,4]; M. Wangler[2,5]; J. Mokry[2]; C. Eng[1,2]; R. Xiao[1,2]

View Session Add to Schedule

Affiliations:
1) Baylor Genetics, Houston, TX; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 3) Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX; 4) Department of Pediatrics, Texas Children's Hospital, Houston, TX; 5) Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX

Background: Clinical whole exome sequencing (WES), an efficient method for diagnosing complex disorders, cannot immediately solve every case with a genetic etiology. Thousands of genes and their effects on human phenotype remain a mystery. Every year, our understanding of gene-disease correlation is expanding. Periodic retrospective analysis of clinical WES, using updated gene-disease associations, provides new diagnoses for previously unsolved cases. These analyses can even uncover new phenotypic patterns for previously uncharacterized genes. Over the past decade, deleterious variants in several members of the BRG1-associated factors (BAF) complex (*ARID1A, ARID1B, ARID2, SMARCA4, SMARCB1, SMARCA2, SMARCE1,* and *ACTB*) were reported to cause Coffin-Siris syndrome or similar disorders, characterized by speech delay, intellectual disability, seizures, motor delay, and facial dysmorphism. The human BAF protein complex, known as SWI/SNF in other organisms, is involved in chromatin remodeling and regulates long-term memory in mice. However, the BAF complex includes many other members whose functions and associations with disease are only recently becoming evident, such as *ACTL6A, ACTL6B, BCL11A, BCL11B, SMARCC1,* and *SMARCC2*. **Methods:** We retrospectively analyzed over 5,000 variants, including possible pathogenic variants and polymorphisms, across several genes of the BAF complex identified through WES. Sanger sequencing was performed for candidate variants in probands and parents to identify *de novo* variants. Clinical correlation was performed on a case-by-case basis. **Results:** Retrospective analyses of our clinical WES database revealed *de novo* pathogenic variants across multiple families in 6 genes of the BAF complex, *ACTL6A, ACTL6B, BCL11A, BCL11B, SMARCC1,* and *SMARCC2*, which previously were not associated with diseases at the time of original analyses. A significant phenotype correlation was identified that includes speech delay, intellectual disability, seizures, motor delay, and facial dysmorphism. **Conclusion:** Revisiting older clinical WES data to search for possible gene-disease associations for lesser known members of an otherwise well-known complex identified causative variants in a multitude of clinical cases, finally providing families with the answers which they have been seeking for years.

A- A+

PgmNr 200: Skewed X-chromosome inactivation in patients with non-diagnostic exome sequencing: Additional diagnostic yield.

Authors:
I.M. Campbell [1]; L. Dong [1]; T. Pinnheiro [2]; J. Yutz [2]; A. Ganguli [2]; E.J. Bhoj [1]

View Session Add to Schedule

Affiliations:
1) Children's Hospital of Philadelphia, Philadelphia, PA.; 2) Univeristy of Pennsylvania, Philadelphia, PA.

X chromosome inactivation is the process by which one mammalian female X chromosome is silenced such that the gene dosage is mostly equivalent to males. Selection of which X chromosome to silence is typically random. However, considerable deviation from random inactivation has been observed to occur both in the general population and among individuals with clinical phenotypes. Such skewing occurs at a tissue or organismal level as a consequence of random chance due to segregation of cells during embryogenesis, cell autonomous growth advantages or disadvantages, or rarely genetic defects of the molecular underpinnings of the inactivation process. Skewed X-inactivation has been frequently reported, both away from deleterious alleles in unaffected carriers as well as towards pathogenic variants causing expression of X-linked recessive disorders.

We hypothesized that skewed X-inactivation could identify those patients with a causative variant on the X chromosome, which would prompt additional investigation of previously dismissed variants. To this end, we enrolled 24 female subjects with negative exome sequencing for unexplained phenotypes which included developmental delay / intellectual disability, craniofacial dysmorphism, brain abnormalities, and major organ malformations.

We determined X-inactivation by CAG repeat polymorphism analysis at the *AR* locus using methylation specific restriction endonuclease. The X-inactivation in our cohort, arbitrarily expressed as the ratio of inactivation of the shorter CAG allele to the longer, ranged from 11:89 to 99:1. In comparison to a population study of 1,005 individuals, our cohort contained significantly more skewing (Kolmogorov-Smirnov test, p=0.00092). The median deviation from random was 24.5 in our cohort compared to 11.6 in the control population. Among those individuals with the highest skewing, we reviewed exome sequencing for variants potentially causing x-linked recessive phenotypes. We identified variants potentially associated with the subjects' phenotypes in genes including *AFF2*, *MED12* and *STAG2,* leading to the identifications of causative variants for these patients.

We suggest that X-inactivation testing may be a useful test to augment variant identification in negative exome or genome sequencing in females.

A- A+

PgmNr 201: Identifying diagnoses beyond the exome: Lessons from challenging cases with compelling clinical phenotypes.

Authors:
A. O'Donnell-Luria [1,2,3]; M.H. Wojcik [1,2]; K.R. Chao [1,3]; J.K. Goodrich [1,3]; L.S. Pais [1,3]; E. England [1,3]; E.G. Seaby [1,3]; A.B. Byrne [1,4,5]; B.B. Cummings [1,3]; R.L. Collins [1,3]; M. Lek [6]; L. Gallacher [7,8]; T.Y. Tan [7,8]; K.M. Bujakowska [9]; E.A. Pierce [9]; P.B. Agrawal [1,2]; C.A. Walsh [1,2]; J.M. Verboon [1,10,11]; V.G. Sankaran [1,10,11]; C. Barnett [12]; H. Scott [4,13]; W.K. Chung [14]; E.A. Estrella [2]; C.C. Bruels [15]; P.B. Kang [15,16]; S. Pajusalu [6,17]; K. Ounap [17]; A.K. Lovgren [1,3]; H.L. Rehm [1,3]; D.G. MacArthur [1,3]

View Session Add to Schedule

Affiliations:
1) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; 2) Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA; 3) Analytic and Translational Genomics Unit (ATGU), Massachusetts General Hospital, Boston, MA; 4) Department of Genetics and Molecular Pathology, Centre for Cancer Biology, SA Pathology, Adelaide, Australia; 5) School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, Australia; 6) Department of Genetics, Yale School of Medicine, New Haven, CT; 7) Victorian Clinical Genetics Services, Murdoch Children’s Research Institute, Melbourne, Australia; 8) Department of Paediatrics, University of Melbourne, Melbourne, Australia; 9) Ocular Genomics Institute, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA; 10) Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA; 11) Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA; 12) Paediatric and Reproductive Genetics Unit, Women's and Children's Hospital/SA Pathology, Adelaide, Australia; 13) Centre for Cancer Biology, University of South Australia and SA Pathology, Adelaide, South Australia, Australia; 14) Division of Molecular Genetics, Department of Pediatrics, Department of Medicine, Columbia University Irving Medical Center, New York, NY; 15) Division of Pediatric Neurology, Department of Pediatrics, University of Florida College of Medicine, Gainesville, FL; 16) Genetics Institute and Myology Institute, University of Florida, Gainesville, FL; 17) Department of Clinical Genetics, Institute of Clinical Medicine, University of Tartu and Tartu University Hospital, Tartu, Estonia

---

Next-generation sequencing has revolutionized clinical genetics, yet our ability to detect pathogenic variants remains incomplete. In many cases, clinicians have strong phenotypic or biochemical data implicating a specific genetic diagnosis but molecular testing, including gene sequencing and deletion/duplication testing, is unrevealing. Should we trust our clinical judgment? Is this testing “complete”? What should we do next?

Through the Broad Institute Center for Mendelian Genomics, we have sequenced and analyzed >5,000 rare disease families including many for which there was a strong clinical suspicion but initial negative testing. Here we review a large series of these cases for which we ultimately identified a diagnosis with exome, genome, and/or transcriptome sequencing. We characterize the spectrum of diagnoses found across ~30 cases and the reasons that the molecular diagnosis was originally missed. Conditions include many phenotypically well-characterized disorders such as Duchenne and Ullrich muscular dystrophy, Axenfeld Rieger, Meckel Gruber, and Marfan syndromes.

Variants in high GC content regions were initially missed by clinical exome (*FOXC1*). Small structural variants (including a partial exon deletion of *FBN1*) were missed on original testing. Copy neutral inversions (*DMD*) were entirely missed by Sanger sequencing, exome, array, and MLPA, and ultimately detected by capturing the breakpoints by genome sequencing. Noncoding variants that result in altered splicing including exon skipping or the creation of pseudoexons were detected by RNA-seq (*COL6A1*). In one case, an epimutation resulted in gene silencing (*MMACHC*). We find the clinical suspicion was correct in most cases, but the underlying causal variant(s) were cryptic to all standard genetic testing methodologies. A small number of cases were due to a phenocopy with a similar condition found to be the eventual diagnosis.

Overall, this cohort provides a view of the full allelic spectrum of pathogenic variants for a set of very well-characterized Mendelian phenotypes and disease genes, demonstrating several classes of pathogenic variation missed by standard genetic testing methodologies. These examples demonstrate the importance of careful clinical phenotyping in guiding the analysis of genetic results, particularly by prioritizing loci for deeper study. Finally, we describe our best practices used to identify elusive pathogenic variants to maximize diagnostic yield.

A- A+

PgmNr 202: Broad-scale untargeted metabolomic profiling improves diagnosis of inborn errors of metabolism.

Authors:
N. Liu [1]; J. Xiao [1]; C. Gijanavekar [1]; K.E. Glinton [1]; B.J. Shayota [1]; Y. Yang [1]; K.L. Pappan [2]; A.D. Kennedy [2]; M.F. Wangler [1]; L.C. Burrage [1]; F. Scaglia [1]; W.J. Craigen [1]; C. Soler [1]; L.T. Emrick [3]; F. Xia [1]; V.R. Sutton [1]; Q. Sun [1]; S.H. Elsea [1]

View Session Add to Schedule

Affiliations:
1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Metabolon Inc., Durham, NC, USA; 3) Texas Children's Hospital, Houston, TX

The patients presenting with nonspecific neurological spectrum including intellectual disability (ID) or global developmental delays (GDD) at late ages are usually screened by the first-line tests including CMA, fragile X and a traditional trio of biochemical analyses: plasma acylcarnitines (ACP), urine organic acids (UOA), and plasma amino acids (PAA). Recently, the steep rises of identified metabolic conditions increase the urgent need for a more comprehensive screening tool. Here, we present a comparison of the diagnostic rate of clinical metabolomics with traditional screening to demonstrate the power of metabolomics that holds the promise of providing new insights into human diseases and serving as a primary screening tool for IEMs. Results from clinical samples received for metabolic screening and plasma metabolomic profiling between 2014-2018 were reviewed, analyzed, and compared to determine the test performance and utility for diagnosis of IEMs. Of the 1483 individuals screened by traditional approach concomitantly, there were 20 cases identified with IEMs, giving a 1.34% diagnostic rate (with 12 different IEMs identified including one condition that is not covered by newborn screening). In a similar time period, we performed metabolomic profiling in 1817 individuals. Patients were primarily pediatric and 80% were referred for neurological phenotypes. Of the 1817 samples, metabolomic screening identified 113 cases, giving a total overall diagnostic rate of 6.22%. Of these 113 positive cases, 39 were further confirmed by targeted quantitative biochemical testing, supporting the reliability of metabolomics. In total, 54 different conditions were identified including 34 conditions that are not covered by newborn screening. In this way, different conditions were identified including, but not limited to, aminoacidopathies, organic acidemias, fatty acid oxidation disorders, vitamin deficiencies, pentose phosphate pathway disorders, peroxisomal disorders, purine disorders, and neurotransmitter abnormalities. In our series, metabolomics provided a ~2-5 fold higher diagnostic yield than the conventional screening approach and identified more metabolic disorders. We believe that these data support the utility of metabolomics in screening for IEMs in children with non-acute, non-specific neurological phenotypes.

A- A+

PgmNr 203: Cascading after peri-diagnostic cancer genetic testing: An alternative to population based screening.

Authors:
K. Offit [1,2,3]; S. Mukherjee [1]; K. Tkachuk [1]

View Session Add to Schedule

Affiliations:
1) Clinical Genetics Service, Memorial Sloan Kettering Cancer Center, New York, New York.; 2) Program in Cancer Biology and Genetics, Sloan Kettering Institute, New York, New York; 3) Departments of Medicine and Health Care Policy and Research, Weill Cornell Medical College, Cornell University, New York, New York

Background: Despite advances in DNA sequencing technology and expanded medical guidelines, the vast majority of individuals carrying pathogenic variants of common cancer susceptibility genes have yet to be identified. Population based genetic screening of healthy individuals will result in substantial costs, counseling for variants of unknown significance, and other complexities. An alternative strategy would exploit the growing trend for genetic testing at time of cancer diagnosis (PMID: 28873162) to guide therapy and prevention, combined with augmented familial diffusion or “cascade” of genomic risk information, which has been accomplished for other hereditary disorders

Methods: We modeled the time course to detect an estimated 3.9 million individuals in the U.S. with a pathogenic variant in one of the 18 cancer susceptibility genes (*APC, ATM, BRCA1/2, CDH1, FLCN, MLH1, MSH2, MSH6, NF1, PALB2, PMS2, PTEN, RET, SDHB, STK11,TP53,VHL*). We estimated the population burden of these variants using an automated curation pipeline (PMID: 30787465) in public datasets. We performed a sensitivity analysis using a multiple linear regression model of the impact of the proportion of incident cancer cases sequenced, the observed frequencies of pathogenic germline variants in cancer cases, the differential rates of diffusion of genetic information in families, and family size, on the time to detect 3.9 million heterozygotes for pathogenic variants in the 18 genes selected.

Results: The model indicates that the time to detect all inherited cancer predisposing variants in the U.S. population is most strongly impacted by family size, followed by prevalence of mutations in cancer cases, the proportion of cancer cases sequenced, and rates of cascade to first, second, and third-degree relatives. Assuming family sizes of two to four siblings per generation, 15% of incident cancer cases (the number treated at comprehensive centers) receive germline sequencing, 10% of cancer cases with germline mutations, and 50% diffusion of genetic information in families, 3.9 million individuals in the U.S. with a cancer susceptibility mutation could be detected in 5-10 years.

Conclusions: Peri-diagnostic cancer genetic testing undertaken with modest changes in medical and reimbursement guidelines and with augmented cascade of genetic information in families would achieve population wide mutation detection with less complexity and cost than sequencing of the general public.

A- A+

PgmNr 204: Evaluation of the cutting efficiencies of sgRNAs in CRISPR/Cas editing experiments utilizing droplet digital PCR (ddPCR).

Authors:
E. Cerveira [1]; M. Ryan [1]; L. Bellfy [1]; A. Mil-Homens [1]; Q. Zhu [1]; C. Lee [1,2]; C. Zhang [1,2]

View Session Add to Schedule

Affiliations:
1) The Jackson Laboratory, Farmington, Connecticut.; 2) The First Affiliated Hospital of Xi’an Jiaotong University, Xi’an, China

The introduction of CRISPR/Cas9 has allowed for ease in gene editing, but accurately determining the cutting efficiency of sgRNAs can be challenging. Traditional methods, such as the Surveyor assay using the gel electrophoresis mutation detection kit, rely on enzymes that recognize and cleave mismatched base pairs after heteroduplexing cut DNA with the wild type DNA. These methods tend to require a large amount of DNA and extensive optimizations as well as lack sensitivity. In order to accurately, quickly, and cost effectively determine the cutting efficiencies of the sgRNAs, we took the advantage of the droplet digital PCR (ddPCR) QX200 system (BioRad).

ddPCR has been used for determining copy number variants (CNVs), mutation frequency, and gene expression levels. In this study, we created a unique drop-off assay to calculate the frequency of cutting events of sgRNAs. The drop-off assay is designed to amplify the region of interest and detect activity at the cut site (target probe) compared to a region in close proximity to the cut site (reference probe). If the template is not cut, both probes are able to bind indicating the presence of wild type DNA. In contrast, if cutting occurs, the probe over the cut site will not bind due to the mismatches incurred during the non-homologous end joining (NHEJ) DNA repair pathway, and signal is generated from the reference probe only. By taking the number of target droplets and dividing it by the total number of droplets (target and reference) we are able to calculate the cutting efficiency of the sgRNAs.

In this study, we successfully used this highly sensitive assay to detect NHEJ presence and quantify cutting efficiency of the sgRNAs. We used DNA inputs as low as 10ng per target and were able to detect sgRNAs that had very low cutting efficiencies, such as 0.5%, with confidence. We also tested different sgRNA and Cas9 inputs and were able to detect a range in cutting efficiency from 11.7% to 75.2%. The results from the ddPCR drop-off assays are highly correlated with results from the Surveyor assay, the current gold standard for measuring the sgRNA cutting efficiency. This ddPCR-based assay offers many benefits compared to the Surveyor assay such as short turn-around time, minimal optimization, and easy analysis that provides quantitative results.

A- A+

PgmNr 205: CRISPR-capture: A novel, low-cost, and scalable method for targeted sequencing.

Authors:
T.L. Mighell [1,2]; C.A. Thornton [2]; B.L. O'Connell [2]; R.M. Mulqueen [2]; C.V. Miller [3]; A.C. Adey [2]; D. Doherty [3]; B.J. O'Roak [2]

View Session Add to Schedule

Affiliations:
1) Neuroscience Graduate Program, Oregon Health & Science Univ, Portland, Oregon.; 2) Department of Molecular & Medical Genetics, Oregon Health & Science Univ, Portland, Oregon.; 3) Department of Pediatrics, University of Washington, Seattle, Washington.

Despite reductions in genome sequencing costs, targeted sequencing methods still have high utility for research and clinical applications; e.g., screening for off-target genome editing or identifying pathogenic mutations in Mendelian disorders. Current target enrichment strategies, whether PCR or hybridization based, still suffer from issues with scalability, bias, and cost (especially for custom targets). Moreover, sequencing a complete gene body remains challenging as non-exonic regions are not typically part of commercial gene panels. To address these challenges, we have developed a novel, low-cost, and scalable method that enables efficient and uniform capture of any set of genomic loci.

Our approach, CRISPR-Capture, is built around the CRISPR-Cas12a system, in which a guide RNA (gRNA) directs the Cas12a endonuclease to a target. Critically, Cas12a double stranded cleavage occurs in a staggered fashion, leaving 4-5 nucleotide overhangs. We reasoned that *in vitro* incubation of genomic DNA with a pool of gRNAs and Cas12a would result in enrichment of “sticky” ligatable ends at programmed sites. Ligation of a biotinylated sequencing adapter is followed by traditional library preparation methods. To further reduce cost, our method supports use of gRNAs created by *in vitro* transcription of low-cost DNA templates. Optimizations led to a single-day capture protocol requiring only 100 ng input DNA, and standard molecular biology equipment and reagents.

To validate this method, we designed a pilot gRNA set targeting 47 known and candidate genes (~3.5Mb) associated with Joubert Syndrome (JS), a genetically heterogeneous, recessive disorder. We tiled 7,176 gRNAs at 500bp intervals, without any design criteria except the presence of a protospacer adjacent motif (PAM). gRNA performance ranged over 1,000-fold and 30% of reads aligned to gRNA specified loci. We next identified critical sequence features governing performance, including GC imbalance and PAM sequence. A linear regression model strongly predicts gRNA performance (r= 0.75). We are currently applying an optimized gRNA set to a large JS cohort (n=578), where ~30% are genetically unsolved, i.e. they lack 2 coding rare variants in the same known gene. With these data, we aim to identify or rule out non-coding mutations in known genes and prioritize families for novel gene discovery. Overall, our CRISPR-Capture platform provides a low-cost and simple workflow for any highly multiplexed sequencing application.

A- A+

PgmNr 206: Interrogating regulatory consequences of genetic variation in DNA associated proteins.

Authors:
C. Wu [1, 2]; S. Shleizer-Burko [2]; A. Goren [2]; M. Gymrek [2, 3]

View Session Add to Schedule

Affiliations:
1) Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA.; 2) Department of Medicine, University of California San Diego, La Jolla, CA; 3) Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA

Understanding the mechanistic impact of a specific mutation is a key challenge in human genomics. Mutations in proteins including transcription factors, chromatin regulators or splicing factors may cause widespread transcriptomic changes. Intriguingly, different mutations in the same gene can cause distinct phenotypic changes, ranging from no impact to severe health consequences. For example, mutations in FOXC1 may lead either to Anterior Segment Dysgenesis with the Rieger subtype or the Axenfield subtype depending on where they fall within the protein. A potential explanation of these diverse phenotypes is that different mutations may act by a variety of mechanisms, including loss/gain of function or reduced/enhanced activity. For example, mutations in the binding domain of a protein may have profound impact on binding affinity, whereas mutations elsewhere in the protein may have little impact. Genome editing allows the study of specific mutations. However, generating cell lines with precise edits remains an inefficient process making it infeasible using standard methods to study more than a handful of mutations. Here we develop a multiplexed genome-editing assay to measure the regulatory effects of dozens of genetic variants in a particular DNA-associated protein simultaneously using single-cell RNAseq (scRNAseq). The pipeline consists of (1) base editing to efficiently introduce multiple edits to the protein of interest in a pool of cells and (2) scRNAseq on the pool of edited cells. scRNAseq data is used to determine which cells received which edit and to identify differentially expressed genes induced by each target mutation. We also develop a simulation framework to determine experimental parameters required to obtain robust results across a range of mutation types and scRNAseq platforms. We tested the feasibility of our approach by introducing a library of 6-8 sgRNAs for three genes harboring known pathogenic mutations (GATA4, EP300, and FOXC1) and perform scRNAseq on the edited pool to identify genome-wide transcriptomic changes. We achieved editing efficiencies of up to 53%, compared to less than 5% using homology-directed repair. We are additionally generating clonal cell lines with known pathogenic mutations as validation data for the full pipeline. Once established, our approach will provide a valuable platform for simultaneously characterizing the transcriptomic impact of dozens to hundreds of pathogenic variants in protein-coding genes.

A- A+

PgmNr 207: Deciphering the histone acetyl code in human cells with dCas9-based epigenome editing.

Authors:
K. Wang [1]; J. Li [1]; I.B. Hilton [1,2]

View Session Add to Schedule

Affiliations:
1) Department of Bioengineering, Rice University, BioScience Research Collaborative, 6500 Main Street, Suite 1030, Houston; 2) Department of Bioscience, Rice University, W100 George R. Brown Hall 6100 Main Street, Houston

---

The acetylation of nucleosomal histones has been positively correlated with eukaryotic gene expression for over 50 years. We previously used dCas9-based epigenome editing to show that the targeted acetylation of endogenous chromatin at human enhancers or promoters results in the activation of gene expression. This work helped to functionally connect the acetylation of human histones to cellular transcription. Although our prior work focused on the programmable acetylation of H3K27, the acetylation of numerous other lysine residues on human histones has also been correlated with active human gene expression. Furthermore, ChIP-seq results show that multiple histone lysine residues are often acetylated contemporaneously on endogenous nucleosomes. Despite these important advances, it remains unclear whether a single acetylated lysine at genomic regulatory elements (e.g. H3K27) is sufficient to cause gene activation, or instead, whether a combination of acetylated histone lysine residues is required to do so. To overcome this mechanistic limitation, we have built a suite of new dCas9-based histone lysine acetyltransferases. These new tools allow us to acetylate specific lysine residues at targeted loci within native human chromatin. Our results show that although targeted histone acetylation generally activates human genes; the acetylation of different histone lysine residues produces different gene-regulatory effects. Moreover, we show that the combinatorial acetylation of human histone lysine residues leads to synergy in targeted human gene regulation. Our findings lay the groundwork for defining a long-sought histone lysine acetyl code regulating human genes.

A- A+

PgmNr 208: Incompletely penetrant rare coding variants account for a major fraction of patients with undiagnosed developmental disorders.

Authors:
K.E. Samocha; P. Danecek; E.J. Gardner; H.C. Martin; P.J. Short; M.E. Hurles; on behalf of the Deciphering Developmental Disorders study

View Session Add to Schedule

Affiliation: Human Genetics, Wellcome Sanger Institute, Hinxton, United Kingdom

While dozens of dominant developmental disorders have been identified over the last decade, we estimate that we have only identified genes that explain approximately half of the cases with dominant-acting *de novo* variants (Martin et al Science 2018). A major contributor to our inability to identify the remaining genes is incomplete penetrance. There are well-known examples of incompletely penetrant copy number variants, such as 16p12.1 and *NRXN1*, but the contribution of incompletely penetrant sequence variants has been largely unexplored.

To pursue the contribution of incompletely penetrant rare coding variants to developmental disorders, we studied exome sequence data from 9,858 parent-child trios collected as part of the Deciphering Developmental Disorders (DDD) study. Additionally, we jointly analyzed the DDD data with the exome sequences of ancestry-matched control individuals from the INTERVAL and UK10K projects (total n=37,898 exomes). We observed two independent lines of evidence that incompletely penetrant variants, both *de novo* and inherited, account for a substantial fraction of the undiagnosed patients in the DDD cohort. First, modeling the residual excess of *de novo* coding variants after removing known disorders indicates that the remaining burden of *de novo* variants is distributed among genes with lower penetrance than the known disorders. Second, we observe a significant excess of inherited rare coding variants ($p<5x10^{-4}$ for protein-truncating variants) across all trios. This burden remains when removing trios with a diagnostic *de novo* variant and those with parents who have similar clinical features ($p=1.1x10^{-5}$). The excess burden of inherited variation is particularly concentrated in genes previously associated with developmental disorders (effect size = 1.39, $p=2.9x10^{-5}$) and genes intolerant of protein-truncating variation (pLI≥0.9; effect size = 1.19, $p=1.0x10^{-7}$). Additionally, we find that the genes with an excess of *de novo* variants also have a significant excess of inherited protein-truncating variants (effect size = 1.65, $p=9.7x10^{-6}$).

Finally, we will use the genotyping information from ~5,000 trios to investigate if the unaffected parents who are transmitting potentially pathogenic variants are protected either via cis-regulatory variation or polygenic scores compared to their affected children. These data will allow us to further explore the interaction between common and rare inherited variation in developmental disorders.

A- A+

PgmNr 209: Dysfunction in neurons and glia reveals that distinct *PIGA* deficiency phenotypes arise from independent cell types.

Authors:
C.Y. Chow [1]; M. Haller [1]; E. Coehlo [1]; J. Plenis [2]; O. Kanca [3]; H. Bellen [3,4]; A. Rodan [2]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics, Univ of Utah, Salt Lake City, Utah.; 2) Department of Internal Medicine, University of Utah, Salt Lake City, UT; 3) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 4) HHMI

Mutations in the *Phosphatidylinositol glycan class A* (*PIGA*) gene cause *PIGA* deficiency, a type of X-linked epilepsy and intellectual developmental disorder. *PIGA* deficiency is an ultra-rare disease with fewer than 12 patients reported. *PIGA* deficiency is characterized by neonatal hypotonia, myoclonic seizures, epilepsy, dysmorphic features, and a number of congenital anomalies. PIGA is involved in the first step of glycosylphosphatidylinositol (GPI) anchor biosynthesis by transferring GlcNAc from UDP-GlcNAc to PI to form GlcNAc-PI. GPI attaches the C-terminus of a protein to the cell surface. GPI-anchored proteins play a number of roles in cell signaling, migration, and immunity. It remains unclear how loss of PIGA function contributes to the phenotypes observed in patients. Because a number of the patient phenotypes include nervous system abnormalities, we used RNAi technology to knockdown *Drosophila PIGA* (*PIG-A*) expression in neurons and in glia. Neuron-specific loss of *PIGA* function results in behavioral and neurological abnormalities, including locomotion defects and sleep disturbances, that are reminiscent of those observed in patients. Because epilepsy is present in all *PIGA* deficiency patients, it is surprising that neuronal knockdown does not result in a seizure-like phenotype. Strikingly, glia-specific knockdown does result in a seizure-like phenotype, but no movement disorder. To understand the molecular underpinnings of these cell type-specific phenotypes, RNAseq analysis was performed on heads from flies with neuronal or glia-specific knockdown of *PIGA* and controls. Transcriptome analysis in neuron vs glia knockdown heads reveals likely mechanisms as to why seizures are observed in glia knockdowns and not in neuronal knockdowns. Finally, we generated 3 patient-specific models of *PIGA* deficiency, each of which carries a different disease allele. These models demonstrate how disease alleles differentially affect neurons and glial and provides insight into the pathogenesis of each *PIGA* disease allele. These models provide a path forward for precision medicine approaches in *PIGA* deficiency, including patient-specific small molecule screens. This study suggests that treatment of the epilepsy and seizure phenotypes observed in *PIGA* deficiency patients will 1) require therapies that target primary and secondary dysfunction in glia and neurons and 2) require precision medicine approaches to treat patients carrying different disease causing alleles.

A- A+

PgmNr 210: The ZBTB7B-RSK3 pathway regulates ATXN1, the protein driving neurodegeneration in an inherited ataxia.

Authors:
W. Lee [1,2,3]; L. Lavery [2,3]; M. Rousseaux [2,3]; I. Al-Ramahi [2,3]; Y. Wan [3,4]; W. Kim [3,5]; C. Adamski [2,3]; V. Bondar [2,3]; H. Orr [6]; Z. Liu [3,4]; J. Botas [2,3]; H. Zoghbi [2,3,4,5]

View Session Add to Schedule

Affiliations:
1) Integrative Molecular and Biomedical Science Program, Baylor College of Medicine, Houston, TX; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 3) Jan and Dan Duncan Neurological Research Institute, Houston, TX; 4) Department of Pediatrics-Neurology, Baylor College of Medicine, Houston, TX; 5) Howard Hughes Medical Institute, Chevy Chase, MD; 6) Institute for Translational Neuroscience, University of Minnesota, Minneapolis, MN

Spinocerebellar ataxia type 1 (SCA1) is an autosomal dominant neurodegenerative disease caused by expansion of polyglutamine-encoding CAG repeats in ATXN1. We found that the expanded polyglutamine tract causes ATXN1 protein accumulation, which induces neuronal toxicity. Reducing Atxn1 by 20% rescues SCA1 phenotypes in mice, however the molecular mechanisms that regulate ATXN1 levels are not well understood. To discover such regulators, we performed a genome-wide shRNA screen of 7,787 potentially druggable targets in human cells and identified 21 novel ATXN1 regulators. Of those, we further validated two closely related BTB-ZF transcription factors, ZBTB7A and ZBTB7B, that positively regulate Atxn1 levels *in vivo*. Of the two, ZBTB7B is a more effective ATXN1 regulator due to its larger effect size. Because ZBTB7B regulates ATXN1 protein rather than RNA levels, and requires the zinc-finger DNA-binding domain for this regulation, we rationalized that the effect is indirect and performed RNA-seq on cells overexpressing ZBTB7B. We found that ZBTB7B regulates expression levels of RSK3, which in turn regulates ATXN1 levels. Knockout of RSK3 nullifies the increase of ATXN1 induced by ZBTB7B overexpression *in vitro*, and heterozygous knockout of Rsk3 decreases Atxn1 *in vivo*. ATXN1 phosphorylation at S776 by MSK1 has previously been shown to stabilize ATXN1 protein. Similarly, we found that RSK3 is able to phosphorylate S776. Combinatorial treatment of RSK3 and MSK1 inhibitors decreases total and phospho-S776 ATXN1 levels to a greater extent than individual application does in SCA1-patient derived neurons. Understanding how the ZBTB7B-RSK3 pathway regulates ATXN1 will provide a framework for future studies aimed at the development of therapeutics to maintain neuronal health in SCA1 patients.

A- A+

PgmNr 211: Rare *CDC20B* variants in juvenile myoclonic epilepsy.

Authors:
P. Barak [1]; P. Satishchandra [4]; S. Sinha [4]; G. Kuruttukulam [3]; M. Kaur [1]; A. Anand [1,2]

View Session Add to Schedule

Affiliations:
1) Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bengaluru 560064, India; 2) Neuroscience Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bengaluru 560064, India; 3) Department of Neurology, Lourdes Hospital, Cochin 682012, India; 4) Department of Neurology, National Institute of Mental Health and Neurosciences, Bengaluru 560064, India

---

Juvenile myoclonic epilepsy (JME) is a prevalent genetic generalized epilepsy characterized by frequent myoclonic jerks which are accompanied by generalized tonic-clonic seizures and absence seizures. Linkage studies have identified about 30 sub-genomic locations which may harbor genes underlying JME. Here, we present, whole genome-wide linkage analysis of a multi-affected JME family which helped identify a previously unknown locus at 5p15-q12. This locus spans 64 mega-bases of the genome and harbors 177 protein coding genes. Whole exome sequencing for two affected members of the family was conducted. This led to the identification of a non-synonymous variant in the gene, *CDC20B* (Cell Division Cycle 20B). In addition, four rare, missense *CDC20B* variants present almost exclusively among a cohort of over 500 JME patients were found. *CDC20B* encompasses 60 kilobase of the genome and encodes a 591 amino acid protein. It has 7 WD repeat domains at its C-terminus. *CDC20B* is paralogous to the *CDC20* gene, which interacts with the anaphase promoting complex and has a major role in cell cycle progression. Cellular and molecular roles of CDC20B are beginning to be examined. Cellular expression and localization of CDC20B protein are being studied in cultured mammalian cells. While during late telophase and cytokinesis, the protein localizes to the midbody, during other cell cycle stages, it is present in the cytoplasm. CDC20B co-immunoprecipitates with gamma tubulin, a protein present abundantly at the midbody during cytokinesis. Over-expression of the CDC20B variants in HeLa cells led to the accumulation of cells at the cytokinesis stage, suggesting cell cycle associated role for the protein.

A- A+

PgmNr 212: Deep tensor factorization characterizes the human epigenome through imputation of thousands of genome-wide epigenomics and transcriptomics experiments.

Authors:
J. Schreiber [1]; J. Bilmes [1,2]; W. Noble [1,3]

View Session Add to Schedule

Affiliations:
1) Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WASHINGTON.; 2) Department of Computer and Electrical Engineering, University of Washington, Seattle, WASHINGTON; 3) Department of Genome Science, University of Washington, Seattle, WASHINGTON

Introduction: The human epigenome has been experimentally characterized by thousands of uniformly processed epigenomic and transcriptomic data sets. These datasets characterize a rich variety of biological activity, such as protein binding, chromatin accessibility, methylation, transcription, and histone modification, in hundreds of human cell lines and tissues (``biosamples''). However, due primarily to cost, the total number of assays that can be performed is limited to a small fraction of potential experiments.

Methods: To address this challenge, we propose a deep neural network tensor factorization method, Avocado, that compresses epigenomic data into a dense, information-rich representation of the human genome. The resulting model can be used to impute the thousands of epigenomic experiments that have not yet been performed, and the learned latent representations are broadly useful for analysis of human epigenomics.

Results: We begin by using 1,014 tracks of chromatin accessibility and histone modification from the Roadmap Epigenomics Consortium to demonstrate that this learned representation of the genome is broadly useful: first, by imputing epigenomic data more accurately than previous methods, and second, by showing that machine learning models that exploit this representation outperform those trained directly on epigenomic data on several prediction tasks.

Next, we applied Avocado to a dataset of 3,814 tracks of data derived from the ENCODE compendium. The resulting imputations cover measurements of chromatin accessibility, histone modification, transcription, and protein binding. To our knowledge, this is the first imputation that jointly models this number of biological phenomena. We comprehensively evaluate these imputations and show significant improvements in protein binding performance compared to the top models in a recent ENCODE-DREAM challenge.

Discussion: Avocado also shows promise in less canonical imputation settings. For example, initial results have shown that a model trained on human epigenomics can be transferred over to other species, allowing for imputations of activity where no experimental data has been acquired yet for that species. Further, when applied to non-overlapping data sets of single-cell measurements, Avocado can operate as an in-silico co-assay, leveraging bulk data to enable the imputation of several forms of biological activity within a single cell.

A- A+

PgmNr 213: Using a deep neural network (DNN) based classifier to predict primary sites of cancers of unknown primary.

Authors:
I. Moon [1,2]; A. Gusev [2,3]

View Session Add to Schedule

Affiliations:
1) Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA; 2) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA; 3) Division of Genetics, Brigham & Women’s Hospital, Boston, MA

When a standardized diagnostic test fails to locate primary site of a metastasis cancer, it is diagnosed as a cancer of unknown primary (CUP). This type of cancer represents about 2 % of all cancers according to the American Cancer Society and presents oncologists with an added difficulty in treating and prognosing the disease. Therefore, there has been a high demand for an accurate method for identifying the primary site of CUP to empower site-specific treatment. While previous methods have focused on classification from whole-genome sequencing or epigenomic data, only tumor panel sequencing is now routinely part of the standard of care and thus has the most translational potential.

To that end, we propose a deep neural network (DNN) based classifier to predict primary sites of CUPs using the tumor panel sequencing data from Dana-Farber Cancer Institute Profile project (N>20,000 tumors) and the AACR Project GENIE (N > 60,000 tumors). We processed the gene mutation and copy number alteration data to create trainable features for each tumor sample, including single variant substitutions, substitution tri-nucleotide context, and somatic copy number alterations. Our DNN classifier has been trained on the processed feature data to capture complex non-linear relationships between the processed features and cancer types. Using 3-fold cross validation, we have validated the proposed classifier achieves overall accuracy of 86.6 % and weighted F1 score of 0.865 on classifying 7 common cancer types for test tumor samples drawn from the DFCI Profile project data.

We are now applying the proposed classifier to a clinical cohort of 824 CUP patient tumors with available somatic, germline, and detailed treatment data to predict primary origin of these tumor samples. Uniquely, we have validated our classifier performance by quantifying germline polygenic risk scores (computed from GWAS of breast, prostate, ovarian, lung, melanoma, and kidney cancers) and seen significant enrichment of PRS for the tumor types classified by the proposed model. Furthermore, we will evaluate overall survival for patients with computationally classified tumors who received matching site-specific therapies compared to patients with empirical or non-matching treatment. Finally, we will further boost performance of the current classifier by incorporating data from diverse tumor panels and exploring other machine learning algorithms and optimizers.

A- A+

PgmNr 214: Using deep learning and genetic big data to predict complex disease risk.

Authors:
D. Li [1]; T. Haritunians [1]; M. Daly [2]; D. McGovern [1]; IIBDGC Consortium

View Session Add to Schedule

Affiliations:
1) Medicine, Cedars-Sinai Medical Center, Los Angeles, California.; 2) Broad Institute

Background: Genetics contributes significantly in many complex diseases. Early identification of patients at high risk could be key in personalized early intervention and prevention.
Methods: We utilized DeepLearning(DL) to build a risk prediction model with genetic data. The model was built using with Convolutional Neural Network(CNN) with L1 and L2 penalization. Two different blocks, one with known variants and the other one with the rest of the genome, were constructed separately and then combined into the output layer. Variable relative importance was calculated using Gedeon's approach based on the weights connecting the input features to the first two hidden layers.

This approach was applied in 13,523 Crohn's Disease (CD) patients and 33,902 controls from IIBDGC as training dataset. Model validation was performed in 2,843 CD cases recruited from a single center and 4568 controls. Both training and validation cohorts were genotyped using ImmunoChip. Performance of DL model was compared to the previously published LDpred algorithm(Amit V. Khera et al, Nat Genet, 2018).
Results: We observed an AUC of 0.841 for DL, significantly better than LDpred (AUC= 0.699, P=5.75E-83). People at the extreme of the predicted DL score have a much higher risk of CD compared to the rest of the sample, with an OR of 26.32 in top 5% and 19.25 in top 10%. Within the top 5% and 10% of the DL score, 93.5% and 90.3% are CD patients, respectively. As a comparison, OR is 10.33 in top 5% and 4.35 in top 10% for LDpred, and the corresponding proportion of CD patients are 78.3% and 69.9%, respectively.

Further analyses indicate that the improved performance of DL is likely through its ability to incorporate non-linear causal effects, including high-order interaction such as interaction between *NOD2* and *CD28*, and deviation from linear additive models in recessive mutations such as *NOD2* frameshift mutation Leu1007fsinsC. Compared to LDPred, individuals carrying those mutations have much higher predicted risk in DL model. Variable importance metrics highlighted 11 novel regions with P<5.0E-8 in a subsequent meta-analysis.
Conclusion: With this novel algorithm we can identify individuals with monogenic-like disease risk only using genetic data, making it a powerful tool for precise and personalized intervention and prevention. Our research also indicates that non-linear genetics effects might contribute significantly to phenotypic variance of complex diseases such as CD.

A- A+

PgmNr 215: A deep learning method for inferring the strength and location of sweeps using local genealogies based on the ancestral recombination graph.

Authors:
H.A. Hejase; A. Siepel

View Session Add to Schedule

Affiliation: The Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

A wide variety of methods that analyze modern DNA sequences have been developed to detect signals of selective sweeps. The hitchhiking effect provides a key signature to detect changes in the spatial pattern of haplotype structure and site frequency spectrum, allowing one to make inferences on whether selection has acted on an allele. Various approaches have been developed to infer selective sweeps including Approximate Bayesian Computation and supervised machine learning. These approaches make use of traditional summary statistics or a combination thereof to infer the strength, location, and timing of selection. Here, we introduce a recurrent neural network (RNN) to infer the strength and location of sweeps. Rather than learning on traditional summary statistics calculated from a multiple sequence alignment, we use the local genealogies sampled from the ancestral recombination graph(ARG) as input to the RNN. While these local genealogies would be extremely high-dimensional, the recent rise of recurrent neural networks for speech recognition suggests that encoding gene trees as “words” and loci (e.g. collection of contiguous gene trees) as “sentences” enable powerful population genetic inferences. Our RNN uses distinguishing local topological features of the ARG as inferred by ARGweaver (e.g. outlier clusters of coalescent events, coalescent time to most recent common ancestry, and number of lineages remaining at discrete time points of the genealogies). It learns patterns of variation in the feature set to infer the location and selection coefficient of a sweep. Using simulations based on the CEU demographic model, our method estimated selection coefficients with a Pearson correlation as high as 90% compared to the ground truth. Additionally, we predicted the location of sweeps with an accuracy as high as 90% across different simulation settings. When applied to local genealogies inferred from the CEU population in the 1000 Genomes dataset, our model detected sweep signals and inferred selection coefficients on the LCT locus ($s = 0.0105$), which affects the lactase persistence trait, pigmentation-related genes (e.g. ASIP with $s = 0.002$), and genes related to fertility and reproduction (e.g. SPAG4 and TEKT2). Our method provides new insight into how selection shaped regions of the human genome, and provides a model that makes use of the full dataset in the form of the ARG rather than using traditional summary statistics.

A- A+

PgmNr 216: Population differences in brain regulatory profiles influence risk of schizophrenia.

Authors:
S. Liu [1]; Y. Chen [1]; F. Wang [1]; Y. Jiang [1,2]; F. Duan [1]; R. Kopp [3]; C. Liu [1,3]; C. Chen [1,4]; PsychENCODE Consortium

View Session Add to Schedule

Affiliations:
1) School of Life Sciences, Central South University, Changsha, China.; 2) Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA.; 3) Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA.; 4) National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South.

Genome-wide association studies (GWAS) have identified different risk variants for schizophrenia (SCZ) in populations of East Asian and European ancestry. Expression quantitative trait loci (eQTLs) map the genomic loci that regulate gene expression; used in conjunction with GWAS, eQTLs can explain GWAS signals and identify risk genes. However, European ancestry data comprises the majority of available postmortem brain data, impeding interpretation of non-Caucasian GWAS signals. Therefore, developing brain regulatory profiles to compare East Asian and Caucasian populations will facilitate a better understanding of racial differences found in SCZ GWAS signals. With data from the China Human Brain Banking Consortium and PsychENCODE Consortium, we analyzed 151 brain genotypes/gene expression data associated with East Asian ancestry and 407 brain genotypes/gene expression data associated with European ancestry. We performed eQTL mapping separately per cohort. Based on the eQTL results, we defined *ancestry-shared* eQTL (common to both populations with the same direction) and *ancestry-specific* eQTL (only one population). The *ancestry-shared* eQTL ratio was low (0.39), showing that eQTLs differ substantially across populations. To explore whether differences in allele frequency between races contribute to *ancestry-specific* eQTL, we compared the distribution of Fst (fixation index, measuring population differences due to genetic structure) value for *ancestry-specific* eQTLs. We found that *ancestry-specific* eQTLs were significantly enriched for population-divergent single nucleotide polymorphisms (SNPs) (Fst >0.25) in population comparisons (P <2.2e-16). This suggests that differences in allele frequency are the driving force behind *ancestry-specific* eQTLs. To identify specific risk genes, we integrated PGC2 and East Asian SCZ GWAS data with eQTL results. Using Summary-data-based Mendelian Randomization, we identified 91 SCZ risk genes in the European ancestry cohort and 62 SCZ risk genes in the East Asian cohort. Only three of the genes were shared. Most of the causal SNPs detected were *ancestry-specific* SNPs. This result indicates that risk genes are involved in *ancestry-related* SNPs. Collectively, our study provides evidence that brain regulatory profiles and SCZ-risk genes are ancestry-specific, due primarily to differences in allele frequency. These results indicate the importance of studying both disease GWAS and brain eQTL in diverse populations.

A- A+

PgmNr 217: Sex-differences in brain development correspond to sex-differences in the age of onset for schizophrenia.

Authors:
C. Jiao [1]; R. Kopp [1]; L. Vivid [1]; C. Chen [2]; C.Y. Liu [1,2]

View Session Add to Schedule

Affiliations:
1) SUNY Upstate Medical University, Syracuse, NEW YORK.; 2) School of Life Science, Central South University, Changsha, Hunan, 410012, China

Objective: Schizophrenia (SCZ) and autism spectrum disorder (ASD) are brain development disorders characterized by earlier onset in males than in females. Their complexity is contributed by several factors, such as synapse, energy metabolism, and the immune system. Each of these factors also relates to neurodevelopmental differences between males and females. Therefore, we hypothesized that sex differences in the age of onset of psychiatric disorders are associated with sex-differences at various neurodevelopmental stages. To test this, we analyzed temporal gene expression profiles in postmortem human brain samples connected with SCZ and ASD related loci.

Materials and Methods: We collected 1,357 samples from 497 postmortem human brains (BrainSpan, Brain Cloud, and HDBR), which spanned in age from 4 post-conceptual weeks to 78 years. We identified seven developmental stages distinguishable by the biological developmental characteristics and gene expression distributions. We detected genes that were highly expressed in specific stages (specific highly expressed [SHE] genes). We also identified “turning point genes,” which showed a significantly different level of expression in neighboring. Enrichment analysis tested whether the above genes were linked with known psychiatric risk loci.

Results: We found that 51.4% of the 11,797 SHE genes (FDR<0.05 and |log (FC)|>2) showed sexual differences in expression and were associated with mTOR and Wnt signaling, as well as synaptic transmission. Furthermore, we found that 19.1% of the SHE genes were more highly expressed in the earlier stages of males than of females. These genes are also related to energy metabolic and immune function and were enriched with astrocyte (p= 0.0014) and endothelial cell (p=4.64e-05) markers.
For SCZ, GWAS signals in males were enriched in SHE genes in an earlier developmental stage (1st trimester, FDR=0.0021) than females (3rd trimester to 1yrs, FDR=0.0038). However, for ASD, we found no sex differences using our methods with SHE genes in the first trimester enriched in nonsynonymous loci (FDR=0.0208). We also found “turning point genes” from first to the second trimester were enriched in GWAS-ASD signals (FDR=0.0303).

Conclusion: Except for validating the importance of the first trimester in brain development, we also found that genes that are highly expressed earlier in males than in females correspond to identified SCZ risk loci.

A- A+

PgmNr 218: Refining cell populations and fine-mapping variants for schizophrenia and bipolar disorder using mouse open chromatin profiles.

Authors:
P.W. Hook; A.S. McCallion

View Session Add to Schedule

Affiliation: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

INTRODUCTION: Schizophrenia (SZ) and bipolar disorder (BD) are common, debilitating, neuropsychiatric disorders. The genetics of SZ and BD have been extensively explored through the use of genome-wide association studies (GWAS). Progress in understanding the molecular mechanisms underpinning observed signals is impeded because causal variants and disease-relevant cell populations, in which those variants act, remain unknown. We address this challenge by performing heritability enrichment analyses and fine-mapping using open chromatin data from purified *ex-vivo* mouse populations.

METHODS: We uniformly processed public ATAC-seq data from 25 mouse cell populations including cortical excitatory and inhibitory neurons, dopaminergic neurons, glia, T-cells, and retina cells. After lifting over open chromatin data to syntenic human sequences, we tested profiles for enrichment of disease heritability using stratified linkage disequilibrium score regression. We then used open chromatin profiles from enriched populations to fine-map 53,785 variants using PAINTOR across 177 SZ loci and screened variants for their predicted impact on transcription factor (TF) binding sites.

RESULTS: We find purified, layer V excitatory neurons show highest enrichment for SZ heritability among discrete populations. Excitatory neurons in all other cortical layers and the dentate gyrus, as well as inhibitory neurons and astrocytes also exhibit SZ heritability enrichment. We observe similar patterns of enrichment for BD. Further, when analyzing a GWAS that compared SZ with BD, we demonstrate enrichment is restricted to excitatory neurons in cortical layers II-III, IV, and V, as well as in the dentate gyrus. Fine-mapping reveals 172/177 SZ loci contain ≥ 1 variant with a posterior inclusion probability (PIP) ≥ 0.1. Of these, 424 variants across 125 loci reside in open chromatin in at least one enriched cell population. One example is rs181813160 which is located in the promoter of the *NGEF* gene; it is highlighted by a PIP of 0.97. It is encompassed by open chromatin in 12/13 enriched cell populations and the risk allele disrupts a binding site for *EGR1*, a TF important to neuronal biology.

CONCLUSION: Leveraging the power and precision of mouse-derived chromatin data, we systematically indict cell populations contributing to SZ and BD. We prioritize and give biological context to variants in 70% (125/177) of SZ loci studied, establishing the settings in which they should be functionally tested.

A- A+

PgmNr 219: From genetic risk variants to convergent protein networks: An integrative approach to elucidate the causal molecular mechanisms of schizophrenia.

Authors:
A. Kim [1,2]; E. Nacu [1,3]; E. Malolepsza [1,2]; N. Petrossian [1]; W. Crotty [1]; T. Suh [1]; J. Riseman [1]; T. Singh [1,2]; R. Liu [1,2]; S. Muller [1,2]; G. Pintacuda [1,3]; M. Johnson [1,4]; K. Sharma [5]; R. Huganir [5]; B. Stevens [1,4]; M. Daly [1,2]; H. Huang [1,2]; M. Schenone [6]; S. Egri [6]; B. Tanenbaum [6]; A. Appfel [6]; C. Stanclift [6]; J. Jaffe [6]; K. Lilliehook [1]; K. Eggan [1,3]; K. Lage [1,2]; SCHEMA Consortium

View Session Add to Schedule

Affiliations:
1) Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA; 2) Department of Surgery, Massachusetts General Hospital, Boston, MA.; 3) Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA; 4) Department of Neurology, Boston Children's Hospital, Boston, MA; 5) Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD; 6) Proteomics Platform, Broad Institute, Cambridge, MA

Recent studies have revealed the highly polygenic nature of schizophrenia, but the specific underlying gene networks (pathways) remain obscure in many cases. This is a key bottleneck towards biological understanding and therapeutic intervention.

Human upper layer cortical excitatory neurons are particularly relevant to schizophrenia. We developed a production framework of homogeneous cell populations at scale (>10 billion neurons per year) and analyzed neuron-specific protein-protein interactions of schizophrenia risk genes using tandem mass spectrometry. In parallel, we developed a computational platform (Genoppi) to QC the data, to integrate proteomic and genetic datasets, and to prioritize emerging pathway models for functional validation experiments.

Many of the high-quality protein interactions we identified are unique to human neurons (i.e., not reported in the literature nor seen in non-brain tissues). By integrating data from the SCHEMA exome sequencing consortium and the East Asia Cohort of the Psychiatric Genomics Consortium, we found that the interaction networks of several schizophrenia risk proteins (e.g., CACNA1C, HCN1, and SYNGAP1) are significantly enriched for common and/or rare risk variants, illustrating how common and rare genetic risk can converge onto the same cellular networks in human neurons. Functional studies of interaction partners identified by our approach show that knockdowns at the gene level result in synaptic phenotypes in primary mouse neurons.

Overall, we describe an integrated experimental and computational framework to 1) map interactomes of schizophrenia proteins in human neurons, 2) model pathway relationships through integration of proteomic and genetic data, and 3) functionally validate key schizophrenia network modules in human neurons, human brain tissue, and animal models. Our approach bridges the evolutionary gap between current model organisms and humans to maximize functional insights from recent genetic data.

A- A+

PgmNr 220: PRINCESS: Framework for comprehensive detection and phasing of SNPs and SVs.

Authors:
M. Mahmoud [1,2]; F.J. Sedlazeck [1]

View Session Add to Schedule

Affiliations:
1) Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas.; 2) Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas.

Short Illumina sequencing is the state of the art for genetics despite the fact that it misses 193 medical relevant genes and other genomic regions (e.g. STR, ALU), which have been associated to diseases. In addition, it cannot provide sufficient phasing information, which is crucial for diseases such as mutations on TPMT to determine the ability of drug metabolism for a patient.

Recently, long-read sequencing technologies such as PacBio and Oxford Nanopore have shown the ability to enhance the detection of genomic variations either being it Single Nucleotide Variants (SNVs), Structural Variants (SVs) or methylation changes. Nevertheless, none of the studies so far detect all genomic variations either focusing on SNVs, SVs or methylation change. Furthermore, only a few studies include phasing information to further ease the prediction of these variations onto the genes and thus phenotypes. Thus, clinical and research studies currently lack a comprehensive view of genomic variations although the information is present in their sequenced data set.

Here we introduce PRINCESS–a method that provides haplotype resolved SNVs, SVs and methylation changes based on a single long read sequencing run from PacBio or Oxford Nanopore. PRINCESS automatically adapts itself to different coverage levels to optimally leverage the data set at hand. Thus, PRINCESS provides cost and time efficient comprehensive insights of haplotype resolved genomic variations. This information can be leveraged to simultaneous study the interaction of the SNVs, SVs and methylation changes and their impact on phenotypic changes.

PRINCESS was evaluated based on Genome in a Bottle (GIAB) Oxford Nanopore standard and ultra-long reads as well as PacBio Continuous Long Reads (CLR) and Circular Consensus Sequencing (CCS) data. On 1 SMRT Cell CLR data, PRINCESS achieved 95% precision and 80% sensitivity for SNVs and 93% precision and 77% sensitivity for SVs on CLR reads reaching 6.9 Mbp N50 phasing of SNVs and SVs. For 1 SMRT Cell CCS data 95% precision and 90% sensitivity for SNVs and 94% precision and 79% sensitivity for SVs reaching a N50 of 225kbp.

PRINCESS applied to 18 PacBio with matching RNA-Seq data samples improved the detection of SVs (on average 22,105), SNVs and phasing (~5 Mbp average N50) and thus allowed the detection of eQTL in an automated, fast and comprehensive fashion.

A- A+

PgmNr 221: A robust and production-level approach to haplotype-resolved assembly of single individuals.

Authors:
S. Garg [1]; C. Fungtammasan [5]; A. Carroll [6]; R. Hall [4]; E. Hatas [4]; M. Mahmoud [2]; F. Sedlazeck [2]; M. Chou [1]; J. Aach [1]; J. Zook [3]; J. Chin [5]; G. Church [1]

View Session Add to Schedule

Affiliations:
1) Harvard Medical School, Boston, MA.; 2) Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston TX 77030; 3) Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899; 4) Pacific Biosciences, Menlo Park, California; 5) DNAnexus, Mountain View, California; 6) Google Genomics, Mountain View, California

---

Reconstructing the complete and phased sequence of every chromosome copy in a human individual is a high priority goal for medical and population genetics. Most current approaches collapse both phased information into a single assembly, discarding phase information. Although efforts have been made to reconstruct these phased sequences, they either require >200 CPU hours or fail to assemble continuous haplotype sequences. There is a pressing need for a streamlined, production-level approach that can reconstruct high-quality phased sequences, and that can be applied to hundreds of human genomes.

Here, we propose an integrative *de novo* assembly and phasing strategy that leverages new forms of long-read and long-range connectivity data in a computationally efficient manner. Specifically, our approach combines complementary high throughput sequencing and connectivity datasets such as PacBio CCS and Hi-C, constructs a preliminary high-quality haploid consensus, and then conducts an optimized partitioning of reads and complete separate assembly of each homolog within a single integrated algorithm. Our approach produces high-quality diploid assemblies (excluding centromeres), is highly scalable, and can be integrated to the cloud platform for production-level assemblies of multiple single genomes. An additional advantage is that it excludes any reference sequence bias that could interfere with discovery of sequences unique to particular individuals or populations and allows for the detection of novel structural variants.

We demonstrate the feasibility of our approach on three genomes from the Personal Genome Project (PGP-1), the Genome in a Bottle project (HG002) and the 1000 Genome Project (NA12878), produce highly continuous haplotype-resolved assemblies with N50 of 15.4 Mb, and show that we require as little as 20x coverage of PacBio CCS and 30x of Hi-C to generate high-quality assemblies. We also discover novel phased sequences not included in GRCh38 and private to each genome. We validate these novel phased sequences against BAC or trio data.

In summary, our novel computational approach efficiently and robustly combines data from new sequencing and genome connectivity mapping technologies to produce high quality diploid assemblies that will support community research goals of producing accurate end-to-end finished human genomes of individuals, and so lead to improvements in personalized medicine and increased understanding of human genome sequence diversity.

A- A+

PgmNr 222: Interrogating and correcting fine-scale genetic structure in large (>36,000 samples) GWAS datasets using scalable haplotype sharing methods.

Authors:
R.P. Byrne [1]; W. van Rheenen [2]; J.H. Veldink [2]; R.L. McLaughlin [1]

View Session Add to Schedule

Affiliations:
1) Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Leinster, Ireland; 2) Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht.

We leveraged a powerful and scalable haplotype painting algorithm, Positional Burrows Wheeler Transform Painting (pbwtPaint), to explore the co-ancestry of 36,052 individuals of European descent from a recent amyotrophic lateral sclerosis (ALS) genome-wide association study (GWAS). The resulting haplotype sharing matrix revealed both striking broadscale genetic structure between samples from different countries and subtle genetic structure within each country. This approach captured population structure within the dataset at a far higher resolution than standard methods using unlinked single nucleotide polymorphism data, making it an appealing option for correcting subtle confounding in GWAS. We explored this possibility by fitting principal components (PCs) of this haplotype sharing matrix as covariates in a logistic regression model GWAS, and comparing metrics of statistical inflation and confounding against a model using standard independent marker PCs as covariates. We observed that both the λ_{GC} and LD-score regression intercept were significantly closer to 1 when using PCs of the haplotype sharing matrix, signifying lower inflation and confounding from population structure. Notably, the GWAS analysis that was corrected using haplotype sharing PCs as covariates retained the power to detect all major hits from the original meta analysis of the data, suggesting that it does not suffer from loss of power to detect true associations. We also detect an additional hit at the established ALS locus *TBK1*, which was sub-threshold in the original analysis, but has since been detected in larger ALS GWAS, implying that this method imparts greater power than traditional approaches in some scenarios. Our results demonstrate that using principal components of the pbwtPaint co-ancestry matrix as covariates in GWAS provides improved correction for confounding from subtle population structure without loss of power.

A- A+

PgmNr 223: Long-read single molecule, real-time (SMRT) sequencing of *NUDT15*: Phased full gene haplotyping and pharmacogenomic allele discovery.

Authors:
Y. Yang [1,2]; M.R. Botton [1,2]; E.R. Scott [2]; Y. Seki [1]; J. Harting [3]; P. Baybayan [3]; N. Cody [1,2]; P. Nicoletti [1,2]; T. Moriyama [4]; T. Lin [4]; S. Chakraborty [3]; J.J. Yang [4]; L. Edelmann [1,2]; E.E. Schadt [1,2]; J. Korlach [3]; S.A. Scott [1,2]

View Session Add to Schedule

Affiliations:
1) Sema4, a Mount Sinai venture, Stamford, CT 06902; 2) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029; 3) Pacific Biosciences, Menlo Park, CA 94025; 4) Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105

The *NUDT15* gene at chromosome 13q14.2 encodes a phosphatase enzyme that metabolizes multiple nucleotide substrates, including a central role in the conversion of the active thiopurine metabolite thioguanosine triphosphate to thioguanosine monophosphate. Importantly, patients that carry variant *NUDT15* alleles are predisposed to excessive thiopurine activation and hematopoietic toxicity when treated with thiopurines for malignant conditions or inflammatory bowel diseases, which prompted recent clinical practice guidelines on *TPMT*-guided thiopurine dosing to also include *NUDT15*. To date, the Pharmacogene Variation (PharmVar) Consortium has catalogued 19 *NUDT15* star (*) alleles; however, determining the phase of *NUDT15* sequence variants for accurate diplotype assignment is not feasible by short-read sequencing or targeted genotyping. Consequently, we developed a novel long-read single molecule, real-time (SMRT) sequencing method using the Pacific Biosciences Sequel platform, which was tested on 8.5 kb amplicons from 100 Ashkenazi Jewish (AJ) individuals. All multiplexed *NUDT15* amplicons were sequenced in a single SMRT Cell (28.5 Gb; mean length: 56.2 kb; average depth: 667X), which identified AJ allele frequencies that were consistent with reported frequencies in Europeans. However, a novel **1* sub-allele (c.-121G>A) and a rare likely deleterious coding variant (p.Pro129Arg) were also detected in the AJ, which were Sanger confirmed and submitted to PharmVar for star (*) allele assignment. A larger 10.5 kb *NUDT15* amplicon subsequently was developed with enhanced 5’ and 3’ coverage, which was validated using accuracy controls (**2*, **3*, **4*, **5*) from clinical samples (peripheral blood, saliva) and cell lines (21.5 Gb; mean length: 56.3 kb; average depth: 1300X). Triplicate *NUDT15* SMRT sequencing of two samples had non-reference genotype concordances of 1.0, and concordance with high-quality short read sequencing variants from four samples was also 1.0. Moreover, *NUDT15* SMRT sequencing identified a **2/*9* diplotype in one control sample previously reported as **1/*2*, indicating that the **9* in-frame deletion variant (p.13_14GlyVal[2]) was not detected by short-read sequencing. Taken together, these data indicate that long-read *NUDT15* SMRT sequencing is an innovative, reproducible, and validated method for phased full-gene characterization and novel allele discovery, which will improve *NUDT15* phenotype prediction for both research and clinical testing applications.

A- A+

PgmNr 224: Diagnostic utility of transcriptome sequencing for rare Mendelian diseases.

Authors:
H. Lee [1,2]; A.Y. Huang [3]; L. Wang [3]; A.J. Yoon [2]; G. Renteria [2]; R.H. Signer [2]; S. Nieves-Rodriguez [2]; C.G.S. Palmer [2,5,6]; J.A. Martinez-Agosto [2,4,5]; S.F. Nelson [1,2,3,4]; Undiagnosed Diseases Network

View Session Add to Schedule

Affiliations:
1) Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, California.; 2) Human Genetics, University of California, Los Angeles, Los Angeles, California.; 3) Institute for Precision Health, University of California, Los Angeles, Los Angeles, California.; 4) Pediatrics, University of California, Los Angeles, Los Angeles, California.; 5) Psychiatry & Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, California.; 6) Institute for Society and Genetics, University of California, Los Angeles, Los Angeles, California.

Clinical whole-exome sequencing (WES) has become a routine diagnostic test for rare Mendelian diseases with a diagnostic rate around 30%. WES-negative cases likely remain unsolved because the causal variants reside in a gene not yet associated with a human disease or is of a type not readily detectable by WES. Whole-genome sequencing (WGS) has the potential to observe variants missed by WES. However, the increase in the diagnostic rate attributed to WGS relative to WES has been modest, largely attributed to structural variant (SV) detection, which, is reported to resolve 3-7% of WES-negative cases. Transcriptome sequencing (RNAseq) can augment identification of causal variants in rare Mendelian diseases. However, the contribution to diagnosis remains unclear particularly for the relatively common area of neurodevelopmental diseases. To determine the value of RNAseq in ascertaining the consequence of DNA variants on RNA transcripts, affected subjects participating in the Undiagnosed Diseases Network (UDN) were selected for RNAseq from one or more accessible tissues, and the data was integrated with WGS for DNA variant interpretation genome-wide. Study participants included 113 probands and 25 similarly affected family members enrolled at the UDN-UCLA clinical site between July, 2014 and August, 2018. Thirteen families were excluded because clinical evaluation determined that genetic studies were not necessary. Of the remaining 100 families, the molecular diagnostic rate within exon coding sequence was 31% including the detection of SVs, and thus DNA was sufficient for diagnosis. In total 8 of these cases were solved by SV detection (26%, 8 of 31 cases). Integration of RNAseq with WGS, to detect mRNA defect attributable to an adjacent DNA variant, resulted in an additional 7 cases with clear diagnosis of a known genetic disease. The mRNA splicing abnormalities detected were exon skipping, intron retention and pseudoexon creation from synonymous and deep intronic DNA variants. Thus, the overall molecular diagnostic rate was 38%, and 18% (7 of 38 cases) of all genetic diagnoses returned required RNAseq to determine variant causality. In this rare disease cohort with a wide spectrum of undiagnosed, suspected genetic conditions, RNAseq increased the molecular diagnostic rate above that possible with WGS alone even without availability of the most appropriate tissue type to assess.

A- A+

PgmNr 2456: RNASeq enhances detection rates of exome sequencing of myocardial tissue from heart transplant patients with dilated cardiomyopathy.

Authors:
S. Amr [1,2]; E. Park [2]; M. Bowser [2]; R. Waikel [3]; M. Guglin [3]; K. Campbell [4]; A. Psychogios [3,5]

View Session Add to Schedule

Affiliations:
1) Department of Pathology, Harvard Medical School-Brigham and Women's Hospital, Boston, MA; 2) Translational Genomics Core, Partners HealthCare Personalized Medicine, Cambridge, MA; 3) Department of Internal Medicine/Division of Cardiology, University of Kentucky, Lexington, KY; 4) Department of Physiology and Division of Cardiovascular Medicine, University of Kentucky, Lexington, KY; 5) Department of Pediatrics/Division of Genetics, University of Kentucky, Lexington, KY

Background
The majority of sequencing variants identified in dilated cardiomyopathy (DCM) patients are classified as variants of uncertain significance (VUS). Transcriptional analysis of myocardial tissue of patients may provide the necessary evidence to clarify the clinical significance of variants. To assess this, we performed exome sequencing (ES) and RNASeq on myocardial tissue from DCM patients who underwent heart transplants to identify disease causing variants.

Methods
The study included myocardial tissue from18 DCM patients who underwent heart transplantation at the Gill Heart and Vascular Institute. Samples were provided by the Cardiovascular Biorepository at the University of Kentucky, and ES and RNASeq was performed at Partners HealthCare Personalized Medicine. Variants identified in ES were prioritized based on an indication-based gene list and classified per ACMG/AMP guidelines. The impact of candidate splice-site variants on relevant gene transcripts was assessed using RNASeq data.

Results
Disease causing variants were identified in 4/18 (22%) of patients, including 3 in the *TTN* gene and 1 in the *LMNA* gene. An additional 4 candidate VUS variants in 3 genes (2 in *TTN*, 1 in *AKAP9*, 1 in *TNNI3K*) were also detected. The two candidate *TTN* variants were splice site variants: c.10361-1G>A and c.36043+5G>C, which have been previously reported in ClinVar; however, their clinical assertions are “conflicting interpretations” and “variant of uncertain significance” respectively. RNAseq data generated using myocardial tissue from the two patients carrying these variants revealed that the variants abolish nearby nascent splice sites and introduce cryptic splice sites, that would lead to disruption of the normal reading frame of the protein. This information provided sufficient evidence to re-classify them as disease causing, thereby increasing the detection rate in our cohort to 33%. The other candidate variants in *AKAP9* and *TNNI3K* are both nonsense variants; however, these two genes have a “limited” association with only a handful of DCM patients reported with variants in these gene. Thus, our findings provide additional evidence to support the association of these genes to DCM.

Conclusion
RNASeq is an effective tool in assessing the impact of splice site variants identified in ES, thereby providing evidence to support clinical interpretation. In addition, we demonstrate the utility of ES in supporting disease-gene associations for candidate DCM genes.

A- A+

PgmNr 226: Characterization of human-derived transdifferentiated neurons for assessing splicing and allele expression in undiagnosed neurodevelopmental and neurological disorders.

Authors:
S. Nieves-Rodriguez [1]; F. Barthelemy [3,4]; J. Wan [1]; A. Huang [5]; L. Wang [5]; H. Lee [1,2]; M.C. Miceli [3,4]; S.F. Nelson [1,2,4]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics, University of California, Los Angeles, CA, USA; 2) Department of Pathology and Laboratory Medicine, University of California, Los Angeles, CA, USA; 3) Department of Microbiology, Immunology, & Molecular Genetics, University of California, Los Angeles, CA, USA; 4) Center for Duchenne Muscular Dystrophy, University of California, Los Angeles, CA, USA; 5) Institute of Precision Health, University of California, Los Angeles, CA, USA

Interpretation of non-coding variants from whole genome sequencing in undiagnosed Mendelian disorders is challenging. RNA sequencing (RNAseq) can be used to determine the consequence of a variant at the transcript level by revealing abnormalities in splicing, changes in gene expression levels, or loss of bi-allelic expression. Thus, RNAseq can be used to reveal novel disease mechanisms and improve diagnostic rate, but only if a gene of interest is expressed in an accessible patient tissue. The most commonly referred clinical indication for exome or genome sequencing are neurodevelopmental diseases, but about half of the known neurodevelopmental disease genes are not expressed in fibroblast, muscle or blood samples that are accessible, restraining the applicability of RNAseq for these diseases. Here, we investigated a rapid lentiviral-based method to transdifferentiate human skin-derived fibroblasts to different neuronal cell types to assess splicing and allelic expression of genes that are only expressed in CNS tissue. It relies on the expression of two neural conversion genes, *ASCL1* and *BRN2*, and a shRNA against a pan-neuronal gene inhibitor in non-neuronal cells, *REST*. Characterization of the sub-neuronal cell types obtained through our transdifferentiation protocol is performed at both RNA (RT-PCR, RNAseq) and protein (immunofluorescence) levels. Our data suggests that the induced cells present molecular signatures and properties of both immature and mature neuronal cells, including morphological changes, expression of *SOX2* (a neural stem cell marker) and upregulation *MAP2* (a mature neuronal marker). In addition, RNAseq data shows that various neuronal genes that are not expressed in accessible tissues are being expressed at adequate levels, including *RBFOX3* and *SYP.* Direct neuronal reprogramming of patient-derived cells provides ease, cost and time efficient advantages over neuronal differentiation from induced pluripotent stem cells. Preliminary data suggests that this direct method provides a gene expression profile that can be useful in the diagnosis of rare neurodevelopmental and neurological diseases. Further work is being performed to evaluate the extent to which transdifferentiation can be applied at scale among patients with undiagnosed neurodevelopmental and neurological diseases to improve diagnostic yield.

A- A+

PgmNr 227: FRASER: A statistical method to detect aberrant splicing events in RNA-seq data.

Authors:
C. Mertes [1,5]; I. Scheller [1,5]; V.A. Yépez [1,2]; Y. Liang [1]; F. Brechtmann [1]; H. Prokisch [3,4]; J. Gagneur [1,2]

View Session Add to Schedule

Affiliations:
1) Department of Informatics, Technical University of Munich, Garching, Germany; 2) Quantitative Biosciences Munich, Gene Center, Ludwig-Maximilians University of Munich, Munich, Germany; 3) Institute of Human Genetics, Helmholtz Center Munich, Neuherberg, Germany; 4) Institute of Human Genetics, Klinikum rechts der Isar, Technical University Munich, Munich, Germany; 5) Contributed equally to this work

---

Aberrant splicing is a major cause of Mendelian disorders. Regardless of significant improvements in sequence-based splicing prediction methods, in most cases no conclusive genetic diagnosis for splice variants can be reached after genome sequencing. Especially deep intronic variants causing splicing defects are hard to interpret and currently considered as variants of uncertain significance (VUS) or even not detected by whole exome sequencing. In recent studies, detecting splicing defects directly on the transcriptome level via RNA sequencing has proved to be an effective complementary avenue to genome sequencing. However, no specialized method exists for the detection of splicing outliers.

Here, we developed FraseR (Find RAre Splicing Events in RNA-seq data), a statistical method to detect rare splicing events. We use an annotation-free approach to detect novel splice sites and consider both exon-exon junction reads and reads overlapping exon-intron boundaries to detect intron retention events. We observed that the splicing metrics percent-spliced-in and splicing efficiency exhibit strong correlation structures between samples for many tissues in the GTEx dataset, although they are based on intra-sample isoform ratios. This suggested the importance of correcting for confounders. To control automatically for such covariation resulting from technical, environmental, or common genetic variations, FraseR is based on a denoising autoencoder. Moreover, the outlier detection is based on statistical testing using the beta-binomial distribution.

Benchmarking our approach on the GTEx dataset against Leafcutter, PCA controlled p-values and z-scores as well as uncontrolled p-values shows improved detection rate of simulated outliers. In addition, FraseR shows higher enrichments of rare splicing variants which shows the importance of RNA-seq as a complementary avenue. Finally, the application to a previously analysed rare disease dataset [1] led to a new diagnostic by identifying an aberrant exon truncation in gene *TAZ*. Altogether, these results indicate the importance of splicing specific outlier detection while controlling for known and unknown confounders based on RNA-seq data in the context of rare diseases.

[1] Kremer et al., Nature Communications (2017)

A- A+

PgmNr 228: Discovery of 310 novel loci for type-2 diabetes and related complications involving 1.4 million participants in a multi-ethnic meta-analysis.

Authors:
M. Vujkovic [1,2]; J. Keaton [6]; J.A. Lynch [4,5]; D. Miller [3]; J. Zhou [7]; S. Damrauer [8]; T. Assimes [9,10]; Y.V. Sun [11,12]; T. Edwards [6]; K. Cho [13]; S. Duvall [4]; P. Wilson [14,15]; D.J. Rader [16]; C.J. O'Donnell [13,17]; J. Meigs [17]; J.M. Gaziano [13,17]; P.S. Tsao [9,10]; L. Phillips [12,14]; P. Reaven [7]; K.M. Chang [1,18]; B.F. Voight [1,19]; D. Saleheen [1,2]; on behalf of the Million Veteran Program

View Session Add to Schedule

Affiliations:
1) Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA; 2) Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 3) ENR Memorial Veterans Hospital, Bedford, MA; 4) Department of Veterans Affairs Salt Lake City Health Care System, Salt Lake City, UT; 5) University of Massachusetts College of Nursing and Health Sciences, Boston, MA; 6) Division of Epidemiology, vanderBilt University Medical Center; 7) Phoenix VA Health Care System, University of Arizona, Arizona State University, Phoenix, AZ; 8) Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 9) VA Palo Alto Health Care System, Palo Alto, CA; 10) Department of Medicine, Stanford University School of Medicine, Stanford, CA; 11) Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA; 12) VA Atlanta Health Care System, Atlanta, GA; 13) VA Boston Health System, Boston, MA; 14) Schools of Medicine, Emory University, Atlanta, GA; 15) Department of Public Health, Emory University, Atlanta, GA; 16) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 17) Massachusetts General Hospital, Harvard Medical School, Boston, MA; 18) Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 19) Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA

Background:
The global epidemic of type 2 diabetes (T2D) is projected to reach 592 million by 2035. T2D is also an established risk factor for micro- and macrovascular diseases; yet the genetic mechanisms leading to T2D complications are largely unknown. To characterize the etiology of T2D and its major complications, we performed genetic association studies in the Million Veteran Program (MVP) and other multi-ethnic populations.

Methods:
The current analyses comprise 228,499 T2D cases and 1,178,783 controls of European, African, Hispanic, and Asian descent. Stratified by self-reported and genetically inferred ancestry, we conducted primary discovery analyses for T2D. We conducted the largest individual level analysis of the X-chromosome to date. Genome-wide interaction analyses were conducted to identify genetic variants whose effect on macrovascular (coronary heart disease [CHD], acute ischemic stroke and peripheral arterial disease [PAD]) and microvascular complications (retinopathy, nephropathy [CKD] and neuropathy) were modified by the presence of T2D. Genome-wide polygenic risk scores based on the DIAMANTE Consortium summary output were used to assess associations with T2D risk and

related complications by comparing the top [80-100%] against the reference quintile [0-20%] in participants with T2D.

Results:
We identified 589 distinct genomic regions in relation to T2D (P 5x10-8); of these, 310 were found to be novel, including 3 loci specific to African Americans and 7 novel loci on chromosome X. Genome-wide interaction analyses identified 19 novel loci that associated with diabetic complications (P 5x10-8), including 4 loci for CHD, 1 for PAD, 4 for ischemic stroke, 7 for retinopathy, 2 for CKD, and 1 locus for neuropathy. T2D polygenic risk scores were associated with an 3.4-fold increased T2D risk (95%CI 3.32-3.54) but modestly increased the risk for complications in diabetics, e.g. retinopathy (OR=1.43, P=4.2E-22), CHD (OR=1.13, P=7.9E-07), PAD (OR=1.11, P=8.9E-05), CKD (OR=1.12, P=3.1E-05), neuropathy (OR=1.09, P=0.002), and stroke (OR=1.08, P=0.084).

Conclusion:
We identified a total of 310 novel T2D risk loci in this large-scale multi-ethnic meta-analysis. The SNPxT2D interaction analysis additionally identified 19 loci that in presence of T2D showed a risk for diabetic complications that was greater than the sum of their individual effects. Associations were observed for a PRS comprised of T2D related variants with the risk of T2D complications.

A- A+

PgmNr 229: The high prevalence of pathogenic mutations in genes causing monogenic diabetes in patients with common type 2 diabetes opens avenues towards precision medicine.

Authors:
P. Froguel [1,2]; M. Boissel [2]; E. Durand [2]; B. Toussaint [2]; E. Vaillant [2]; S. Gaget [2]; R. Roussel [3]; B. Balkau [4]; M. Marre [3]; S. Franc [5]; G. Charpentier [5]; M. Vaxaillaire [2]; M. Canouil [2]; A. Bonnefond [1,2]

View Session Add to Schedule

Affiliations:
1) Imperial College London, London, United Kingdom; 2) CNRS UMR8199, Lille, France; 3) Inserm U1138, Centre de Recherche des Cordeliers, Paris, France; 4) Inserm U1018, Institut Gustave Roussy, Center for Research in Epidemiology and Population Health, Villejuif, France; 5) CERITD (Centre d'Etude et de Recherche pour l'Intensification du Traitement du Diabete), Evry, France

---

Genome-wide association studies have identified > 200 independent loci associated with type 2 diabetes (T2D) risk. Despite this success, the translation of these discoveries into advances in precision medicine has been modest. In contrast, the genetic investigation of monogenic forms of diabetes has revealed several key regulators of insulin secretion, leading to textbook cases of genomic medicine. Indeed, patients with a pathogenic mutation in *HNF1A* are optimally treated with sulfonylureas and patients with a pathogenic mutation in *GCK* do not require any hypoglycemic treatment. Here, in unselected patients with common T2D, we assessed the prevalence of pathogenic mutations across 33 genes causing monogenic diabetes, compared to normal glucose controls. Through next-generation sequencing (NGS), genes were sequenced in a case-control study including 6,348 individuals. As NGS-based targeted DNA sequencing has well-known caveats, we applied stringent quality control steps before any analyses. The pathogenicity of variants was assessed using the American College of Medical Genetics and Genomics criteria. Among the 6,348 individuals, we accurately identified 200 pathogenic or likely pathogenic variants. Among these variants, 37% were protein-truncating variants and 58% had previously been established as pathogenic. Using the MiST method adjusted for age, sex, body-mass index and ancestry, we found a significant association between pathogenic or likely pathogenic variants and increased T2D risk ($p<0.001$; with 6.5% mutation carriers among cases). This significant association was mostly explained by the mutations found in *GCK* ($p<0.001$) and *HNF1A* ($p<0.001$); i.e., the most frequently mutated genes in monogenic diabetes. Among the patients with common T2D, we found that the carriers of a (likely) pathogenic variant were only modestly leaner and developed T2D slightly earlier, but the use of antidiabetic drugs and the family history of T2D were not different between carriers and non-carriers. In conclusion, we showed high prevalence of pathogenic mutations in monogenic diabetes genes among common T2D patients who did not present with clinical characteristics of monogenic diabetes (i.e. MODY) and thus could not be identified without NGS. These results open avenues towards precision diabetic medicine, and to savings in health expenditure via inexpensive ($50) systematic genetic characterization of newly diagnosed patients.

A- A+

PgmNr 230: Chromatin accessibility patterns of a hiPSC model of islet development highlight type 2 diabetes risk loci in beta cell differentiation.

Authors:
M. Pérez Alcántara [1]; M. Jansen [1]; M. Thurner [1,2]; A.L. Gloyn [1,2,3]; A. Mahajan [1]; A. Wesolowska-Andersen [1]; M.I. McCarthy [1,2,3]

View Session Add to Schedule

Affiliations:
1) Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 2) Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford, United Kingdom; 3) Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

---

Most variants associated with type 2 diabetes (T2D) risk in genome-wide association studies (GWAS) act through reduced insulin secretion, due to deficiencies in islet development and/or mature islet function. While functional studies have focused on the latter, we studied patterns of open chromatin (ATAC-seq) in three human iPSC lines (from different donors) differentiated *in vitro* towards pancreatic beta cells, to highlight T2D-associated variants likely to exert their effects during islet development. We detected 128,427 ATAC-seq peaks open at one or more of the 7 stages of islet development. These peaks were clustered using weighted gene co-expression network analysis (WGCNA), grouping those with similar chromatin accessibility profiles across the stages. This defined 12 modules of correlated ATAC-seq peaks of which 4 were enriched in T2D SNPs from the European ancestry DIAMANTE analysis (~ 900,000 participants) using fGWAS, including modules with peak accessibility in middle and final developmental stages. These modules were enriched in binding sites of transcription factors (HOMER *p*-value $\leq 10^{-10}$) implicated in islet development (e.g. *PDX1*, *FOXA2*) and within T2D-associated loci (e.g. *MNX1, TCF7L2, HNF6, HNF4A*), whose expression at those stages was confirmed by RNA-seq. We then used the fGWAS estimates from the enriched modules as priors to redefine the sets of variants at each locus that collectively account for 99% posterior probability of association (PPA), which increased the PPA of individual signals over a 70% threshold at 8 loci including *DGKB* and *IGF2BP3*.
To predict changes in ATAC-seq peaks based on the DNA sequence within them we trained a convolutional neural network. We predicted effects of T2D-associated variants on open chromatin during development and found 51 candidates among variants with PPA>=10%. The most prominent was rs17712208, a secondary association signal at the *PROX1* locus: its alternative A allele predicted to decrease chromatin accessibility, particularly in mid-development (*p*: 7.59×10^{-26}). Finally, we found that ATAC-seq peaks are significantly closer to genes presenting highly correlated expression patterns across development, suggesting that active enhancers marked by these peaks are preferentially regulating genes in cis. In conclusion, relating ATAC-seq with RNA-seq patterns could link enhancers to effector transcripts and help elucidate the link between genome regulation during islet development and T2D pathogenesis.

A- A+

PgmNr 231: Identifying mechanisms of type 1 diabetes risk by integrating genetic fine-mapping, high-throughput transcription factor binding and chromatin accessibility.

Authors:
P. Benaglio [1]; J. Chiou [1,2]; S. Corban [1]; M. Okino [1]; A. Aylward [1,3]; J. Yan [4,5]; N. Nariai [6]; B. Ren [4,6]; K. Frazer [1,6]; K. Gaulton [1]

View Session Add to Schedule

Affiliations:
1) Pediatrics, University of California San Diego, La Jolla, CA; 2) Biomedical Sciences University of California San Diego, La Jolla, CA; 3) Bioinformatics & Systems Biology, University of California San Diego, La Jolla, CA; 4) Ludwig Institute for Cancer Research, La Jolla, CA; 5) Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden; 6) Institute of Genomic Medicine, University of California San Diego, La Jolla, CA

Genetic studies have identified >60 type 1 diabetes (T1D) risk loci, but the identification of their underlying mechanisms is challenging because associated variants are mainly non-coding. To identify causal T1D risk variants and understand their mechanisms, we performed fine-mapping of T1D loci using a large GWAS meta-analysis of 16,272 T1D cases and 382,828 controls from 7 cohorts imputed into the HRC panel. We identified a total of 83 independent association signals through conditional analyses of known T1D loci, and performed Bayesian fine-mapping of each signal. The 99% credible sets of these signals contained on average 63 variants, and 41 signals contained at least one fine mapped variant with >25% posterior probability. We next employed multiple strategies to prioritize functional regulatory variants within T1D credible sets. First, we tested 94k variants at T1D loci for allelic differences in transcription factor binding by generating high-throughput SELEX-seq data for 530 TFs. Second, we generated ATAC-seq data from pancreatic islets before and after stimulation with pro-inflammatory cytokines (IL-1, IFN-γ, TNF-α), and combined this dataset with published ATAC-seq from stimulated and unstimulated immune cells. By integrating SELEX-seq and ATAC-seq data, we identified drivers of regulatory variant activity in stimulated and unstimulated immune and islet cells such as TFs in the JUN, BACH, IRF and NFATC families. Variants in stimulated chromatin were more enriched for T1D association compared to unstimulated sites using LDSC partitioned heritability. Stimulated chromatin sites were also strikingly similar between immune and islet cells, suggesting that a subset of T1D risk variants might affect stimulated chromatin in both tissues. We prioritized fine-mapped T1D variants located in stimulated chromatin in both immune cells and islets that showed differential TF binding, and identified variants with allele-specific activity at the *RBPJ* and *RAD51B* loci using gene reporter and EMSA assays in T cell (Jurkat) and beta cell (MIN6) lines. Our results combine comprehensive fine mapping and detailed functional annotation to reveal novel mechanisms of T1D risk and highlight dual roles for a subset of T1D variants affecting both immune and beta cell function.

A- A+

PgmNr 232: Frequency of pathogenic germline variation in pediatric pan-cancer survivors: A report from the Childhood Cancer Survivor Study (CCSS).

Authors:
D.M. Gianferante [1]; J. Kim [1]; D. Karyadi [1]; S. Hartley [1]; M. Frone [1]; L. Robinson [2]; G. Armstrong [2]; S. Bhatia [3]; M. Dean [1,4]; M. Yeager [4]; B. Zhu [1]; L. Song [1]; S. Brodie [1]; K. de Andrade [1]; K. Santiago [5]; A. Goldstein [1]; P. Khincha [1]; M. Machiela [1]; M. McMaster [1]; M. Nickerson [1]; L. Oba [1]; A. Pemov [1]; M. Pinheiro [1]; M. Rotunno [1]; M. Tucker [1]; L. Morton [1]; S. Chanock [1]; S. Savage [1]; D. Stewart [1]; L. Mirabello [1]; NCI-DCEG Cancer Genomics Research Laboratory

View Session Add to Schedule

Affiliations:
1) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA; 2) Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, TN, USA; 3) Institute for Cancer Outcomes and Survivorship, University of Alabama at Birmingham, Birmingham, AL, USA; 4) Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research, Frederick, MD, USA; 5) International Research Center, A.C. Camargo Cancer Center, São Paulo, Brazil

Introduction. Pediatric cancer is the leading cause of death by disease in children despite improved survival rates. Here, we conducted the largest pediatric pan-cancer study to date using exome sequencing (ES) of 5,451 long-term survivors to quantify the frequency of germline pathogenic variation in cancer-susceptibility genes (CSG).

Methods. The CCSS is a multi-center retrospective cohort of children diagnosed <21 years of age who survived at least 5 years from a diagnosis of leukemia, lymphoma, brain tumors, neuroblastoma, Wilms tumor, bone cancers, and soft tissue sarcomas. Germline DNA underwent ES; analyses focused on rare variants in 237 previously published CSGs. Pathogenic/likely pathogenic (P/LP) variants were identified using ClinVar, HGMD with manual review, and InterVar. We compared the frequency of P/LP variants with the same pathogenicity criteria in 5,105 European-ancestry cases (CEU >0.8) vs. 51,377 gnomAD non-Finnish European cancer-free controls (ES data) using the Fisher's exact test and corrected for multiple testing using a false discovery rate (FDR; q<0.05 considered significant). An unadjusted Kaplan-Meier analysis was used to estimate the differences in overall survival in children with and without P/LP variants.

Results. In 5,105 European-ancestry pediatric cancer survivors, 11% harbored a P/LP variant in a dominant CSG (n=176) vs. 9% in controls (P<0.0001). Eight dominant genes (*NF1*, *WT1*, *TSC1*, *REST*, *KMT2D*, *EZH2*, *CDKN2A*, *MEN1*) had significantly more P/LP variants in cases vs. controls (FDR q<0.05). We identified a novel germline cancer risk association in *KMT2D* (odds ratio [OR] 16.8, 95% CI 4.5-63.5, vs. controls). *KMT2D*, underlying Kabuki Syndrome, has only previously been associated somatically with cancer. In *EZH2* (OR 4.7), *CDKN2A* (OR 8.6), and *MEN1* (OR 20.1), we found multiple novel pediatric cancer associations. Children carrying a P/LP variant had worse survival compared to children without a P/LP variant (P=0.001). Children with one P/LP variant in *SBDS*, an autosomal recessive CSG, had a significantly increased risk of cancer (FDR q<0.05).

Conclusion. To our knowledge, this is the largest pediatric pan-cancer study of pathogenic germline variation to date. In long-term cancer survivors, we found multiple novel gene-cancer associations and worse survival in children with a P/LP variant. These findings could have implications in cancer risk stratification and genetic counseling for patients and families.

A- A+

PgmNr 233: Evaluating the prevalence of pathogenic variants in cancer predisposition genes among children with newly diagnosed rhabdomyosarcoma: A report from the Children’s Oncology Group.

Authors:
H. Li [1]; S.D. Sisoudiya [2,3]; D.S. Hawkins [4]; G.C. Kendall [5]; J.F. Amatruda [5]; S.X. Skapek [5]; S. Dugan-Perez [1]; D.A. Marquez-Do [2]; M.E. Scheurer [2]; D. Muzny [1]; R.A. Gibbs [1,3]; S.E. Plon [1,2,3]; P.J. Lupo [2]; A. Sabo [1]

View Session Add to Schedule

Affiliations:
1) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 2) Section of Hematology-Oncology, Department of Pediatrics, Baylor College of Medicine, Houston, TX; 3) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 4) Seattle Children’s Hospital, Seattle, WA; 5) UT Southwestern Medical Center, Dallas

Background: Rhabdomyosarcoma (RMS) is a highly malignant tumor and the most common soft tissue sarcoma in children. Cancer predisposition syndromes are among the strongest risk factors for RMS; however, there have been no large-scale efforts to systematically annotate the prevalence of these variants in an unselected cohort of children with RMS.

Methods: We are analyzing germline DNA samples from 900 patients with newly diagnosed RMS that were consented to a Children’s Oncology Group biology and banking protocol. To date, we have completed whole-exome sequencing on 270 samples. For this interim assessment, we evaluated the prevalence of rare (minor allele frequency <0.01) pathogenic variants in 60 autosomal dominant cancer predisposition genes (CPGs). We also compared the prevalence of germline pathogenic variants between the two most common RMS histological subtypes: embryonal (ERMS) and alveolar (ARMS), which is largely characterized by *PAX-FOXO1* fusions.

Results: We identified 18 patients that carry rare variants previously reported as pathogenic or likely pathogenic (P/LP) in the ClinVar database within 10 of the 60 CPGs: *BRCA2* (n=2), *CBL*, *DICER1* (n=2), *HRAS* (n=3), *MSH6*, *NF1* (n=3), *PMS2*, *PTEN*, *SDHA*, and *TP53* (n=3). In addition, we identified novel putative loss of function variants in *NF1, SHDC* and *RIT1*. These P/LP variants account for 7.8% of newly diagnosed RMS patients. Notably, 12.1% (14/116) of children with ERMS harbored these variants compared to 3.6% (3/84) among those diagnosed with ARMS (p=0.04). Among those ARMS cases where *PAX-FOXO1* status was known, 10% of fusion-negative cases harbored P/LP variants, whereas none of the fusion-positive cases had a P/LP in the 60 CPGs. Notably, three ERMS patients had well-described oncogenic variants in *HRAS*: two with G12V and one with G12S. These variants are associated with Costello syndrome, a rare disorder associated with risk of RMS and other embryonal tumors.

Conclusion: Our initial results demonstrate that genetic risk of RMS results from pathogenic variants in a wide spectrum of CPGs. Completion of this study should confirm the extent of germline findings in children with RMS, including the prevalence of RMS patients who have molecular features of Costello syndrome, and will also provide the opportunity for new gene discovery. Additionally, our interim findings suggest that germline testing of RMS patients should not be restricted to a limited set of genes.

A- A+

PgmNr 234: The Healthy Oregon Project: Initial laboratory findings for inherited cancer predisposition screening in a large population in the Pacific Northwest.

Authors:
T.D. O'Brien [1]; A.B. Potter [1]; A. Kulkarni [1]; G. Goh [1,2]; J.H. Letaw [1,3]; C.S. Dahl [1]; J. Pleyte [1]; J. Thanner [4]; T.J. McFarland [5]; S. Medica [2,5]; C.A. Harrington [5,6]; K.J. Hamman [6]; J. Sampson [6]; K. Johnson-Camacho [2]; J. Shannon [7,8]; P.T. Spellman [2,6]; C.S. Richards [1,6]

View Session Add to Schedule

Affiliations:
1) Knight Diagnostic Laboratories, Oregon Health & Science University, Portland, OR; 2) Cancer Early Detection Advanced Research Center, Knight Cancer Institute, Oregon Health & Science University, Portland, OR; 3) Department of Computational Biology, Oregon Health & Science University, Portland, OR; 4) Information Technology Group, Oregon Health & Science University, Portland, OR; 5) Integrated Genomics Laboratory, Oregon Health & Science University, Portland, OR; 6) Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR; 7) OHSU-PSU School of Public Health, Portland, OR; 8) Knight Cancer institute, Community Outreach and Engagement, Oregon Health & Science University, Portland, OR

Here we report the initial genetic findings from the Healthy Oregon Project (HOP), an IRB-approved clinical research study sponsored by the Cancer Early Detection Advanced Research Center (CEDAR) at the OHSU Knight Cancer Institute. HOP was initiated to survey and influence the health of the Oregon population. The overarching aim of HOP is to develop a massive population database of information including health habits, family history, and genetic testing results (when available) to learn how these factors influence personal risk for disease and to provide data to participants to help guide their health management. The initial driving project is the assessment of participants' inherited cancer predisposition. HOP participants consented to be in the study and filled in questionnaires about health practices and personal and family medical history. Participants could choose to have genetic testing to assess their cancer risk, and if so, provided a mouthwash sample to the Knight Diagnostic Laboratory. Here we report our laboratory's challenges and solutions in developing and implementing a cost-effective clinically validated next generation sequencing (NGS) screening panel for 32 genes with strong evidence of an associated inherited cancer predisposition, as well as results from the pilot. For DNA preparation and NGS testing of thousands of samples, automation using multiple robotic systems was critical. Unique bioinformatics approaches were developed to rapidly identify known pathogenic or novel likely pathogenic variants (other types of variants are not reported), and an in-house database was built to analyze and report results. While negative reports are returned directly to the participant using a HIPAA compliant smartphone app, positive results are confirmed by a second sample using our standard clinical test and returned to participants by direct contact with a genetic counselor. To date approximately 2800 participants consented to enroll in HOP, and approximately 80% chose to have the genetic screening test. Our laboratory has now performed testing for over 1200 samples and anticipates completing 2000 prior to this presentation. Our initial findings show a higher than expected positive rate of disease-causing variants of approximately 7% but only 5% are reportable based on our return of results strategy. Laboratory insights and lessons learned about management of a large scale population-based clinical study will be discussed.

A- A+

PgmNr 235: Limitations of direct-to-consumer genetic screening for hereditary breast, ovarian, and colorectal cancer risk.

Authors:
E.D. Esplin; E. Haverfield; S. Yang; B. Herrera; M. Anderson; R. Nussbaum

View Session Add to Schedule

Affiliation: Invitae, San Francisco, California.

Background
Genetic screening for hereditary cancer risk is a growing opportunity for personalized preventive medicine. The FDA authorized a direct-to-consumer (DTC) screen to report on 3 *BRCA1/BRCA2* Ashkenazi Jewish (AJ) founder variants, and 2 *MUTYH* variants common in those of northern European [NE] descent (c.536A>G, c.1187G>A). We determine how often DTC genetic screening for these 5 variants would falsely reassure individuals of a low risk for a hereditary cancer syndrome (HCS).

Methods
We analyzed de-identified data from: 1) An indication-based cohort of 270,806 patients referred by healthcare providers for large panel genetic testing including *MUTYH* due to personal/family history of any cancer, 2) An indication-based cohort of 119,328 patients referred by healthcare providers for *BRCA1/2* genetic testing due to personal/family history of breast or ovarian cancer. Cohorts were stratified by self-reported ethnicity.

Results
In cohort 1), 5,929 patients had a pathogenic or likely pathogenic (P/LP) variant in *MUTYH*, of which 4,552 had one of the NE variants. In total, 40% of patients homozygous or compound heterozygous for *MUTYH*, and an additional 22% of carriers, would have been missed by screening limited to the NE *MUTYH* variants, and clinically false-negative. By ethnicity, clinical false negative rates for *MUTYH* would be: Asian 100%, African-American 75%, Hispanic 46%, Caucasian 33%.

In cohort 2), 4,733 had a P/LP variant in *BRCA1/2*. Among patients with P/LP variants in *BRCA1/2*, 12% carried one of the 3 AJ founder mutations: 81% of AJ had one of the 3 founder mutations, but only 6% of non-AJ patients. This results in a potential clinical false-negative rate of 19% and 94% among AJ and non-AJ individuals, respectively. By ethnicity, clinical false-negative rates for *BRCA1/2* would be: Asian 98%, African-American 99%, Hispanic 94%, Caucasian 94%.

Conclusions
We found the vast majority of individuals would have received clinical false negative results in *MUTYH* and *BRCA1/2* when screening was limited to the NE *MUTYH* and AJ founder variants. This approach is an even greater disservice to underrepresented non-NE populations, with predicted clinical false-negative rates approaching 100% in certain populations. The limitations of DTC screening and FDA recommendations for confirmatory testing of every individual who screening, positive or negative, may not be well understood by consumers and should therefore be used with caution.

A- A+

PgmNr 236: Functionally-informed fine-mapping improves polygenic localization of complex trait heritability.

Authors:
F. Hormozdiari [1]; O. Weissbrod [1]; C. Benner [2]; R. Cui [3]; J. Ulirsch [3]; A. Schoech [1]; S. Gazal [1]; Y. Reshef [1]; B. Geijn [1]; C. Márquez-Luna [1]; L. O'Connor [1]; H. Finucane [3]; A. Price [1]

View Session Add to Schedule

Affiliations:
1) Harvard T.H. Chan School of Public Health, Harvard Univ, Boston, Massachusetts.; 2) Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland,; 3) Broad Institute of MIT and Harvard, Cambridge, MA

Fine-mapping aims to identify causal variants in GWAS. Several recent methods improve fine-mapping accuracy by prioritizing variants in enriched functional annotations. However, these methods can only use information at significant GWAS loci, limiting the benefit of functional data.

We propose PolyFun, a statistical framework to improve fine-mapping using genome-wide functional data. PolyFun prioritizes variants in enriched functional annotations by defining prior causal probabilities for fine-mapping methods such as FINEMAP (Benner *et al.* 2016 Bioinformatics) or SuSiE (Wang *et al.* 2018 bioRxiv). PolyFun robustly estimates prior causal probabilities by (1) estimating expected per-SNP heritabilities based on a broad set of annotations from the baseline-LF model (Gazal *et al.* 2018 Nat Genet) via L2-regularized stratified LD-score regression (S-LDSC) using odd (resp. even) chromosomes; (2) partitioning SNPs into 20 bins with similar per-SNP heritabilities; (3) re-estimating per-SNP heritabilities in each bin using even (resp. odd) chromosomes; and (4) setting prior causal probabilities proportional to per-SNP heritabilities. In simulations with *N*=337K British-ancestry UK Biobank individuals (imputed MAF>0.1% SNPs), PolyFun + SuSiE was well-calibrated and identified >31% more fine-mapped SNPs with posterior causal probability >0.95 vs. SuSiE or functionally-informed CAVIARBF.

We applied PolyFun + SuSiE to 28 UK Biobank traits and identified 2,346 fine-mapped SNPs with posterior causal probability >0.95, a >20% improvement vs. SuSiE; 529 SNPs were fine-mapped for multiple traits, indicating pervasive pleiotropy. The number of fine-mapped SNPs ranged from 24 (age at menarche) to 345 (height). In a reduced analysis of *N*=107K UK Biobank individuals, PolyFun + SuSiE identified >33% more fine-mapped SNPs vs. SuSiE, with similar rate of replication in *N*=337K SuSiE results.

We used posterior per-SNP heritabilities from PolyFun + SuSiE to estimate the number of SNPs causally explaining a given proportion of heritability, re-estimating per-SNP heritabilities in bins of similar magnitude using S-LDSC to avoid winner's curse. The number of SNPs causally explaining 50% of common SNP heritability ranged widely from 0.1K (LDL cholesterol) to 243K (BMI), greatly improving on prior results based solely on functional annotations (192K and 330K, respectively). Our results provide insights into complex trait architectures and can help prioritize variants for functional follow-up.

A- A+

PgmNr 237: Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations.

Authors:
Li [1,2,3,4,5]; Y. Luo [1,2,3,4,5]; X. Wang [6]; S. Gazal [3,7]; J.M. Mercader [3,8]; B.M. Neale [3,9]; J.C. Florez [3,8,10]; A. Auton [6]; A.L. Price [3,7,11]; H.K. Finucane [3]; S. Raychaudhuri [1,2,3,4,5,12]; 23andMe Research Team, SIGMA Type 2 Diabetes Consortium

View Session Add to Schedule

Affiliations:
1) Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 2) Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 3) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4) Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; 5) Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 6) 23andMe, Inc., Mountain View, California, USA; 7) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; 8) Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; 9) Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA; 10) Department of Medicine, Harvard Medical School, Boston, MA, USA; 11) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; 12) Arthritis Research UK Centre for Genetics and Genomics, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

Non-Europeans populations, especially admixed populations, like African Americans and Latinos, have been so far underrepresented in current genetic studies. Most statistical methods for understanding the genetic mechanisms of complex traits assume homogeneous populations. For example, summary statistics-based methods for estimating heritability (h_g^2) and its enrichments, such as linkage disequilibrium score regression (LDSC), rely on accurate LD estimation. However, LD in admixed populations is confounded by admixture, which makes LDSC and other summary statistics-based methods not suited for these populations.

Here, we introduce covariate adjusted LDSC (cov-LDSC), a novel extension of LDSC, to accurately estimate h_g^2 and its enrichment in admixed populations. In cov-LDSC, we regressed the covariates out of the raw genotypes and calculated the covariate-adjusted LD scores on the adjusted genotype matrix. Through extensive simulations in both simulated genotypes and the SIGMA cohort (The SIGMA Type 2 Diabetes Consortium, 2014), we concluded that original LDSC, as expected, underestimated h_g^2 by 10%-60%; while cov-LDSC obtained unbiased h_g^2 estimates.

We then applied cov-LDSC on 161,894 Latino, 46,844 African American and 134,999 European research participants from 23andMe to estimate h_g^2 and enrichments, making this, to our knowledge, the most comprehensive heritability-based analysis of admixed individuals. For the seven traits we tested (height, BMI, morning person, left handedness, motion sickness, nearsightedness, and age at menarche), we observed that most of the h_g^2 estimates were similar across populations except significant difference in age at menarche ($p=7.1*10^{-3}$ between Latinos and Europeans). We extended cov-LDSC to estimate partitioned heritability in sets of genes that are specifically expressed in

different tissue and cell types (Finucane, et al., 2018). We observed highly consistent tissue-type enrichments across three populations: central nervous system (CNS) for BMI, musculoskeletal and connective tissues for height, and CNS for morning person. Our results also recapitulated known biological mechanisms, for example, limbic system is the most enriched tissue in BMI (per-standardized-annotation effect size=0.18, 0.16, 0.28, 0.18 in Europeans, Latinos, African Americans, and meta-analysis respectively). Our results demonstrate that cov-LDSC is a powerful way to analyze genetic data for complex traits from underrepresented populations.

A- A+

PgmNr 238: Efficient variance components analysis across millions of genomes.

Authors:
A. Pazokitoroudi [1]; Y. Wu [1]; K. S. Burch [2]; K. Hou [3]; A. Zhou [1]; B. Pasaniuc [3,4,5]; S. Sankararaman [1,4,5]

View Session Add to Schedule

Affiliations:
1) Department of Computer Science, UCLA, Los Angeles, California; 2) Bioinformatics Interdepartmental Program, UCLA, Los Angeles, California; 3) Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California; 4) Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, California; 5) Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California

---

Variance components analysis has emerged as a versatile tool to probe the genetic basis of complex traits, with applications ranging from heritability estimation to association mapping. While the application of these methods to large-scale genetic datasets has the potential to reveal important insights into genetic architecture, existing methods for fitting multiple genetic variance components on these large datasets are impractical.
Here, we present a randomized extension of Haseman-Elston regression for multiple variance components estimation and demonstrate its utility in estimating and partitioning SNP-heritability of complex traits. Our proposed algorithm requires only a few hours to estimate hundreds of variance components from millions of individuals and SNPs. Across a wide range of simulations, we show that our method yields unbiased estimates of genome-wide SNP-heritability by fitting multiple genetic variance components to account for frequency and LD-dependent effects. We then estimate SNP-heritability for 22 complex traits in the UK Biobank (N=290K) and find that methods such as stratified LD-score regression and SumHer yield SNP-heritability estimates that are higher by 2.5% and 25% on average, respectively, compared to our approach. In addition, we partition heritability by minor allele frequency (MAF) and LD bins and observe that SNPs with lower LD tend to have higher heritability enrichment than SNPs with higher LD for both common (seven-fold more enrichment on average over 22 traits) and low-frequency SNPs (eight-fold more enrichment on average over 22 traits), consistent with reports of the impact of negative selection on complex traits. Comparing the quartile with the lowest LD score to the quartile with the highest LD score, height showing similar increase in heritability enrichment for common and low-frequency SNPs (7.7 fold and 5.6 fold for low-frequency and common SNPs) while systolic blood pressure shows a greater increase in the low-frequency SNPs relative to common SNPs (18 fold for low-frequency vs 5 fold for common SNPs).

A- A+

PgmNr 239: The role of dominance in complex traits in the UK Biobank.

Authors:
D.S. Palmer [1,2,3]; W. Zhou [1,2,3]; A. Bloemendal [1,2,3]; L. Abbott [1,2,3]; N.A. Baya [1,2,3]; B.M. Neale [1,2,3]

View Session Add to Schedule

Affiliations:
1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts.; 2) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, Massachusetts.; 3) Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, , Massachusetts.

For GWAS, dominance genetic effects are defined as the deviation from a purely additive (or dose response) genetic effect. Dominance effects are well documented in model organisms, although the evidence in humans is limited to a handful of traits, particularly those with strong single locus effects such as hair color. Here, we perform a systematic evaluation of the impact of dominance effects on the contribution to phenotypic variance in >3,000 traits in the UK Biobank. To do so, we develop and apply a new method in the LD-score framework to quantify the “dominance heritability”. To estimate dominance heritability, we require dominance encoded association test statistics and dominance encoded LD scores. The dominance encoding of the genotypes is orthogonal to the additive encoding.

Here we present an overview of the impact of dominance on complex traits. First, we performed >3,000 dominance GWAS scans, identifying 337 loci at genome-wide significance, recapitulating many well known signals in phenotypes with dominant and recessive patterns of inheritance (e.g., hair color), as well as a number of novel dominance loci. Second, we estimate the dominance heritability for all traits, showing limited evidence of a substantial dominance contribution to phenotypic variance. We also estimate the sample sizes necessary to capture clear evidence of dominance. Finally, we introduce dominance fine-mapping to enable more accurate estimation of causal variants in the presence of a dominance signal. We have made all dominance summary statistics and heritability estimates publicly available for download. These results provide the most comprehensive evaluation in humans of dominance to trait variance to date.

PgmNr 240: Characterizing portability of complex trait associations across the diverse populations of the PAGE Study.

Authors:
C. Gignoux [1]; M. Graff [2]; S. Bien [3]; R. Tao [4]; J. Haessler [3]; H. Highland [2]; Y. Patel [5]; L. Hindorff [6]; S. Buyske [7]; C. Haiman [5]; L. Le Marchand [8]; R. Loos [9]; T. Matise [7]; U. Peters [3]; K. North [2]; C. Avery [2]; C. Kooperberg [3]; E. Kenny [9]; G. Wojcik [10]; The PAGE Study

View Session Add to Schedule

Affiliations:
1) Colorado Center for Personalized Medicine, Univ of Colorado, Anschutz Medical Campus, Aurora, Colorado.; 2) University of North Carolina, Chapel Hill, NC; 3) Fred Hutchinson Cancer Research Center, Seattle, WA; 4) Vanderbilt University Medical Center, Nashville, TN; 5) Keck School of Medicine, University of Southern California, Los Angeles, CA; 6) National Human Genome Research Institute, Bethesda, MD; 7) Rutgers University, New Brunswick, NJ; 8) University of Hawaii, Honolulu, HI; 9) Icahn School of Medicine at Mount Sinai, New York NY; 10) Stanford University, Stanford, CA

As sample sizes for studies of human health and disease extend into the hundreds of thousands to millions, the field of human genetics continues to focus largely on populations of European descent. Understanding the portability of genetic associations to other populations is a critical step in any downstream analyses and interpretation of GWAS findings, such as trans-ethnic fine mapping, functional -omics imputation or polygenic scores.

We extend the findings of the Population Architecture using Genomics and Epidemiology (PAGE) Study by assessing the probability of replicating variants in the GWAS Catalog across 11 large ethnic groups (N=400 to 17,299 from 49,839 individuals in total). To investigate ancestry-related patterns influencing population-stratified associations, we formulate a post-hoc logistic regression to predict replication of GWAS findings in each of these 11 populations in a model accounting for sample size, ancestry cluster proportions, and minor allele frequency. The regression results offer a quantitative measure of the influence of ancestry on the probability of replication. Among other findings, we demonstrate that ancestry shapes portability of findings across populations in a trait-specific manner, where we observe stronger signals when stratifying by trait than in aggregate across all traits. We show ancestry-specific effects of LDL variant portability, with increasing European ancestry increasing portability (OR per decile of ancestry: 1.20, 95% CI 1.04-1.37), and Indigenous American ancestry decreasing portability (OR: 0.88, 95% CI 0.79-0.98). In contrast, for HDL we observe no significant effects. Similarly, we see a large effect of percent African ancestry affecting portability of white blood cell count findings (OR: 1.54, 95% CI 1.06-2.25), with no similar ancestry affect for platelet count. We further extend this work to characterize the joint roles of global and local ancestry on portability, examine aspects of phenotypic measurement throughout time in longitudinal cohorts, and characterize findings across a number of cardiac, kidney, lipid, inflammatory, and lifestyle traits. Together, these effects help characterize and predict the degree of portability of genetic findings across traits, and provide unique information only present in a large, multi-phenotype, multi-population dataset such as PAGE.

A- A+

PgmNr 241: Application of European-derived polygenic risk score to admixed populations reveals strong biases by ancestry.

Authors:
G. Wojcik [1]; S.A. Bien [2]; M. Graff [3]; C.L. Avery [3]; S. Buyske [4]; C. Haiman [5]; L. Le Marchand [6]; C. Kooperberg [2]; R.J.F. Loos [7]; T.C. Matise [4]; K.E. North [3]; U. Peters [2]; S.S. Rich [8]; C.R. Gignoux [9]; C.D. Bustamante [1]; E.E. Kenny [7]

View Session Add to Schedule

Affiliations:
1) Stanford University, Stanford, California.; 2) Fred Hutchinson Cancer Research Center, Seattle, WA; 3) University of North Carolina at Chapel Hill, Chapel Hill, NC; 4) Rutgers University, Piscataway, NJ; 5) University of Southern California, Los Angeles, CA; 6) University of Hawaii, Honolulu, HI; 7) Icahn School of Medicine at Mount Sinai, New York, NY; 8) University of Virginia, Charlottesville, VA; 9) University of Colorado Anschutz Medical Campus, Aurora, CO

In recent years, the polygenic risk score (PRS) has become an important tool to capture the architecture of disease for clinical risk prediction. However, the majority of risk scores are developed in European-descent populations and this may exacerbate health disparities. To understand the relationship between population substructure and the performance of PRS, we applied a recently published PRS of obesity (2.1 million variants; Khera et al, 2019) to the multi-ethnic Population Architecture using Genomics and Epidemiology (PAGE) Study (Hispanic/Latino (H/L; n=21,982), African American (AA; n=17,040), Asian (n=4,672), Native Hawaiian (NatHw; n=3,907), Native American (NatAm; n=643)). We find significantly different distributions of the standardized risk score between groups (P<0.001), with the highest mean scores in Asian participants (+0.31 SD), despite their having the lowest mean BMI (25.3 kg/m^2). These discrepancies are reflected in the varying correlation (R^2) between the risk score and BMI (Asian: 0.017, H/L: 0.029, AA: 0.017, NatHw: 0.003, NatAm: 0.035). In addition to differential accuracy between racial/ethnic groups, we find confounding by ancestry within admixed groups, with strong correlation between PRS and proportion of Native American ancestry within H/L (r=0.32) and proportion of African ancestry within AA (r=0.21). Among Hispanic/Latino participants, adjustment for proportion of Native American ancestry increases risk score correlation with BMI (R^2: 0.038 to 0.044). Adjustment also changes risk classification, most strikingly at extremes with the top PRS decile as “high risk” and the bottom PRS decile as “low risk”. After adjustment, 27.33% of the high risk group are downgraded to normal risk, of which 92.94% are in the highest quintile of Native American ancestry. Conversely, 20.67% of the low risk group are upgraded to normal risk after adjustment, of which 73.06% are in the lowest quintile of Native American ancestry. Without consideration for admixed and diverse populations, the application of the genome-wide risk score underestimates risk in individuals with more European ancestry and overestimates risk in individuals with more non-European ancestry. Differential performance of the PRS between and within racial/ethnic groups underscores the possibility of exacerbating known health disparities as PRS become translated to clinical care, demonstrating the need for inclusion of diverse populations in their development and evaluation.

A- A+

PgmNr 242: The role of linkage disequilibrium structure in transferability of polygenic risk scores.

Authors:
P. Orozco del Pino; S. Zöellner

View Session Add to Schedule

Affiliation: Biostatistics, University of Michigan, Ann Arbor, MI.

Polygenic risk scores (PRS) are gaining importance in genetics research from statistical genetics methods to diagnosis and treatment of diseases. However, most PRS are built based on European population samples, reducing predictive power in non-European populations and limiting the benefit of medical breakthroughs for these populations. While this reduction in predictive power has been demonstrated multiple times qualitatively, the quantitative impact is challenging to predict. Two drivers for this reduced predictive power can be considered: First, the effect sizes of the causal variants may differ (e.g. due to different environmental risk factors). Second, patterns of linkage disequilibrium (LD) differ between populations and the true risk variant can not be clearly ascertained. Here we quantify the impact of the second driver by estimating the bias from transferring a PRS defined in a European population to a non-European population. As we assume the effect size of the causal variant is constant between populations, we identify the lower bound for the error generated from transferring a PRS. To asses this error, we simulated a discovery study for each variant with minor allele frequency >0.1 (power >0.3) using haplotypes from Europeans in the 1000 Genome Project. If a significant marker near the simulated risk variant was discovered, we estimated the predictive ability of this marker in African, East Asian, South Asian and American populations. Thus we create a map of expected loss of information across the genome. By sampling combinations of risk variants from this map, we predict the overall loss of predictive power. We show that for only 32.7 % of all variants, predictive power is unbiased in the non-European population, resulting in substantial biases in PRS calculate from large marker sets. This shows that differences in LD are sufficient to substantially lower the estimation accuracy of the PRS in the understudied population, even in the absence of effect size heterogeneity.

A- A+

PgmNr 243: Investigating the lack of transferability of polygenic risk scores in cohorts with admixed ancestry.

Authors:
B. Domingues Bitarello; I. Mathieson

View Session Add to Schedule

Affiliation: Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PENNSYLVANIA.

Polygenic risk scores (PRS) can be used to summarize the results of genome-wide association studies (GWAS) into a single number representing the risk of disease. For some traits (for example, cardiovascular disease, breast cancer) PRS allows us to identify individuals with clinically actionable levels of risk in the tails of the PRS distribution. One barrier to the use of PRS in clinical practice is that the majority of GWAS come from cohorts of European ancestry, and predictive power is lower in non-European ancestry cohorts. There are many possible reasons for this decrease; here we investigate the performance of PRS in admixed cohorts to identify some of these reasons.

We focus on the performance of PRS for height (a model polygenic trait) in cohorts with admixed African and European ancestry. Having multiple ancestry components in the same genome allows us to test for ancestry-related differences in PRS prediction while controlling for environmental differences. We first show that that the predictive power of height PRS increases linearly with European ancestry (partial R^2ranges from 0.015-0.15 for 0-100% European ancestry). This effect is unaltered when we re-estimate effects-sizes using sibling pairs, ruling out residual population structure as an explanation. Second, we show that this pattern persists when PRS is computed using subsets of SNPs in regions of both high and low linkage disequilibrium (LD), indicating that differences in LD are not the only cause. Third, we show that frequency differences of associated variants between African and European ancestry backgrounds explain only up to 25% of the observed reduction in predictive power. Finally, we find that there is no association between ancestry and phenotypic variance, indicating that there is no relationship between ancestry and genetic variance, and that the reduction in PRS predictive power cannot be explained by causal variants that are specific to the African ancestry background.

In conclusion, no single factor we investigated can explain the difference in predictive power across ancestries, hinting that other factors – for example heterogeneity in effect size – or a combination of multiple factors is responsible for this pattern. This study further highlights the need for more diversity in GWAS, as well as a better understanding of the complexities of variant discovery and portability across cohorts and ancestries.

A- A+

PgmNr 244: Multi-omic analysis maps the genomic landscape of ovarian cancer to reveal mutational mechanisms and functional pathways that drive chemoresistance.

Authors:
M.R. Jones [1]; S.G. Coetzee [1]; N.M. Gull [1]; K. Dabke [2]; C.A. Kalita [3,4]; A.L.P. Reyes [1]; J.T. Plummer [1]; B.D. Davis [1]; S. Chen [1]; J.P.B. Govindavari [5]; J. Lester [6]; K. Lawrenson [7]; D. Hazelett [1]; A. Gusev [3,4]; S. Parker [8]; B.P. Berman [1,9]; B. Karlan [6]; S.A. Gayther [1,10]

View Session Add to Schedule

Affiliations:
1) Center for Bioinformatics and Functional Genomics, Department of Biomedical Sciences, Cedars Sinai Med Ctr, Los Angeles, CA; 2) Graduate Program in Biomedical Sciences, Department of Biomedical Sciences, Cedars Sinai Med Ctr, Los Angeles, CA; 3) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA; 4) Division of Genetics, Brigham & Women's Hospital, Boston, MA; 5) Dept of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, CA; 6) Department of Obstetrics and Gynecology, David Geffen School of Medicine at UCLA, Los Angeles, CA; 7) Women's Cancer Program, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA; 8) Advanced Clinical Biosystems Research Institute, Smidt Heart Institute, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, 90048; 9) Department of Developmental Biology and Cancer Research, Faculty of Medicine, The Hebrew University of Jerusalem, Ein Kerem; 10) Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA

High Grade Serous Ovarian Cancer (HGSOC) is typically diagnosed at late stage with a 5-year relative survival rate of only 31%. Tumor recurrence occurs in 80% of patients, usually at the same site(s) as the primary cancer, and most patients experience chemo-resistant disease that is fatal. To identify the mutational mechanisms underlying chemoresistance in HGSOC we performed whole genome sequencing (WGS), ultra-deep RNA-Seq, whole genome bisulfate sequencing (WGBS) and proteomics in paired chemotherapy naïve primary and recurrent tumors from 34 patients (12 BRCA1/2 mutation carriers; 24 patients without any identified genetic risk). We generated a genomic landscape of the changes that occur as a result of chemoresistance in each tumor available (n=73). WGS analysis showed the genomes of both primary and recurrent HGSOC tumors to be highly unstable; BRCA1/2 mutant tumors were deficient in homologous repair while tumors from patients without BRCA1/2 mutations carry high numbers of foldback inversions. Somatic SNV, Indel and SV analysis reveal monoclonal and polyclonal seeding of recurrent tumors that are highly related to chemoresistant clones present in the primary tumor that survive the purifying selection process of chemotherapy. Mutational signatures were dominated by BRCA and Age signatures, with little change between primary and recurrent tumors. Two patients showed a unique sub-population of clones that were able to leave the peritoneal cavity and metastasize to the brain, defined by a specific set of gene expression and methylation. Allele specific expression identified gene sets with expression skewed to avoid mutations damaging coding mutations. Combinatorial analysis identify novel networks of transcription factors that preferentially bind hypomethylated enhancers in primary (*ETV5*, *ETV7*, *FLI1*, *SPIC*, *SPI1*) and recurrent (*NFIA*, *NFIC*, *FOXI1*, *NFIB*) with methylation significantly correlated with gene expression of target genes that are significantly enriched in replication fork stability and DNA repair pathways. Using this multi-omic combinatorial approach we have identified novel pathways that are altered post chemotherapy treatment and confer chemoresistance to HGSOC tumors. These tumors show broad resistance to second line therapies currently used in the treatment of HGSOC, however our novel analysis approach has identified potential second line therapies that could provide effective treatment of platinum-resistant HGSOC.

A- A+

PgmNr 245: Prediagnostic serum RNA levels are highly dynamic in lung cancer.

Authors:
R. Lyle [1,9]; S. Umu [2]; H. Langseth [2]; A. Keller [3,4]; E. Meese [5]; A. Helland [6,7,8]; T. Rounge [2]

View Session Add to Schedule

Affiliations:
1) Department of Medical Genetics, Oslo Univ. Hosp, Oslo, Norway; 2) Department of Research, Cancer Registry of Norway, Oslo, Norway; 3) Department of Clinical Bioinformatics, Saarland University, Saarbruecken, Germany; 4) Department of Neurology and Neurological Sciences, School of Medicine, Stanford University, USA; 5) Department of Human Genetics, Saarland University, Homburg/Saar, Germany; 6) Department of Oncology, Oslo University Hospital, Oslo, Norway; 7) Institute for Cancer Research, Oslo University Hospital, Oslo, Norway; 8) Institute of Clinical Medicine, University of Oslo, Oslo, Norway; 9) PharmaTox Strategic Research Initiative, School of Pharmacy, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway

The majority of lung cancer (LC) patients are diagnosed at a late stage and survival is poor. Circulating RNA molecules are known to have a role in cancer, however, their involvement before diagnosis remains an open question. To understand their potential as non-invasive early-detection biomarkers of cancer, analyses of RNAs in prediagnostic samples are required. In this study, we investigated circulating RNA dynamics of serum samples prior to lung cancer (LC) diagnosis to identify if and when disease-related signals are present.

The Janus Serum Bank (JSB) is a population-based cancer research biobank containing 674,386 serum samples from 318,628 Norwegians. By linking JSB data to the Cancer Registry of Norway we identified 542 prediagnostic serum samples from LC patients donated up to 10 years before diagnosis. As controls, we selected 519 serum samples from donors who were cancer-free at least 10 years after sample collection. Cases and controls were frequency-matched on sex, age at donation and blood donor group (BDg). LC histology and stage were retrieved from the Cancer Registry of Norway, and smoking information from health surveys.

We sequenced circulating small RNAs from these samples and analysed the data classified according to histology, stage and time-to-diagnosis. The results showed dynamic changes in circulating RNA levels specific to histology and stage. The greatest number of differentially expressed RNAs were identified around 7 years before diagnosis for early stage LC and 1 to 4 years prior to diagnosis for locally advanced and advanced stage LC, regardless of LC histology. The majority of differentially expressed RNAs were associated with cancer-related pathways, and the dynamic nature of these changes pinpoint different phases of tumor development over time.

In conclusion, we identified highly dynamic RNA LC signals that were time-to-diagnosis, stage and histology dependent. Our results improve the molecular understanding of carcinogenesis and indicate substantial opportunity for screening and improved treatment. However, the dynamic nature of these changes, which identify different phases of tumor development over time, suggest challenges for prediagnostic biomarker discovery.

A- A+

PgmNr 246: Iterative deconvolution around a single gene allows improved accuracy in comparing MSLN high and low PDAC tumors.

Authors:
E. LaPlante; E. Lurie; D. Liu; Q. Yao; A. Milosavljevic

View Session Add to Schedule

Affiliation: Baylor College of Medicine, Houston, TX

The limited efficacy of current standard treatment regimens for pancreatic ductal adenocarcinoma (PDAC) has motivated the search for targeted therapies. One promising therapeutic target is mesothelin (MSLN), a cell surface protein not expressed in the normal pancreas, but multiple clinical trials have failed to show an increase in disease free survival. A lack of basic biological understanding of MSLN function in patient cancer cells may be precluding more effective targeting or combination therapies. Gaining such cancer cell specific insight from patient tumors is complicated as they represent an amalgam of expression from multiple different cell types. To better understand MSLN biology, we attempted to separate samples based on *MSLN* expression to compare differences between *MSLN* high and low tumors to identify its possible roles and interactions. However, classifying TCGA samples based on bulk *MSLN* expression misclassified several samples with low *MSLN* expression as having high expression because of differences in tumor purity. To address the challenge of differences in tumor purity, we developed a novel extension of Epigenomic Deconvolution (EDec), an in silico method that infers cell-type composition and RNAseq profiles from bulk sequencing, to classify samples based on cell-type specific expression rather than bulk signal. By deconvoluting randomly separated groups, then comparing each sample to the cancer cell specific profile and reassigning samples to the group with best fit, we saw accurate separation of *MSLN* high and low cell types, as shown by simulations. We then determined via simulations that comparison of the *MSLN* high and low profiles estimated by EDec was more accurate than using DESeq2 with tumor purity as a covariate on the bulk samples to determine differentially expressed genes. Comparison of the TCGA based EDec cancer cell profiles allowed us to identify 88 genes, which were dysregulated in cancer cells with *MSLN* expression, two of which were then confirmed via western blot and qRT-PCR in *MSLN* high and low cell lines. This study shows deconvolution based around a gene of interest gives more accurate differential expression which allows identification of possible drug targets not discovered via other methods.

A- A+

PgmNr 247: Network-based identification of key master regulators for immunologic constant of rejection in cancer.

Authors:
R. Mall [1]; M. Saad [1]; J. Rolands [2]; W. Hendrickx [2]; M. Ceccarelli [3]; D. Bedgonetti [2]

View Session Add to Schedule

Affiliations:
1) Data Analytics, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Doha, AL Rayyan, Qatar; 2) Department of Immunology, Inflammation and Metabolism, Division of Translational Medicine, Research Branch, Sidra Medicine, Doha, Qatar; 3) Computational Biology, Computational Oncology and Immunology (CIAO), AbbVie Biotherapeutics Inc., Redwood City, California

---

Molecular alterations governing mechanisms leading to immune exclusion are largely unknown. Availability of large-scale biomedical data offers an opportunity to assess the effect of key cellular determinants for an observed phenotype. We develop a network-based approach to identify key transcription factors (TF) associated with poor immunologic responsiveness. A cancer phenotype displaying the coordinated expression of T-helper-1 (Th1) chemokine, interferon, and immune-effector function genes, is associated with favorable prognosis and responsiveness to immunotherapy. This disposition is summarized by a signature that we termed as Immunologic Constant of Rejection (ICR). Based on expression of ICR genes, cancers are classified as immune active (ICR4), or immune silent (ICR1).

We use The Cancer Genome Atlas (TCGA) RNA-seq data for 12 cancer types (~ 2,500 samples) to: (1) build gene regulatory networks via Regularized Gradient Boosting Machines (RGBM), (2) determine each TF's regulon, which are sets of genes regulated by the TF, (3) determine activity matrix of TFs for all samples, and (4) run functional gene set enrichment analysis to identify the top TFs, named Master Regulators (MR), that discriminate ICR1 vs ICR4. MRs such as *L3MBTL1*, *HDAC11*, and *SALL2* were coherently associated with the immune-silent phenotype (ICR1) across 12 cancers. Downstream analysis of MRs specific to ICR1 resulted in identification of NOTCH signaling pathways, chromatin regulation, transcriptional regulation of oncogene TP53, and several cancer-related signaling pathways that can represent novel targets to reprogram the immune suppressive tumor microenvironment. In summary, this is the first report that identified MRs associated with an immune-excluded cancer phenotype.

PgmNr 248: Assembling a *de novo* human genome in 100 minutes.

Authors:
A. Khalak [1]; C.-S. Chin [2]

View Session Add to Schedule

Affiliations:
1) Foundation for Biological Data Science, Belmont, CA, USA.; 2) DNAnexus, Mountain View, CA, USA

De novo genome assembly provides comprehensive, unbiased genomic information and makes it possible to gain insight into new DNA sequences not present in reference genomes. Many de novo human genomes have been published in the last few years, leveraging inexpensive short-read and single-molecule long-read technologies. As technologies improve the scalability of DNA sequencing, the computational burden of generating assemblies persists as a critical factor. The most common approach to long-read assembly, using an overlap-layout-consensus paradigm, requires all-to-all read comparisons, which quadratically scales in computational complexity with the number of reads. Even with various advanced techniques for fast handling of non-overlapped pairs and repeats, the most efficient current methods still require hundreds to thousands of CPU hours. We assert that the current approaches have not optimally addressed computation cost savings possible using characteristics (e.g. accuracy ~99% and read length ~11-15k) of the latest sequencing technologies for accurate long reads.

A novel genome assembler Peregrine is introduced (https://github.com/cschin/Peregrine), which uses a new approach for indexing reads. This approach overcomes the computationally expensive all-to-all read comparison step. Instead, read pairs with high overlap probability are gathered in a single step and compared. Peregrine can assemble 30x human PacBio CCS read datasets in less than 20 CPU hours and around 100 wall-clock minutes to high contiguity assembly (N50 > 20Mb). The continued advance of sequencing technologies in terms of read length and base accuracy coupled with high performance assemblers like Peregrine will enable routine generation of human de novo assemblies. This will allow for more comprehensive representation of the full scope of genomic variations on a population scale -- beyond SNPs and small indels. We evaluate the Peregrine for both computational performance and base accuracy with public available sequencing data for a few human genomes: NA12878, HG002, HG005, and CHM13. For resolving haplotype specific variants, Peregrine can also generate haplotigs in combination with other tools that phase SNPs to partition the read set into haplotype specific groups for separate assembly. We also design and implement a toolset along with Peregrine to enable interactive computation for comparing assemblies to a reference genome for quality control and variation discoveries.

A- A+

PgmNr 249: RaPID: An ultra-fast method for identifying IBDs in biobanks.

Authors:
D. Zhi [1,4]; X. Liu [2]; K. Tang [3]; S. Zhang [3]; A. Naseri [1]

View Session Add to Schedule

Affiliations:
1) School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas; 2) USF Genomics, College of Public Health, University of South Florida, Tampa, Florida; 3) Department of Computer Science, University of Central Florida, Orlando, Florida; 4) Department of Epidemiology, Human Genetics & Environmental Sciences, The University of Texas Health Science Center at Houston, Houston, Texas

The availability of very large cohorts of genotypes presents a challenge to analyze them without extensive computational costs. Finding related individuals in a cohort is one of the essential tasks in genetic studies. A fundamental measure to determine related individuals is Identity by Descent (IBD) segments. IBD segments are identical segments between two relatives that have been inherited from a common ancestor. IBD detection has a wide range of applications such as studying population history or association studies for finding disease-causing markers. The existing tools for detecting IBD segments, however, require extensive resources and/or time to call IBD segments in a large cohort comprising millions of individuals, and it’s almost infeasible to search for IBDs in very large cohorts without extensive resources. We developed a tool called RaPID (Random Projection-based IBD Detection) that is able to detect IBD segments in cohorts containing millions of individuals without requiring extensive resources.

We applied RaPID on UK-Biobank containing genotype data of ~500k individuals from the UK population. RaPID was able to call IBD segments in all autosomal chromosomes with a length more than 10 cM in one day using only a single CPU. Given the low memory consumption RaPID, multiple chromosomes can be searched at the same time which will take only a few minutes to call all IBD segments (with a length ≥ 10 cM) in a large panel such as UK Biobank. The program is also able to handle varying genetic distances across the chromosome which increases the detection power in regions with high recombination rate and also results in more accurate calls in regions with low recombination rate.

RaPID is available freely for academic use at https://github.com/ZhiGroup/RaPID.

A- A+

PgmNr 250: Efficient approximation of GWAS resampling for method validation on massive datasets.

Authors:
N.A. Baya [1]; R.K. Walters [1,2]; J. Bloom [1]; B.M. Neale [1,2,3]

View Session Add to Schedule

Affiliations:
1) Broad Institute of MIT and Harvard, Cambridge, MA; 2) Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA

Rigorous testing of statistical methods in genomic datasets on the scale of large biobanks (>300k individuals) is an essential aspect of evaluating performance. The large scale can robustly identify biases that may not be evident in smaller samples, and empirical evaluation may identify relevant features not present in simpler theoretical or simulation settings. For example, testing statistical methods that use GWAS summary statistics as input may require running linear regression many times on a single population, and may benefit from testing with real patterns of linkage disequilibrium, population structure, and genetic architecture. However it is cost prohibitive to run linear regression on hundreds of thousands of individuals, tens of thousands of times, to evaluate the distribution of estimates and test statistics. We propose a method for efficiently approximating resampling of GWAS in a large genomic dataset and use this approach to evaluate biases in LD Score Regression (LDSC).

Our method partitions individuals in a dataset into equally sized groups and then calculates linear regression summary statistics for each group. To piece together a GWAS of a subset of the dataset, the summary statistics of a certain number of groups are combined using inverse-variance weighted meta-analysis. Our method decreases the marginal cost in CPU-hours of generating a sample GWAS with >100k individuals by more than 98%.

We demonstrate the value of this method applied to validating LDSC heritability and genetic correlation estimation in the UK Biobank (UKB). Empirically, we observe a slight trend towards lower heritability estimates at small sample sizes, and average genetic correlation in male vs. female GWAS of a given trait slightly less than 1. For evaluating heritability, our method allows efficient evaluation of GWAS in downsampled subsets of UKB. For genetic correlation, we can assess whether the genetic correlation is 1 for random binary splits of the UKB, simulating a scenario where we expect perfect genetic correlation. These resampling experiments show support for the existence of the minor biases suggested by the empirical results. We then evaluate the features of the GWAS sample that can affect the magnitude of these biases.

We estimate that our method for approximating resampling of GWAS saves on the order of tens of thousands of dollars of computational cost in the process of evaluating these biases.

A- A+

PgmNr 251: An artificial intelligence approach for nearly instant diagnosis of Mendelian diseases by deep phenotyping and whole-genome or exome sequencing.

Authors:
M. Yandell [1,2]; M. Falconi [3]; B. Moore [1,2]; S. Nohzadeh-Malakshah [3]; E. Frise [3]; E. Kiruluta [3]; M. Reese [3]; F. De La Vega [3]

View Session Add to Schedule

Affiliations:
1) University of Utah, Salt Lake City, UTAH.; 2) Utah Center for Genetic Discovery,Salt Lake City, UTAH.; 3) Fabric Genomics, Oakland, CA 94612, USA.

Sequencing of genomes and exomes is now widely used for clinical diagnoses of Mendelian diseases, for idiopathic disease, and for diagnosis for newborns in NICUs. Pressures remain to increase diagnostic rate while reducing cost. Major cost drivers include the cost of sequencing, the computational complexity of data processing pipelines, and time required for clinical interpretation. Expert interpretation typically consists of iterative filtering coupled with evidence review of candidate variants. We have previously developed two tools designed to be used in tandem to formalize and speed the review process: VAAST – a probabilistic disease-gene finder, and PHEVOR, which leverages the Human Phenotype and Gene ontologies for integration of deep phenotype data with genome sequence information. We have previously demonstrated in a set of 1,963 fully reviewed cases from Genomics England 100K Project that VAAST and PHEVOR rank causative variants within the top-20 for 75% of solved cases. Here we present a novel artificial intelligence (AI) based approach that integrates the outputs of VAAST and PHEVOR, proband and parental genotypes, with knowledge from OMIM, GnomAD, and ClinVar databases to identify disease-causing genes in patient’s genomes. The output includes the predicted disease, inheritance mode, the likely pathogenic variants, and a Bayesian confidence value. We validated our method by analyzing causative variants spiked in a number of control genomes for a variety of disease and modes of inheritance, and for dozens of previously solved clinical cases. We show that our method quickly identifies disease genes and variants responsible for the patient’s phenotype with significantly greater precision than VAAST and PHEVOR; for about 80% of the clinical test cases the output includes only the correct gene and no others, while in the remaining cases the correct gene is among 2-3 candidates offered. Our method still retains the ability of VAAST and PHEVOR to identify novel disease genes, but significantly reduces the time spent on reviewing candidate causative genes and variants. In summary, our AI-method provides a significant performance improvement over our already proven VAAST and PHEVOR algorithms reducing interpretation turnaround times to nearly instant diagnosis. We are integrating this tool in a cloud-platform for clinical genome interpretation which provides a rich set of annotations and tools for clinical reporting.

A- A+

PgmNr 252: Single-cell transcriptomics reconstructs developmental tree from embryonic stem cell to insulin producing cells.

Authors:
C. Weng; J. Xi; H. Li; Y. Li; F. Jin

View Session Add to Schedule

Affiliation: Case Western Reserve University, Cleveland, Ohio.

In vitro differentiation from human embryonic stem cell (hESCs) into insulin producing cells offers new opportunities for disease modeling and potential diabetes therapy. However, the precise molecular changes during the differentiation process remain unclear, leaving it a long-standing challenge to fully understand β cell lineage specification, and thus poses a substantial obstruction to further optimizing differentiation strategy. Here, we analyzed 87,769 single cell transcriptome data at 12 time points across 7 differentiation stages from hESCs to insulin producing cells. We identified multiple cell populations in most of stages. Particularly, we highlighted a distinct cell population (β-like) in the last stage that is closely, but still not fully, resemble the bond fide β cells from human donor. We developed a time-course-supervised lineage reconstruction algorithm TreCCA, by which we built the whole developmental tree that contains a considerable number of side lineages in addition to the main lineage that differentiate into β-like cells. Along the tree, we identified 64 gene modules that are highly variable and show different lineage specificities. Cross-lineage analysis exhibits gene signatures to distinguish endocrine lineage from non-endocrines, and also β-like lineage from α-like. By removing all side lineages, we specifically characterized the transformation of transcriptome landscape into β-like lineage. Interestingly, we identified a number of genes, including 117 transcription factors, showing double-wave expression changes rising and falling at different time, implying temporal dependent transcription factor reuse and potential gene dual functions during differentiation. Overall, we demonstrated the power of using differentiation-timeline-supervised single cell transcriptome for lineage tree reconstruction, providing a blueprint to better understand the cell fate decisions for β cell maturing *in vitro*.

A- A+

PgmNr 253: Characterizing cellular communication in human central nervous system by single-cell RNA sequencing.

Authors:
Z. Miao [1,2]; G. Hu [2]; K. Wang [2]; A. Nguyen [3]; Y. Lyu [2]; J. Lakkis [2]; C. Strang [4]; C. Curcio [5]; D. Stambolian [6]; E. Lee [3]; M. Li [2]

View Session Add to Schedule

Affiliations:
1) Graduate Group in Genomics and Computational Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104; 2) Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104; 3) Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104; 4) Department of Psychology, University of Alabama at Birmingham, Birmingham, AL 35294; 5) Department of Ophthalmology and Human Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104; 6) Department of Ophthalmology and Visual Sciences, University of Alabama at Birmingham, Birmingham, AL 35294

In the Central Nervous System (CNS), cell-cell communication includes transmission of signals and responses to stimulation by neurons and glial cells. However, in disease states such as Alzheimer's Disease (AD), cell-cell communication is disrupted, which may lead to cognitive deficits and memory impairments. To study differences of cell-cell communication between healthy and AD brains, we conducted computational analysis using single-nucleus RNA-seq (snRNA-seq) data from 15 brains with or without AD. In this dataset, 29,869 cells from healthy brains and 116,568 cells from AD brains were classified as five broadly defined cell types including microglia, astrocytes, GABAergic neurons, glutamatergic neurons, and oligodendrocytes. To quantify cell-cell communication between cell types, we focused on the ligand-receptor type of interactions. For each known ligand-receptor pair, we defined a "communication score" between two cell types as the product of the percentage of a ligand gene being expressed in one cell type and the percentage of the corresponding receptor gene being expressed in another cell type. To assess statistical significance, the communication scores were compared with scores generated by randomly re-assigning cell type labels to each cell. Our results indicate that cellular communication patterns are dynamic and susceptible to cellular state and genetic background. For example, in AD patients with TREM2 R47H mutation, cell-cell communication within the microglia sub-populations are globally enhanced compared to those without this mutation. To further assess the reliability of our method, we analysed a single-cell RNA-seq (scRNA-seq) dataset generated from human retina obtained from donors without AD or other neurological disorders. The retina is another CNS tissue with well-characterized cell types that can be affected by AD. The same analysis was conducted on the retina dataset that includes 92,385 cells comprising 11 major cell types. We found cell types that are spatially close to each other tend to have more co-expressed ligand-receptor pairs, indicating the validity of our method. Taken together, we have developed a computational approach to quantify cellular communication using scRNA-seq (or snRNA-seq) data and applied this method to the analysis of two human CNS datasets. We found that cell-cell communication patterns are dynamic and are related to genetic background, cellular states, as well as spatial locations of the cells.

A- A+

PgmNr 254: Reconstructing pseudotemporal trajectories in single-cell RNA-seq data using neural networks.

Authors:
J. Lakkis [1]; G. Hu [2]; H. Zhang [3]; C. Xue [3]; M. Reilly [3]; M. Li [1]

View Session Add to Schedule

Affiliations:
1) Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania.; 2) Information Theory and Data Science, Nankai University, Tianjin, China; 3) Division of Cardiology, Columbia University, New York, New York

Single-cell RNA-seq (scRNA-seq) profiling can quantify transcriptional dynamics in temporal processes, such as cell differentiation, using computational methods to label each cell with a “pseudo-time.” Pseudotemporal analysis models the gradual transition of a cell rather than assigning it to a discrete cluster. It is important to identify genes that are differentially expressed over pseudo-time by disease condition. If trajectory reconstruction is driven by a biological process, then these genes are potential targets for therapy. However, most popular pseudotemporal trajectory reconstruction methods such as Monocle 3 and TSCAN are unsupervised. In such cases, it is unclear whether factors driving trajectory construction are biologically meaningful. To overcome these limitations, we present PseudoNet, a supervised approach for pseudo-time reconstruction to model a non-branching temporal cell process between two nodes (e.g. cell types). PseudoNet is a targeted pseudo-time reconstruction method in which the user can incorporate prior knowledge about the underlying process by defining cells that are in the start and end nodes. A feedforward neural network is trained with these specified cells to predict the probability that a cell is in the end node of the temporal process based on its gene expression, and the predicted probability is used as a proxy for pseudo-time. To evaluate the performance of PseudoNet, we analyzed 7 benchmark datasets for cell cycle and cell differentiation, where true cell states and real times are known. We show that PseudoNet orders cells with higher resolution than Monocle 3 and TSCAN in each of the 7 datasets. We further applied this method to a scRNA-seq dataset on monocytes for 16 human subjects, totaling ~60,000 single cells and show that PseudoNet identifies genes differentially expressed between healthy subjects and those with cardiometabolic disease risk factors. Lastly, we extended PseudoNet to reconstruct trajectories that allow branching. This extension is built on top of an unsupervised deep embedding algorithm for scRNA-seq clustering. It first finds high confidence cell clusters, and then leverages PseudoNet to order cells between pairs of nodes. This method requires minimal assumptions and is robust to the underlying trajectory structures. As the growth of single-cell studies increases, we expect our methods will substantially improve the accuracy in pseudotemporal reconstruction of single cells.

A- A+

PgmNr 255: Integrative single-cell and bulk RNA-seq analysis in human retina identified cell type-specific composition and gene expression changes for age-related macular degeneration.

Authors:
Y. Lyu [1]; R. Zauhar [2]; N. Dana [3]; C. Strang [4]; K. Wang [1]; Z. Miao [1]; P. Gamlin [5]; C. Curcio [5]; D. Stambolian [3]; M. Li [1]

View Session Add to Schedule

Affiliations:
1) Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 2) Department of Chemistry and Biochemistry, The University of the Sciences in Philadelphia, Philadelphia, PA; 3) Dept of Ophthalmology and Human Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 4) Department of Psychology, University of Alabama at Birmingham, Birmingham, AL; 5) Department of Ophthalmology and Visual Sciences, University of Alabama at Birmingham, Birmingham, AL

Age-related macular degeneration (AMD) is a leading cause of central vision loss among elderly. Clinical, epidemiologic and pathology studies suggest that AMD preferentially affects distinct cell types and topographic regions in retina. To characterize the impact of AMD on gene expression changes across retinal cell types and regions, we conducted integrative analysis of single-cell RNA-seq (scRNA-seq) data from 92,385 cells generated from 2 donors, and bulk RNA-seq data from another set of 15 donors in macular and peripheral retina. The scRNA-seq data revealed 11 major cell types. Among 75 previously reported AMD risk genes, 29 (38.7%) show cell type- and/or region-specific expression patterns. For example, *CFH* is specifically expressed in endothelium cells, whereas *VTN* and *MMP9* are specifically expressed in photoreceptor cells, and *ABHD12B* and *TRPM3* are preferably expressed in Müller cells particularly for macular retina. To understand the impact of AMD on cell-type composition in retina, we performed cell-type deconvolution analysis in the bulk RNA-seq data using the scRNA-seq data as a reference. Notably, rods constitute 70% of detectable cells in normal macula , but were barely detectable in late AMD macula. Ganglion cells were undetectable in all samples. Our results highlight pronounced changes in cell type composition as AMD progresses. Significant changes were loss of rod photoreceptors in both regions especially macula, and increase of microglia and endothelial cells in macula only. To investigate the cell-type-specific response to AMD, we developed a calibration-based method, which allowed us to identify 1,158 AMD associated genes that are DE only in specific cell types. Among these genes, 126 are specific to rods with 41 up-regulated and 85 down-regulated. Interestingly, the down-regulated genes (e.g. *IMPG2*, *RP1* and *PED6*), are enriched in visual perception, sensory perception of light stimulus and detection of light stimulus, whereas those up-regulated genes (e.g. *EPS8*, *YBX1* and *FBXO32*), are enriched for negative regulation of cell death, cellular response to oxidative stress and protein refolding. Taken together, our results reveal changes in cell type composition and gene expression in the macula that are absent in peripheral retina. Our study also provides novel methods for integrating scRNA-seq data with bulk RNA-seq data to elucidate the molecular mechanisms in whole tissue.

A- A+

PgmNr 256: Identifying the role of genetics and neurodevelopment in sporadic late onset Parkinson’s disease.

Authors:
S. Kumar [1]; J.E. Curran [1]; J.M. Peralta [1]; A.C. Leandro [1]; D.M. Lehman [2]; D.C. Glahn [3,4]; J. Blangero [1]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics and South Texas Diabetes and Obesity Inst. University of Texas Rio Grande Valley School of Medicine, Edinburg and Brownsville, TX, USA; 2) Department of Medicine, University of Texas Health Science Center, San Antonio, TX, USA; 3) Department of Psychiatry, Boston Children’s Hospital and Harvard Medical School, Boston, MA, 02115, USA; 4) Olin Neuropsychiatry Research Center, Institute of Living, Hartford Hospital, Hartford, CT, 06106, USA

Parkinson's disease (PD) is the second most common age associated neurodegenerative disorder. The defining PD motor symptoms only appear after 50-60% of the substantia nigra dopaminergic (DA) neurons are lost. However, the disease often precedes motor symptoms by years or even by decades. Unfortunately, in most sporadic cases (which account for > 90% of the total burden of the disease) diagnosis is often late, and disease etiology is multifactorial. Based on the remarkable similarity between PD’s early nonmotor symptoms and those observed in impaired neurogenesis and several of the PD associated genes also playing role(s) in embryonic or adult neurogenesis, a significant neurodevelopmental component to the PD has been hypothesized. In our Genetics of Brain Structure and Function Study (GOBS) pedigree-based cohort, we have identified two sporadic late onset PD cases, with iPSC generated DA neurons exhibiting a high accumulation of α-synuclein and phosphorylated α-synuclein, consistent with PD. To identify genes that may influence sporadic PD and/or early neurodevelopment, we performed genome-wide RNA sequencing analysis of iPSC derived neural stem cells (NSCs) of these 2 PD cases and their 10 currently unaffected relatives, who also have participated in GOBS. Differential gene expression (DGE) analysis identified 193 genes that were significantly differentially expressed (DE) between the PD cases and their unaffected blood relatives (moderated *t* statistics *p*-value ≤ 0.05, fold change (FC) absolute ≥ 4.0). Functional enrichment analysis showed significantly high enrichment of DE genes in neurological diseases (*p*-value range $4.4x10^{-2}$ to $5.3x10^{-5}$) and nervous system development and function (*p*-value range $4.5x10^{-2}$ to $4.5x10^{-8}$). Fifteen of these DE genes have been previously reported as associated with PD. To identify the underlying cause of these gene expression variations, we performed a systematic analysis of the genes involved in neuro-developmental processes and identified an expression phenotype (i.e. highly upregulated expression of *HES3* and *HES5* genes and nominal downregulation of *HES1* gene), which is consistent with *HES1* gene dysfunction. The *HES* genes are critical for the development of the nervous system and for adult neurogenesis. A further analysis using iPSC based *in-vitro* modeling shows that this expression phenotype affects NSC maintenance, neural crust cells development, and the differentiation and maintenance of DA neuron.

A- A+

PgmNr 257: Parkinson's disease derived monocytes show alteration in the phago-lysosomal pathway.

Authors:
E. Udine [1,2]; E. Navarro [1,2]; M. Parks [1,2]; G. Riboldi [4]; K. Lopes [1,2]; B. Schilder [1,2]; T. Sikder [1,3]; K. Watkins [1,2]; M. Zhang [1,2]; D. Raymond [5]; S. Elango [5]; E. Wieder [5]; S. Simon [5]; S. Bressman [5]; J. Crary [1,3]; S. Frucht [4]; R. Saunders-Pullman [5]; T. Raj [1,2]

View Session Add to Schedule

Affiliations:
1) Ronald M. Loeb Center for Alzheimer’s Disease, Nash Family Department of Neuroscience, Icahn School of Medicine at Mt. Sinai Hospital, New York, NY; 2) Department of Genetics and Genomics, Icahn School of Medicine at Mt. Sinai Hospital, New York, NY; 3) Department of Pathology, Icahn School of Medicine at Mt. Sinai Hospital, New York, NY; 4) The Marlene and Paolo Fresco Institute for Parkinson’s and Movement Disorders, NYU Langone Health, New York, NY; 5) Mount Sinai Beth Israel, New York, NY

Parkinson's disease (PD) is a neurodegenerative disorder that causes motor and cognitive impairment. Human genetic studies have identified over 90 loci and have suggested that several innate immune genes may have a role in PD. We hypothesize that PD susceptibility genes modulate disease risk by dysregulation of gene expression networks and cellular function within the innate immune cells (central nervous system microglia and peripheral monocytes). Our aims are to (i) understand causal gene networks in PD monocytes; (ii) identify common genetic variants that may alter gene expression; and (iii) perform functional validation of the altered pathways. Our previous transcriptome-wide association study of PD prioritized genes in peripheral monocytes (Li, Wong et al. 2019). Here, we have generated independent genomic and transcriptomic profiles using RNAseq from peripheral monocytes of 120 PD cases and 110 age-matched controls from a clinically well-characterized cohort. We have also generated single-cell RNAseq (sc-RNAseq) and performed functional studies (including stimulation with LPS and IFN) in a subset of samples. Using this dataset, we performed integrative analysis using PD phenotypes, expression QTL, network analysis, and functional validation. Transcriptomic analysis from this cohort shows dysregulation of gene expression, highlighting the role of the innate immune system in PD. Pathway analysis shows significant enrichment for genes associated with lysosomal, mitochondrial, and phagocytosis function. A large proportion (~8-10%) of PD heritability is explained by genes in the lysosomal co-expression network. We report evidence that some disease-associated common variants affect the expression of nearby genes or *cis*-eQTLs (including *LRRK2, CTSB, GPNMB, TMEM175, BST1,* and *SNCA*). We have validated these findings via sc-RNAseq and we have shown that PD macrophages have a less acidic lysosomal pH in comparison to controls. Overall, we demonstrate that there is gene expression dysregulation in peripheral monocytes from PD patients, some of the genes are genetically regulated, and our findings point to the phago-lysosomal pathway as a key driver of the disease.

A- A+

PgmNr 258: Compound heterozygous mutations in *ATP10B* compromising lysosomal glucosylceramide flippase activity are associated with Parkinson’s disease.

Authors:
S. Smolders [1,2,3]; S. Martin [4]; C. Van den Haute [5,6]; B. Heeman [1,2,3]; S. van Veen [4]; D. Crosiers [1,2,7]; I. Beletchi [4]; A. Verstraeten [1,2,3]; G. Gelders [5]; P. Pals [2,7]; N. Hamouda [4]; S. Engelborghs [2,8]; J.J. Martin [2]; J. Eggermont [4]; P.P. De Deyn [2,8]; P. Cras [2,7]; V. Baekelandt [5,6]; P. Vangheluwe [4]; C. Van Broeckhoven [1,2,3]; the BELNEU consortium

View Session Add to Schedule

Affiliations:
1) Center for Molecular Neurology, VIB, Antwerp, Belgium; 2) Institute Born-Bunge, Antwerp, Belgium; 3) Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium; 4) Laboratory of Cellular Transport Systems, Department of Cellular and Molecular Medicine, KU Leuven, Leuven, Belgium; 5) Laboratory for Neurobiology and Gene Therapy, KU Leuven, Leuven, Belgium; 6) Leuven Viral Vector Core, KU Leuven, Leuven, Belgium; 7) Department of Neurology, Antwerp University Hospital, Edegem, Belgium; 8) Department of Neurology and Memory Clinic, Antwerp Hospital Network, General Hospitals Middelheim and Hoge Beuken, Antwerp, Belgium

---

Background Parkinson’s disease (PD) causal genes and risk factors provided valuable insights into the underlying disease mechanisms and delivered new therapeutic targets. Together, causal and risk genes represent only a small fraction of the genetic etiology of PD, leaving both PD families and sporadic PD patients genetically unexplained.
Methods We performed whole exome sequencing (WES) in 53 unrelated early-onset PD (EOPD) patients (AAO ≤ 50 years) to identify novel genes for PD; whole genome sequencing (WGS) in one EOPD patient-unaffected parents trio; targeted resequencing in 618 PD patients (mean AAO 59.9 ± 11.8 years) and 597 control individuals (mean AAI 70.4 ± 8.0 years); qPCR on mRNA isolated from human brain tissue; and functional assays of wildtype and mutants in *in vitro* and cellular models.
Results In three EOPD patients with unaffected parents, we identified *trans* compound heterozygous mutations in the ATPase class V type 10B gene (*ATP10B*), compatible with recessive inheritance. WGS in one EOPD family confirmed *ATP10B* as the disease cause with no other mutations in PD associated genes. Targeted resequencing of *ATP10B* revealed three additional sporadic PD patient carriers of compound heterozygous mutations. *ATP10B* mRNA is enriched brain, specifically in the *substantia nigra* and *medulla oblongata*, and is significantly decreased in these brain regions in patients *versus* control individuals. We further established that functional ATP10B encodes for a late endo-/lysosomal lipid flippase responsible for the translocation of glucosylceramide and phosphatidylcholine from the exoplasmic to the cytosolic membrane leaflet. Disease-associated ATP10B mutants are catalytically inactive and fail to provide cellular protection against rotenone and manganese, two PD environmental risk factors. In isolated cortical neurons, loss of ATP10B leads to lysosomal dysfunction and cell death.
Discussion We identified loss of function recessive mutations in *ATP10B* increasing risk for PD by disturbed lysosomal export of glucosylceramide. Lysosomal functionality and integrity is well known to be implicated in PD pathology and linked to multiple causal PD genes and genetic risk factors. Strikingly, both ATP10B and the well-known PD risk factor GBA play essential roles in the fate of

lysosomal glucosylceramide, and dysfunction of both results in intra-lysosomal accumulation of glucosylceramide.

A- A+

PgmNr 259: An integrated genomic approach to dissect the genetic landscape regulating the cell-to-cell transfer of a-synuclein.

Authors:
E. Kara [1]; A. Crimi [1]; M. Emmenegger [1]; C. Manzoni [2,3]; S. Bandres Ciga [4]; J. Botia [3,5]; A. Wiedmer [1]; M. Carta [1]; D. Heinzer [1]; M. Avar [1]; A. Chincisan [1]; C. Blauwendraat [4]; S. Garcia Ruiz [1]; D. Pease [1]; L. Mottier [1]; A. Carrella [1]; D. Schneider [1]; A. Magalhaes [1]; C. Aemisegger [6]; Z. Fan [7]; J. Marks [7,8]; S. Hopp [9]; P. Lewis [2,3]; M. Nalls [4]; M. Ryten [3]; J. Hardy [3]; B. Hyman [7]; A. Aguzzi [1]

View Session Add to Schedule

Affiliations:
1) Institute of Neuropathology, University of Zurich, Zurich, Switzerland; 2) School of Pharmacy, University of Reading, London, United Kingdom; 3) Department of Neurodegenerative disease, University College London, London, United Kingdom; 4) Laboratory of Neurogenetics, National Institutes of Health, Bethesda, United States; 5) Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Murcia, Spain; 6) Center for Microscopy and Image Analysis , University of Zurich, Zurich, Switzerland; 7) Department of Neurology, Harvard Medical School, Boston, United States; 8) Mayo Medical School, Rochester, Minnesota; 9) Department of Pharmacology, UT Health San Antonio, San Antonio, United States

Objectives: It is thought that a-synuclein pathology in Parkinson’s disease (PD) spreads throughout the brain by neuron-to-neuron transfer of the misfolded protein. Identification of genetic modifiers of this process would advance the understanding of the pathogenesis of PD.

Methods: We have cloned a construct encoding GFP-2a-synuclein-RFP. In a transient transfection tissue culture system, the cells that have been transfected are positive for GFP and RFP fluorescence, whereas cells that have taken up a-synuclein through transfer are positive only for RFP fluorescence. A commercially available library (ThermoFisher) containing 64,752 siRNAs targeting 21,584 genes was used in a 384 well pooled arrayed format for the genome wide screen which was undertaken on a HEK cell line stably overexpressing a-synuclein. High content imaging was undertaken on a GE InCell analyzer 2500HS. The cell-to-cell transfer ratio was calculated by dividing the number of RFP+GFP- cells to the number of GFP+ cells.

Results: The top 1,000 genes, with a p-value cut-off of 6.8*10-4 were selected for validation through a secondary screen in which single siRNAs for each gene were assessed in technical triplicates. 152 genes were confirmed to modulated a-synuclein transfer. Subsequently, the pooled version of the screen was repeated independently 3 times for those 152 genes plus 80 randomly selected genes, and the results were overlaid with those of the RNA sequencing on the cell line used for the screen. 38 genes passed the Bonferroni corrected p-value of 0.05 in all 3 screens, exhibited the same effect directionality in all experiments, and were expressed in the cell line used for the screen. 36 (95%) of the hits were expressed in human brain according to the GTEx dataset, 27 (71%) clustered within the same modules as known PD Mendelian genes and risk factors according to WGCNA analysis of GTEx data, and 4 genes (11%) were located within 500kb of SNPs identified as significantly associated with PD in the latest PD GWAS metaanalysis. Functional enrichment analysis through g-profiler showed that the hits were implicated in cell cycle regulation, intracellular organization, protein metabolism and waste disposal. Assessment of the cumulative effect of rare variants through imputation of PD GWAS data from 45,866 individuals within the identified genes grouped by WGCNA modules is currently underway.

Conclusions: We identified numerous novel genes regulating propagation of a-synuclein.

A- A+

PgmNr 260: *Titin*, the most challenging human gene, requires a multidisciplinary extensive approach.

Authors:
M. Savarese [1,2]; M. Johari [1,2]; A. Vihola [1,2,3]; P.H. Jonson [1,2]; H. Luque [1,2]; T. Qureshi [1,2]; S. Välipakka [1,2]; S. Koivunen [1,2]; M. Arumilli [1,2]; J. Sarparanta [1,2]; P. Hackman [1,2]; B. Udd [1,2,3,4]

View Session Add to Schedule

Affiliations:
1) Folkhälsan Research Center, Helsinki, Finland; 2) Department of Medical Genetics, Medicum, University of Helsinki, Helsinki, Finland; 3) Neuromuscular Research Center, Department of Genetics, Tampere, Finland; 4) Department of Neurology, Vaasa Central Hospital, Vaasa, Finland

---

Few other genes have benefited from the massive use of NGS technologies more than human titin gene *(TTN),* containing 364 exons and encoding a giant protein that acts as a backbone of the sarcomeric structure and a molecular spring determining elasticity of the muscle.
The first *TTN* mutation responsible of a muscle disease (Tibial muscular dystrophy) was identified in the Finnish population back in 2002 and few other mutations had been identified until the NGS development. In the last few years, an increasing number of ultra-rare variants have been detected resulting in a wide spectrum of *TTN*-related diseases characterized by a different age of onset, progression and muscle involvement. Titin truncating variants, for example, have been associated to dilated cardiomyopathy (DCM), a cardiac disease affecting approximately 1 in 250 individuals
Because of its sheer size, its repetitive modular structure, its high number of different tissue- and developmental stage- specific splicing isoforms, *TTN* is one of the most significant challenges related to NGS investigation in the field of medical genetics.
For an improved approach to the study of *TTN* and titinopathies, we set up an exhaustive approach. By a systematic analysis of rare variants in a large multi-center cohort of patients with muscle disorders (over 2500) and in publicly available databases, we identified possible improvements for the current variant interpretation guidelines. We also evaluated the exon usage in anatomically different human muscle tissues and in different developmental stages, refining the *TTN* splicing pattern. At the same time, we refined the current genotype-phenotype correlation, identifying a clear correlation between the location of carboxy terminal truncating variants and the severity of the disease.
Our experience suggests that a collaborative effort, involving different specialists working on titin, is essential for an improved interpretation of *TTN* variants.

A- A+

PgmNr 261: Protein function-specific structural insights into the effect of Mendelian disease variants in 1,330 human genes.

Authors:
S. Iqbal [1,2,3]; E. Perez-Palma [4]; J. Jespersen [5,6]; P. May [7]; D. Hoksza [7]; H. Heyne [3,8]; S. Ahmed [9]; Z. Rifat [9]; S. Rahman [9]; K. Lage [5,6]; A. Palotie [1,2,8]; J. Cottrell [1]; F. Wagner [1]; M. Daly [1,2,3,8]; A. Campbell [1]; D. Lal [1,4,10,11]

View Session Add to Schedule

Affiliations:
1) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 2) Medical and Population Genetics program, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 3) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 4) Cologne Center for Genomics, University of Cologne, Cologne, Germany; 5) Department of Surgery, Massachusetts General Hospital, Boston, MA, USA; 6) Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark; 7) Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg; 8) Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland; 9) Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, ECE Building, West Palashi, Dhaka-1205, Bangladesh; 10) Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, USA; 11) Genomic Medicine Institute, Lerner Research Institute Cleveland Clinic, Cleveland, US

Amino-acid substitution due to a single missense variant can cause a Mendelian phenotype. However, not every variant is pathogenic. For every Mendelian disorder-associated gene, both disease-causing pathogenic and benign missense variants have been reported. Current machine learning based variant classification algorithms are approaching a good level of accuracy. However, these methods do not provide insights into the molecular pathology, and therefore, are not bridging the gap between genomic variation, molecular consequence on protein function and drug development. Only a few studies have applied data-driven approaches to identify protein features of pathogenic missense variants. Due to the challenge that missense variants cannot yet be mapped at scale on protein 3D structure, features of missense variants related to protein structure, function, and dynamics remain underexplored in the interpretation of disease etiology associated with missense variants.

Here we describe the aggregation and analysis of 460,586 missense variants in 1,330 disease-associated genes (>80% annotated in OMIM) on >14000 molecularly-solved human protein structures. Comparing the burden (Fisher’s Exact test) of missense variants from the general population (gnomAD database, N=164,915) and patients (ClinVar and HGMD databases, N=32,924) on 40 structural, physicochemical, and functional protein features of amino-acids with 3D coordinates, we found 18 and 14 features that are significantly associated to pathogenic and population variants across all 1,330 genes, respectively. Additionally, we grouped 1,330 genes into 24 protein functional classes and performed separate class-specific analyses, revealing 240 and 185 function-specific pathogenic and population variant-associated features, respectively. We validated the identified benign and pathogenic features using independent missense variant set as well as functional read-out data from high throughput saturation mutagenesis experiments.

We then developed an online resource, Missense variant to proteIn StruCture Analysis web SuiTe (MISCAST, http://miscast-beta.broadinstitute.org/), for interactive exploration of features and missense variants on 1D and 3D protein structures (freely downloadable for offline analysis). Our results and MISCAST can serve as a powerful resource for translation of genomics to medicine: can help in variant interpretation and selection for functional assay, and in formulating hypotheses for drug development.

A- A+

PgmNr 262: *NOTCH3* cysteine altering variants among the 92,456 whole exome sequenced participants of the Geisinger DiscovEHR initiative.

Authors:
R.J. Hack [1]; N. Pearson [3]; J.W. Rutten [1]; J. Li [3]; A. Khan [2]; M.A. Iqbal [2]; J. Hornak [2]; V. Abedi [3]; Y. Zhang [4]; M.T.M. Lee [4]; C. Griessenauer [2]; . Regeneron Genetics Center [5]; S.A.J. Lesnik Oberstein [1]; R. Zand [2]

View Session Add to Schedule

Affiliations:
1) Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands; 2) Neuroscience institute, Geisinger Health System, Danville, Pennsylvania, USA; 3) Biomedical and Translational Informatics Institute, Geisinger, Danville, Pennsylvania, USA; 4) Geisinger Genomic Medicine Institute, Danville, Pennsylvania, USA; 5) Regeneron Genetics Center, Tarrytown, New York, USA

Background: Cysteine altering missense variants in *NOTCH3* have been exclusively associated with CADASIL, a hereditary cerebral small vessel disease (SVD). Recently, we discovered that these variants, especially those located in the NOTCH3 protein’s epidermal growth factor-like repeat (EGFr) domains 7-34, have an unexpectedly high frequency in the general population worldwide, namely 1:300. The goal of this study was to investigate the phenotype of participants with *NOTCH3* cysteine altering variants in the Geisinger DiscovEHR initiative.

Methods: We utilized whole exome sequence data from 92,456 Geisinger-Regeneron DiscovEHR participants. We selected individuals with a cysteine altering mutation in one of the NOTCH3 EGFr domains. The control group, matched for age and sex, had only synonymous variants in *NOTCH3*. We reviewed and recorded all the patients’ demographic and clinical information, as well as neuroimaging characteristics. Group comparisons were done using the χ^2 test for categorical variables, and unpaired t-test for normally distributed continuous variables. SPSS 24.0 was used for all statistical analyses.

Results: We identified 135 individuals with a *NOTCH3* cysteine altering missense variant (frequency 1:685), of which 134 had a variant in one of EGFr domains 7-34. Clinical records were available for 118 cases, with a mean age of 58.1 ± 16.9 years. The control group consisted of 184 individuals, mean age 57.9 years ± 16.6. In the case group, 12.7% had a history of stroke, compared to 4.9% of controls (p=0.014). Age at onset of stroke did not differ between cases and controls: 67,8 years ± 20.4 vs 65,0 years ± 11.72. There was no significant difference in the frequency of migraine headache, depression, and dementia. Twenty-nine (25%) cases and 45 (24%) controls had an interpretable MRI. Compared to controls, cases more frequently had large areas of confluent white matter hyperintensities (Fazekas 3) (p=0.012) and lacune count was significantly higher (p=0.042).

Conclusion: *NOTCH3* cysteine altering variants in the population, almost exclusively located in EGFr domains 7-34, are associated with an increased prevalence of stroke and SVD markers on MRI. This suggests that NOTCH3 EGFr 7-34 variants are a novel risk factor for SVD in the general population, but are not associated with a classical CADASIL phenotype.

Study supported by: Regeneron, Geisinger Health System, ZonMw

Disclosures: None

A- A+

PgmNr 263: Identification of functionally essential sites using 3D single protein and multiprotein complexes across 73 neurodevelopmental disorder-associated genes.

Authors:
T. Brünger [1]; S. Iqbal [2,3]; E. Perez-Palma [1]; M.J. Daly [2,3,4]; A.J. Campbell [2]; P. May [5]; D. Lal [1,2,6,7]

View Session Add to Schedule

Affiliations:
1) Cologne Center for Genomics (CCG), University of Cologne, Cologne, 50931, Germany; 2) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 3) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 4) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, FI-00014, Finland; 5) Luxembourg Centre for Systems Biomedicine, University Luxembourg, Luxembourg; 6) Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44106, USA; 7) Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44106, USA

Interpretation of missense variants is challenging in the phenotypically and genetically heterogeneous group of neurodevelopmental disorders (NDD). Recent large-scale genomic screens have identified >100 missense variant intolerant NDD-genes. Interpretation of variants on protein structure represents an opportunity to gain mechanistic insights into the molecular pathology. However, only a few genetic variants have been mapped on NDD-genes so far. For most NDD-genes, human protein structures are not available, subsequently, a systematic screen to identify essential sites in NDD-proteins and multiprotein complexes has not been performed.

Here we systematically collected molecular-solved, modeled and predicted protein structures for 73 NDD-associated genes, which are intolerant for missense variants. To explore the utility of these structures and to identify essential sites within proteins, we normalized the linear sequenced based missense variant constrained score (MTR) and an amino acid level paralog conservation score using 12 Å distance spatial windows for each of the 73 protein structures. The corresponding 3D constrained and 3D paralog conserved protein sites show a higher burden for ClinVar and HGMD ascertained pathogenic variants compared to the conserved sites defined by the (original) linear scores. Next, we mapped pathogenic variants (ClinVar and HGMD databases) and population variants (gnomAD database) onto 3D structure and identified pathogenic-variant-enriched amino acids (3D-hotspots) for which no patient variant has yet been reported. Across 73 genes, we identified 192 3D-hotspots for pathogenic variants (mean: 2.7 ± 9.3 per NDD-gene) without a reported variant. For a subset of NDD-proteins, we were able to perform the spatial enrichment analyses for single proteins and multiprotein complexes and observed an increased number of identified 3D-hotspots in the complexes.

In summary, we present the first large scale human protein structure-based analysis of missense variants in NDD-genes. We show that incorporation of solved and predicted human protein structures, as well as multiprotein complexes, represent a useful tool for variant interpretation.

A- A+

PgmNr 264: Million Veteran Program Return Of Actionable Results - Familial Hypercholesterolemia (MVP-ROAR-FH) Study: Considerations for variant return to mega-biobank participants.

Authors:
J.L. Vassy [1,2,3]; N. Alexander [1]; T. Assimes [4,5]; C.A. Brunette [1]; T.E. Callis [6]; K.D. Christensen [2,3]; M. Danowski [1]; Q. Hui [7]; J.W. Knowles [5]; V.A. Morrison [8]; A.C. Sturm [9]; Y. Sun [7]; M. Vatta [6]; V.L. Venne [8]; for the MVP-ROAR-FH Study

View Session Add to Schedule

Affiliations:
1) VA Boston Healthcare System, Boston, MA.; 2) Harvard Medical School, Boston, MA; 3) Brigham and Women's Hospital, Boston, MA; 4) VA Palo Alto Health Care System, Palo Alto, CA; 5) Stanford University, Stanford, CA; 6) Invitae, San Francisco, CA; 7) Emory University, Atlanta, GA; 8) VA Salt Lake City Health Care System, Salt Lake City, UT; 9) Geisinger Health, Danville, PA

Background

Professional consensus is emerging around disease-gene curation, genetic variant interpretation, and unanticipated secondary genetic findings to be considered for return to patients and research participants. Operationalizing these principles is not straightforward. Here, we illustrate considerations for selecting genetic variants to return to participants in a nationwide biobank study.

Methods

The Million Veteran Program (MVP) is a multi-ethnic biobank study of >750,000 U.S. Veterans receiving VA healthcare. Upon enrollment, participants are genotyped with an Affymetrix Axiom Biobank Array (MVP chip) of ~723K variants enriched for known disease-associated SNPs. As consensus around variant interpretation and actionability has emerged, MVP investigators are now planning a pilot study to return genetic results associated with familial hypercholesterolemia (FH) via a new research protocol, the MVP Return Of Actionable Results (ROAR) FH Study. Here, we consider 3 overlapping categories of variants in FH genes (*LDLR*, *APOB* and *PCSK9*) on the MVP chip: 1) variants classified by at least one submitter in ClinVar as pathogenic (P) or likely pathogenic (LP) in Sun (ASHG 2017) and as of May 2019; 2) variants associated with high LDL cholesterol in MVP data, 3) variants interpreted as P/LP by Invitae, a commercial clinical laboratory.

Results

Of 58 variants classified as P/LP in ClinVar in 2017, 2 moved from conflicting interpretation (CI) to P/LP, 2 moved from P/LP to CI, and 2 moved from CI to variant of uncertain significance by 2019. Of the original 58 variants, 16 are not monomorphic and present in at least 30 MVP participants. Of these, 8 are significantly associated with high LDL in MVP data; only 2 are currently classified as P/LP in ClinVar. Of the original 58 variants, 26 are currently classified as P/LP by Invitae while another 13 remain unclassified by Invitae. Eleven variants are classified as P/LP by Invitae but are not yet in ClinVar.

Discussion

Evolving variant interpretation poses a challenge for returning unanticipated genetic results. In considering which FH-associated variants to return, MVP-ROAR-FH investigators are monitoring the work of the ClinGen FH Variant Curation Expert Panel, which is developing FH-specific rules for applying ACMG variant classification guidelines. As biobank studies consider recontacting participants and returning genetic results, the clinical validity of variant interpretation is of paramount importance.

A- A+

PgmNr 265: Comprehensive secondary findings analysis of parental samples submitted for exome evaluation yields a high positive rate.

Authors:
E.V. Haverfield; E.D. Esplin; S. Aguilar; S. Yang; R. Truty; R.L. Nussbaum; S. Aradhya

View Session Add to Schedule

Affiliation: Invitae Corporation, San Francisco, CA

The American College of Medical Genetics and Genomics (ACMG) recommends providing results representing secondary findings in 59 genes associated with medically actionable monogenic disorders in individuals undergoing clinical diagnostic whole exome (WES) or whole genome (WGS) sequencing. Healthy parents tested alongside the proband for a trio analysis can also choose to receive secondary findings. This study evaluates the frequency of medically actionable findings in this group of genes in probands as well as their unaffected parents.

We analyzed de-identified data from 4,325 individuals who requested secondary findings as part of their WES analysis. These analyses represent evaluations of probands and parental samples. We analyzed every individual for all 59 genes and provided a personalized secondary findings report for each individual, independent of the proband's WES result.

Pathogenic/likely pathogenic (P/LP) variants in genes conferring increased risk for hereditary breast and ovarian cancer (HBOC), Lynch syndrome, familial hypercholesterolemia and other conditions were identified in 200 of 4,325 (4.6%) individuals, including probands and parents. If heterozygous P/LP variants in *MUTYH* and the *APC* I1307K increased risk allele were excluded, the detection rate of secondary findings was 142 of 4,325 (3.3%). Of the 2,195 families evaluated (proband only, duo and trio analyses),154 had a positive finding. Of these, 57 were in a proband-only (37.0%), 38 were in a proband and one parent (24.7%), 50 were in a parent-only (32.5%), and 9 (5.8%) families showed more than one positive finding.

Medically significant secondary findings were identified in 4.6% of individuals undergoing WES. Notably, analysis of the 59 genes identified unknown personal and familial risk in a parent-only in one-third of positive findings, including the conditions recommended by the CDC for population genomic screening. Although screening for these 59 genes is performed opportunistically in individuals undergoing genomic sequencing for other reasons, these data highlight the potential benefit of using a targeted gene panel to screen for hereditary disease risk in the general population. Along with rapidly decreasing costs of genomic testing, the medical actionability of findings in these genes for early detection and disease prevention supports responsibly providing physician-directed access to genetic screening.

A- A+

PgmNr 266: Secondary inherited cardiac condition findings from genome sequencing: Variant interpretation, assessment of phenotype, and impacts of disclosure.

Authors:
A.R. Harper [1,2]; K. Thomson [1,2,3]; M. Mackley [1]; H. Watkins [1,2]; E. Ormondroyd [1]

View Session Add to Schedule

Affiliations:
1) Radcliffe Department of Medicine, Division of Cardiovascular Medicine, University of Oxford; 2) Wellcome Centre for Human Genetics, University of Oxford; 3) Oxford Molecular Genetics Laboratory, Oxford University Hospitals NHS Foundation Trust

The concept of ‘secondary findings’ (SF) relates to the detection of genomic variants that are deemed pathogenic in individuals not known to manifest the associated disease. It has been speculated that the penetrance of SF is low. With widespread adoption of genome sequencing and lack of policy recommendation consensus, there is a need for robust evidence to inform interpretation and disclosure of SF. Evaluation of SF in inherited cardiac conditions (ICC), encompassing inherited cardiomyopathies and long QT syndrome, is most tractable. An absence of clinical changes in adulthood suggests a risk of serious complications equivalent to background population risk. ICC SF are actionable at a single point in time, as opposed to cancer SF which require life-long surveillance.

We present an approach to evaluate SF for ICC. A recall-by-genotype, double-blind, randomized, case-control study design for 7,203 participants, with no history of an ICC diagnosis, within the NIHR BioResource for Rare Disease was implemented. 25 genes robustly associated with ICC were screened for variants considered pathogenic and likely pathogenic. Eligible variant carriers were randomised 1:1 with age and sex matched non-variant carriers, and contacted using a 2-stage process to protect their ‘right not to know’. Genetic counselling and medical evaluation, including cardiac phenotyping, was performed. Investigators and participants were blinded to genotype status during investigation. Genotype status was unblinded and disclosed with clinical findings on the same day; participants were referred onto a clinical pathway, including facilitation of family cascade testing, as appropriate.

42 variant carriers were identified (0.58% [95% CI: 0.43–0.79%]), of whom 20 were eligible (based on age (18–80 years old) and comorbid status) and agreed to participate. Four variant carriers (mean age=60) and seven controls (mean age=55) have participated so far. Clinical evidence of cardiomyopathy was detected in 1 variant carrier, but not in the other 3, nor in controls. 1 qualitative interview per participant has been performed to explore impacts of disclosure.

In summary, using stringent classification of variants in ICC genes, we report a SF prevalence of 0.58% [95% CI: 0.43–0.79%] within a UK cohort. Our study provides a framework upon which future studies can evaluate the clinical significance and utility of SF. This is critical to provide the necessary policy recommendations to support clinical care.

A- A+

PgmNr 267: Secondary findings in non-ACMG 59™ genes.

Authors:
A.E. Katz [1]; J. Paschall [2]; H. Shiferaw [1]; X. Liu [3]; W.S.W. Wong [3]; T.P. Conrads [4]; G.L. Maxwell [4]; D.P. Ascher [5]; L.G. Biesecker [1]; The Genomic Ascertainment Cohort (TGAC)

View Session Add to Schedule

Affiliations:
1) Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, Bethesda, MD, USA; 2) Computational and Statistical Genomics Branch, National Human Genome Research Institute, NIH, Bethesda, MD, USA; 3) Division of Medical Genomics, Inova Translational Medicine Institute, Falls Church, VA, USA; 4) Department of Obstetrics and Gynecology, Inova Schar Cancer Institute, Inova Fairfax Hospital, Falls Church, VA, USA; 5) Inova Children's Hospital, Inova Health System, Falls Church, VA, USA

An estimated 2-4% of individuals harbor a reportable secondary finding among the ACMG recommended minimum list of 59 genes. The rate of pathogenic or likely pathogenic variants in genes with known disease association outside of the ACMG 59™ is largely unknown. We estimated the collective rate of secondary findings in 42 genes not included in the ACMG 59™ by reviewing sequence data from a large cohort of adults unselected for any phenotype. To determine our genes of interest, we aggregated existing gene lists for return of findings identified on genome or exome sequencing among large screening studies. We excluded the ACMG 59™ genes from this aggregated list. We then removed genes that were only expected to cause disease in an autosomal recessive or X-linked inheritance pattern. Using the resultant 42 genes, we queried our cohort of 4,872 adults who had previously undergone research exome or genome sequencing. We identified all rare missense and predicted loss of function (pLOF) variants in our cohort by applying a variant allele frequency cutoff of 0.001. We then manually reviewed the variants, applying ACMG variant pathogenicity criteria as if they had been identified via clinical sequencing. No phenotypic information about the individuals in our cohort was used in the variant interpretation. We considered variants to be secondary findings if there was sufficient evidence to classify them as pathogenic or likely pathogenic. About 0.60% (29/4,872, 95% CI 0.38-0.81%) of individuals had a secondary finding in one of the 42 genes of interest. Twenty-nine unique variants were identified as secondary findings in our cohort. No individual had more than one variant we considered to be a secondary finding. We identified seven missense and 22 pLOF variants. Secondary findings were identified in 19 genes out of the queried list of 42. The phenotypes known to be associated with these genes include mendelian forms of cardiovascular disease, cancer predisposition, and hematologic disorders. Preliminary electronic health record review of these individuals shows findings corresponding to the predicted phenotype. We provide an estimate of the rate of secondary findings among 42 disease-associated genes among adults unselected for any disease phenotype. The clinical utility of a secondary finding in these 42 genes is unknown. Targeted evaluation of individuals identified in our cohort can help determine the clinical yield of these findings.

A- A+

PgmNr 268: Expanded CGG repeat RNA and FMRpolyG in the oocyte leads to impaired response to gonadotropin hormones in a murine model of the *FMR1* premutation.

Authors:
K. Shelly[1]; N. Candelaria[2]; Z. Li[3]; P. Jin[4]; D. Nelson[1]

View Session Add to Schedule

Affiliations:
1) Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030; 2) Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas 77030; 3) Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30329; 4) Human Genetics, Emory University, Atlanta, Georgia 30329

Women heterozygous for an expansion of CGG repeats in the 5'UTR of *FMR1* are at risk to develop Fragile X-associated Primary Ovarian Insufficiency (FXPOI). While approximately 20% of female premutation carriers will be clinically diagnosed with POI, the molecular underpinnings remain understudied. We previously established that ectopic expression of an expanded CGG tract is sufficient to drive ovarian dysfunction and reproductive senescence in a murine model of the *FMR1* premutation. Using heterozygous females from two conditional mouse lines expressing either CGG RNA-only (RNA-only) or CGG RNA and its aberrantly translated polyglycine product FMRpolyG (FMRpolyG+RNA) we demonstrate in this study that deficits in ovarian function are apparent in young mice and are intrinsic to the ovary. Further, we leverage this early impaired ovarian response in CGG-expressing females to determine specific molecular processes that lead to reduced oocyte ovulation. By tracking histological markers and transcriptional changes following ovulation induction, we observed impaired cumulus expansion in the preovulatory follicles of RNA-only and FMRpolyG+RNA ovaries. This is accompanied by a reduction of meiotic resumption in the oocytes normally triggered by ovulation. Additionally, we demonstrate cell-specific expression of FMRpolyG+RNA in oocytes (under *Gdf9*-Cre) recapitulates reduced oocyte ovulation but not expression in granulosa cells (under *Cyp19a1*-cre). Expression of RNA-only in either oocytes or granulosa cells did not reduce ovulated oocytes. Interestingly, analysis of cumulus expansion-enabling factors downstream of the predominant oocyte signals GDF9 and BMP15 indicated some differences but does not account for all histological findings. Expanding the search for mechanism we utilized RNA-sequencing of unstimulated immature ovaries to identify alternative signaling pathways for cumulus expansion, including effectors of EGFR and MAP3K1 signaling. Cumulatively, our results show that early perturbations in ovarian function can be recapitulated using cell-specific expression of CGG repeats and that these effects are apparent only when both CGG RNA and its translated polyglycine product are expressed. This study is the first to identify cumulus expansion and meiotic resumption as critical processes altered by CGG expression. It is also the first describe differential effects of FMRpolyG+RNA and CGG RNA alone in these ovarian functions.

A- A+

PgmNr 269: Exome sequencing reveals *de novo* mutations and deletions in severe unexplained male infertility.

Authors:
M. Oud[1]; R.M. Smits[2]; F.K. Mastrorosa[3]; H. Smith[3]; M.J. Xavier[3]; G.S. Holt[3]; H. Sheth[3]; B.J. Houston[4]; T. Luan[4]; R. Burke[4]; M.K. O'Bryan[4]; P.F. de Vries[1]; B. Alobaidi[3]; H. Ismail[3]; A. Garcia-Rodriguez[3]; A. Mikulasova[5]; G. Astuti[1]; C. Gilissen[1]; L.E.L.M. Vissers[1]; C. Friedrich[6]; F. Tuttelmann[6]; K. McEleny[7]; J. Coxhead[8]; S. Cockell[9]; D.D.M. Braat[2]; K. Fleischer[2]; G.W. van der Heijden[1,2]; L. Ramos[2]; J.A. Veltman[1,3]

View Session Add to Schedule

Affiliations:
1) Human Genetics, Radboudumc, Nijmegen, Netherlands; 2) Obstetrics and Gynaecology, Radboudumc, Nijmegen, the Netherlands; 3) Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom; 4) School of Biological Sciences, Monash University, Melbourne, Australia; 5) Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne, United Kingdom; 6) Institute of Human Genetics, University of Münster, Münster, Germany; 7) Newcastle Fertility Centre, Newcastle University, Newcastle upon Tyne, United Kingdom; 8) Genomics Core Facility, Newcastle University Newcastle upon Tyne, United Kingdom; 9) Bioinformatics Support Unit, Newcastle University, Newcastle upon Tyne, United Kingdom

One in every 12 men is infertile and approximately 40% of these men produce no or insufficient numbers of sperm (azoospermia/severe oligozoospermia). We recently performed a clinical validity assessment for all reported male infertility genes (Oud et al. Hum. Reprod. 2019), demonstrating that there is a lack of diagnostically relevant genes for this disorder. *De novo* mutations (DNMs) are known to play a prominent role in early-onset disorders with reduced fitness. However, the role of dominant DNMs in male infertility remains unexplored, partly due to the difficulty in obtaining parental samples. Here we report on the first exome sequencing study to investigate the role of *de novo* mutations in male infertility.

We examined and sequenced 101 patients suffering from non-obstructive azoospermia or severe oligozoospermia (<5 million sperm/ml) and their fertile parents. In total, we identified and validated 90 protein-altering DNMs, which show an enrichment in STRING protein network edges ($p=6.17x10^{-4}$). In addition, we found an enrichment of loss-of-function (LoF) variants in extremely LoF intolerant genes ($pLi_{median} = 1.00$, $p=7.90x10^{-3}$). Of all DNMs, 52 are likely to disrupt normal gene function and affect genes expressed in the testis. Of those, 22 lie in genes involved in sperm production such as *TOPAZ1* and *ODF1*, but none of these are known human male infertility genes. Complementarily, we detected rare *de novo* copy number variants in 2 patients affecting multiple genes involved in cell replication and spermatogenesis.

While we are currently replicating these results in an additional cohort of 100 patient-parent trios, it is clear that much larger cohorts are required to comprehensively identify and characterize recurrently mutated genes. This is one of the reasons for establishing the International Male Infertility Genomics Consortium (imigc.org). We have also initiated *Drosophila melanogaster* knockdown studies for orthologs of 37 out of 52 new candidate genes. Testis somatic cell knockdown of *LEO1*, *EXOSC9*, *ATP8A1*, *FOXF2* and *HOXA1* resulted in male sterility which indicates that these genes may play an evolutionary conserved role in spermatogenesis.

In conclusion, our data provide the first indications that DNMs may play an important role in severe male infertility. A DNM approach will likely identify novel genes and help to increase the number of diagnostically relevant genes, which will bring us closer to a molecular diagnosis for male infertility patients.

A- A+

PgmNr 270: Aneuploidy and recombination across chromosomes, gametes, and individuals from large-scale single-sperm sequencing.

Authors:
A.D. Bell [1,2]; C.J. Mello [1,2]; J. Nemesh [1,2]; S.A. Brumbaugh [1,2]; A. Wysoker [1,2]; A. Leung [3,4]; D. Sakkas [3,4]; S.A. McCarroll [1,2]

View Session Add to Schedule

Affiliations:
1) Department of Genetics, Harvard Medical School, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA; 3) Boston IVF, Waltham, MA; 4) Division of Reproductive Endocrinology and Infertility, Department of Ob/Gyn, Beth Israel Deaconess Medical Center, Boston, Massachusetts

Human meiosis, though critical for reproduction, is error-prone and varies across individuals, gametes from the same individual, and human sexes. Aneuploidy, the loss or gain of a chromosome, leads to miscarriage and morbidity, and occurs frequently during oogenesis and at lower but appreciable frequency during spermatogenesis. To investigate aneuploidy and recombination in spermatogenesis, we previously developed a method to sequence many sperm genomes at once and used it to sequence >30,000 cells across 20 sperm donors, identifying >800,000 crossovers and 787 aneuploidies. Here we describe insights from new analyses of these data.

In addition to observing that individuals and chromosomes were variably vulnerable to aneuploidy, we found that most autosomal chromosome gains were of sister chromatids (74.8%), presumably deriving from nondisjunction in meiosis II (MII). Most sex chromosome gains were of homologous chromosomes (68.9%), deriving from nondisjunction in meiosis I (MI). Rates of nondisjunction in MI and MII were uncorrelated across chromosomes and sperm donors, consistent with distinct *cis* and *trans* factors underlying nondisjunction in MI and MII.

Crossovers were associated with proper disjunction in MI: chromosomes gained during MI had 36% fewer crossovers than matched non-aneuploid chromosomes. Neither sperm cells with aneuploidy nor sperm donors with high aneuploidy rates had detectably lower crossover rates.

Crossover phenotypes were also related to one another. Both sperm donors and sperm cells with higher crossover rates placed proportionally more crossovers in centromere-proximal regions and made their crossovers closer together. This parallelism suggests that similar processes may underpin variation at two levels: among cells and among people.

These relationships might be explained in terms of differential compaction of meiotic chromosomes. Because crossover interference acts over physical (micron) rather than genomic (basepair) distances, physically longer (less tightly compacted) chromosomes allow more crossovers, coupling crossover number, location, and separation. The compaction of chromatin varies among spermatocytes and is correlated across different chromosomes in the same nucleus; we suggest it also differs on average across individuals.

We are currently applying high-throughput single-sperm sequencing to clinical infertility, sequencing sperm from couples undergoing *in vitro* fertilization where a male factor may be causal.

A- A+

PgmNr 271: Disruption of genes essential for Müllerian duct/Wolffian duct development in Mayer-Rokitansky-Küster-Hauser syndrome (MRKHS).

Authors:
N. Wu [1,2,3]; N. Chen [4]; S. Zhao [1,2]; H. Pan [5]; L. Wang [1,2]; A. Jolly [3]; Z. Wu [1,2,6]; P. Liu [3,7]; J. Posey [3]; J. Lupski [3,8,9]; L. Zhu [4]

View Session Add to Schedule

Affiliations:
1) Beijing Key Laboratory for Genetic Research of Skeletal Deformity, Beijing 100730, China.; 2) Department of Orthopedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100730, China.; 3) Department of Molecular and Human Genetics, Baylor College of Medicine; Houston, TX 77030, USA.; 4) Department of Obstetrics and Gynaecology, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100730, China.; 5) Department of Obstetrics and Gynaecology, The 3rd Affiliated Hospital of Shenzhen University, Luohu hospital, Shenzhen, Guangdong 518000, China.; 6) Medical Research Center, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China; 7) Baylor Genetics, Houston, TX 77021, USA.; 8) Departments of Pediatrics, Texas Childen's Hospital and Baylor College of Medicine, Houston, TX 77030, USA.; 9) Texas Children's Hospital, Houston, TX 77030, USA.

Introduction
Mayer-Rokitansky-Küster-Hauser syndrome (MRKHS) is a birth defect with congenital absence of uterus, cervix, and the upper part of the vagina in females with normal karyotype (46, XX). Disrupted development of the Müllerian ducts (MD)/ Wolffian ducts (WD) has been proposed to manifest in MRKHS. However, the underlying biological and developmental pathways remain to be elucidated as only one established disease-causing gene (*WNT4*) has been identified as a cause of MRKHS with hyperandrogenism, thus far.

Methods
To explore the biological pathways potentially perturbed and the genetics underlying MRKHS, we ascertained a discovery cohort of 442 Chinese patients with MRKHS from two medical centers. A replication cohort of Chinese patients with hypothyroidism and an isolated Caucasian MRKHS case were also collected. Exome sequencing (ES) was performed on the discovery cohort and 941female controls. We performed family-based genomics, taking into consideration a potential sex-limited disease trait, and analyzed the mutational burden and variant spectrum of genes required for MD/WD duct development.

Results
The discovery cohort consisted of 442 MRKH patients including 337 singletons and 105 trio families. Among them were 330 (74.7%) patients with MRKHS type I (isolated) and 109 (25.3%) with MRKHS type II (syndromic). By analyzing 19 candidate genes, we identified twelve protein-truncating variants in seven genes: *PAX8*, *BMP4*, *BMP7*, *TBX6*, *HOXA10*, *EMX2*, and *WNT9B*, while no truncating variants in any of the candidate genes were detected in 941 female control samples (P = 1.27E-06). Two patients with available parental samples demonstrated paternal inheritance of the truncating variants in *PAX8*. A DNA-binding assay revealed two additional *PAX8* deleterious variants. In our replication

cohort, we identified two patients to be affected with MRKHS from five Chinese females with hypothyroidism and molecular diagnoses of *PAX8* pathogenic variants. We also identified a pathogenic *PAX8* missense variant in a single Caucasian case affected with MRKHS.

Conclusion

Loss-of-function variants in MD/WD duct developmental pathway genes were significantly enriched in MRKHS patients. *PAX8*, *BMP4* and *BMP7* are candidate disease-associated genes in MRKHS, in which *PAX8* represents the most significant disease-associated gene underlying the etiology of patients in the discovery cohort and 3 additional patients in replication cohorts.

A- A+

PgmNr 272: Sexual dimorphism in genetic associations of testosterone and sex-hormone binding globulin with cardiometabolic diseases.

Authors:
J. Hu [1]; J. Li [2,3]; K.M. Rexrode [1]; L. Liang [3]

View Session Add to Schedule

Affiliations:
1) Division of Women's Health, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.; 2) Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA.; 3) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA.

Explaining sex disparities in the risk of cardiometabolic diseases, such as coronary heart disease (CHD) and type 2 diabetes (T2D), may lead to better understanding in etiology research. Prior epidemiological evidence shows that prospective associations between testosterone (T) and sex-hormone binding globulin (SHBG) with risk of CHD and T2D may be different between men and women. However, whether there is sexual dimorphism in the genetic regulation of T and SHBG, and whether such genetic heterogeneity contributes to sex differences in CHD and T2D development, are largely unknown. We hereby carried out a sex-stratified genome-wide association study (GWAS) for CHD, T2D, T, and SHBG, in 20,9978 men and 14,4921 postmenopausal women from the UK Biobank study. Using linkage disequilibrium score regressions, we observed significant inverse genetic correlations between SHBG with CHD and T2D; the correlations were stronger in women (CHD: r_g=-0.33, T2D: r_g=-0.62; P<3.1×10^{-9}) than in men (CHD: r_g=-0.20, T2D: r_g=-0.36; P<4.4×10^{-9}). However, genetic correlations for T were only significant in men (CHD: r_g=-0.17, T2D: r_g=-0.34; P<3.1×10^{-7}) but not in women. Sex-stratified GWAS identified 17 loci showing genome-wide significant associations with CHD in at least one sex but with significant heterogeneity between sexes (including known CHD loci *CDKN2B*, *PCSK9*, *APOE* and *LPA*; P_{het}<5.9×10^{-5}). Five loci were identified showing between-sex heterogeneity for T2D (including the known T2D locus *TCF7L2*). While only eight-percent loci that were associated with SHBG levels in at least one sex showed significant between-sex heterogeneity, over half of the loci that were associated with T levels in at least one sex exhibited significant sex differences with most having associations in opposite directions. We identified 4 loci showing sex-specific associations for a biomarker (T and/or SHBG) and a disease (CHD and/or T2D), and further examined their potential functions and involved pathways, including associations with tissue-specific expressions. In conclusion, we observed significant sex-differences in genetic correlations between T and SHBG with CHD and T2D, and identified substantial genetic loci showing between-sex heterogeneity in associations with CHD, T2D, T, and SHBG. The between-sex heterogeneity in genetic effects may contribute to sex-differences in cardiometabolic disease etiology, which may partially be expressed through differential regulation of the sex-hormone pathways.

A- A+

PgmNr 273: A sex-stratified meta-analysis of 537,602 individuals identifies sex-specific effects on susceptibility to inguinal hernia.

Authors:
E. Jorgenson [1]; J. Yin [1]; A. Sohota [2,3]; R. Mostaedi [4]; N. Ahituv [2,3]; H. Choquet [1]

View Session Add to Schedule

Affiliations:
1) Kaiser Permanente Northern California (KPNC), Division of Research, Oakland, CA 94612, USA; 2) Institute for Human Genetics, University of California San Francisco (UCSF), San Francisco, CA 94143, USA; 3) Department of Bioengineering and Therapeutic Sciences, UCSF, San Francisco CA 94158, USA; 4) KPNC, East Bay Richmond Medical Center, CA 94801, USA

Purpose: Inguinal hernias can lead to serious medical morbidity, including bowel incarceration and strangulation, and emergency hernia repair surgery is associated with a substantial risk of morbidity and mortality. Men have a higher lifetime prevalence of inguinal hernias (20–27%) compared to women (3–6%); it is not clear why this difference exists. To date, 4 genetic loci have been associated with hernia risk, but they explain a proportion of variance in risk of only 1.0–1.4% in men and 1.3–2.8% in women. Array-based heritability estimates found a stronger contribution of genetic risk factors in women (20.8 to 25.5%) compared to men (13.2 to 18.3%), suggesting that sex-specific genetic effects may underlie some of the difference in risk.
Methods: We conducted a sex-stratified genome-wide association meta-analysis of inguinal hernia risk, combining results from the Genetic Epidemiology Research in Adult Health and Aging (GERA) and the UK Biobank (UKB) cohorts. Genetic association analyses of inguinal hernia were performed in men and women separately using logistic regression adjusted for age and ancestry principal components. Additionally, we characterized hernia-associated loci using *in silico* tools, and RNA-seq and ChIP-seq experiments.
Results: We identified 33 genome-wide significant loci (P<5x10^{-8}) in men, including 29 that were novel. A total of 8 of the 33 loci identified in men also reached a Bonferroni level of significance in women. While we had a considerably smaller number of cases in women than in men, we identified 3 genome-wide significant loci (*LYPLAL1*, *EFEMP1*, and *WT1*), including 1 novel, each of which was also identified in men. Of the three loci significantly associated in both men and women, variation in the *LYPLAL1* locus has a significantly stronger effect in women than men (lead variant rs2820446, OR = 0.77, P=2.3x10^{-16} in women; OR = 0.94, P=1.9x10^{-12} in men). The *LYPLAL1* locus has previously been associated with variation in waist-to-hip-ratio adjusted for BMI in women, but not in men, suggesting that some of the sex difference in inguinal hernia susceptibility may be due to sex differences in anatomical structure.
Conclusions: This largest study conducted to date on the genetic contributors to inguinal hernia susceptibility identified many novel genetic loci and sex-specific genetic effects, providing novel biological insights into hernia etiology.

A- A+

PgmNr 274: Sex-biased gene expression in fetal brain is distinct from adult brain and gives new insights on neurodevelopmental disorders.

Authors:
C. Benoit-Pilven [1]; J. Leinonen [1]; J. Karjalainen [1,2]; M. Daly [1,2]; T. Tukiainen [1]

View Session Add to Schedule

Affiliations:
1) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Despite the established sex differences in human brain organization and in the prevalence of several neurodevelopmental disorders, little is known about how and when these sex differences emerge. Many brain disorders are shown to be rooted in early development and so investigations of adult brain may not reflect the relevant differences. Thus, to gain more insight on the sex differences in brain, we analyzed sex biases in gene expression in human forebrain both during early development (5-17 post-conceptional weeks) and in adulthood, using RNA-seq data from the Human Developmental Biology Resource (N=279) and GTEx (N=1054).

Pointing to the early developmental origins of sex differences, we detected 2053 genes as differentially expressed between sexes (sex-DE) in prenatal forebrain (qvalue<0.1). This sex-bias was largely distinct from adult brain. For autosomal genes, the sex-DE consistency was close to what is expected by chance (53% of genes with the same sex-bias direction). However, for genes escaping from X-chromosome inactivation (XCI), the consistency was high (91%). This suggests an early regulation of the sex-biased expression of XCI genes, in contrast to the variability of autosomal genes' sex-bias across lifespan.

To better understand the biological implications of these sex biases for the developing brain, we performed enrichment analyses for Gene Ontology terms and for genes associated with neurodevelopmental diseases. Several pathways related to neurogenesis, transcription, and translation were found as male-biased (P<0.05), potentially reflecting a higher cell proliferation rate in male forebrain during development, as well as a different cell type composition. Male-biased genes were also enriched for genes associated with developmental disorders ($P=5.1\text{x}10^{-4}$), whereas, female-biased genes, interestingly, showed an enrichment for genes associated with autism ($P=2.6\text{x}10^{-4}$), a disorder with a higher male prevalence. The higher female expression of autism genes could reflect the suggested female-protective effect, as the deleterious effects of mutations in these genes could be buffered by their globally higher expression. Notably, these enrichments were only detected in fetal and not in adult forebrain.

Our study demonstrates the importance of analyzing brain at various life stages and suggests that sex-DE genes during early development may contribute to the observed differences in the prevalence of some neurodevelopmental disorders between sexes.

A- A+

PgmNr 275: Sex-dependent glia-specific changes in the epigenome landscape of the *APOE* region in Alzheimer’s disease brains.

Authors:
O. Chiba-Falek [1, 2]; L. Song [1, 2]; J. Barrera [2]; A. Safi [2]; Y. Jun [1, 2]; M. Garrett [3]; A. Ashley-Koch [3]; G. Crawford [2]

View Session Add to Schedule

Affiliations:
1) Neurology, Duke Univ, Durham, North Carolina.; 2) Center for Genomic and Computational Biology , Duke Univ, Durham, North Carolina.; 3) Medicine,Duke Univ, Durham, North Carolina.

Sex differences in clinicopathological progressions of late-onset Alzheimer’s disease (LOAD) have been described and women are often reported to have increased incidence of LOAD. Thus, there is an unmet need for better understanding of sex differences in LOAD. Towards this goal we investigated the sex-dependent epigenome landscape including, chromatin accessibility and DNA-methylation, in single cell-types from LOAD vs. normal brains to determine LOAD-specific changes. We applied NeuN Fluorescence Activated Nuclei Sorting (FANS) to sort neuronal vs. non-neuronal nuclei using frozen brain cortex from 19 LOAD-patients and 21 matched controls. This strategy characterizes cell-type specific molecular profiles and overcomes shortcoming of bulk brain tissue analyses such as a bias due to sample-to-sample variability in the proportion of the different cell-types. The NeuN$^+$ and NeuN$^-$ sorted nuclei were used for ATAC-seq, methylationEPIC microarray complemented by pyrosequencing to profile chromatin accessibility and DNA-methylation, respectively, followed by data integration with known LOAD-risk loci. We identified nearly 88,000 neuronal specific ATAC-seq sites and ~83,000 non-neuronal specific sites. Comparisons between LOAD and control brains discovered approximately 221 significant differential chromatin accessibility sites in neuronal nuclei. While we didn’t identify LOAD-specific sites in non-neuronal nuclei, when we stratified the analysis by sex, we identified 842 LOAD-specific chromatin accessibility sites in females only. Seven LOAD-specific female-specific non-neuronal open chromatin sites and 4 neuronal sites coincide with LOAD-GWAS regions, indicating candidate regulatory elements that are likely causative in LOAD. Similarly, we observed sex-dependent differential DNA-methylation between LOAD and control brains in non-neuronal nuclei that overlay with the *APOE* region encompassing the two SNPs that determine the *APOE* e2/3/4 alleles. In the post-GWAS era the key challenge is to translate association to causation to interpret the genetic etiologies of LOAD. Here, single-cell technologies revealed the effects of sex-dependent glia-specific epigenomic alterations on LOAD pathology, which may explain the molecular mechanisms through which several LOAD-GWAS loci including the *APOE* region exert their pathogenic effects. In summary, single brain cell-type omics profiling combined with sex-specific analyses are imperative for decoding LOAD genetics discoveries.

A- A+

PgmNr 276: Large scale whole genome sequencing reveals the mutational and clonal landscape of the IBD colon.

Authors:
S. Olafsson [1]; R. McIntyre [1]; T. Coorens [1]; P. Robinson [1]; H. Lee-Six [1]; M. Sanders [1,2]; T. Butler [1]; K. Arestang [3]; C. Dawson [3]; Y. Hooks [1]; M.R. Stratton [1]; I. Martincorena [1]; M. Parkes [3]; T. Raine [3]; P.J. Campbell [1]; C.A. Anderson [1]

View Session Add to Schedule

Affiliations:
1) Wellcome Trust Sanger Institute, Hinxton, UK; 2) Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands; 3) Gastroenterology Research Unit, Cambridge Biomedical Campus Addenbrooke's Hospital, Hills Road, Cambridge, UK

To understand the development of cancer, it is crucial to understand the mutagenic processes affecting the non-neoplastic normal tissue from which cancer develops. In recent years, we and others have characterized the driver landscape, mutation burden and mutagen exposure of tissues such as skin, endometrium, esophagus and more. Here, we build on our work on normal tissue to study the effect of a complex disease on the somatic mutation landscape of the colon before the development of cancer. Inflammatory bowel disease (IBD) is a debilitating disease characterized by repeated flares of inflammation in the colon. Individuals with IBD have an elevated risk of developing colorectal cancer, they require regular screening and may even undergo prophylactic colectomy to mitigate this risk.

To better understand the changes in the mutational- and clonal landscapes of the colon that predispose IBD patients to colorectal cancer we have dissected and whole-genome sequenced ~350 individual colonic crypts from actively inflamed, previously inflamed and never inflamed tissue from ~40 IBD patients. We compare the somatic mutation burden, mutational signature exposure, clonal structure and cancer driver mutation landscape in actively and previously inflamed regions with crypts from never-inflamed regions of the same individuals, and with a control cohort from our recent study of the mutational landscape of the normal colon (Lee-Six et al, in review).

Preliminary results indicate that clonal expansions beyond the confines of the crypt, which are rare in the normal colon, are common in IBD. These clonal expansions however, are rarely driven by mutations in known colorectal cancer genes, which might be hypothesized to confer selective advantage on carrier crypts. We estimate a mutation burden of 65 (CI = 38-92), 66 (CI = 38-94) and 61 (CI = 29-98) somatic mutations per crypt per year for the left, right and transverse colon, respectively. Mutational signature analysis does not identify a new signature specific to IBD but indicates that the relative contribution of known signatures may be altered after disease onset. In particular, our results suggest that mutagenesis by reactive oxygen species is accelerated compared with mutagenesis due to cell proliferation or age (beta = 0.28 (CI=0.12-0.45), $P = 0.00052$). Finally, we detect a mutational signature of purine treatment in a subset of patients receiving purine analogue immunosuppressive treatment.

A- A+

PgmNr 277: Whole genome sequencing puts Cas9 off-target mutagenesis into the context of genetic drift.

Authors:
L.M.J. Nutter [1]; S. Khalouei [2]; J.D. Heaney [3]; D.G. Lanza [3]; S.M. Murray [4]; K. Peterson [4]; J.R. Seavitt [3]; J.A. Wood [5]; A. Ramani [2]

View Session Add to Schedule

Affiliations:
1) The Centre for Phenogenomics, The Hospital for Sick Chlidren, Toronto, ON, Canada; 2) The Centre for Computational Medicine, The Hospital for Sick Children, Toronto, Canada, M5G 1X8; 3) Baylor College of Medicine, Houston, TX, 77030; 4) The Jackson Laboratory, Bar Harbor, ME, 04609; 5) Mouse Biology Program, University of California Davis, Davis, CA, 04609

There are reports demonstrating that Cas9 introduces off-target mutations and others that off-target mutation rates are low. The majority of reports investigate one or a few guide RNAs, which may result in sequence or chromosome location bias. The Knockout Mouse Phenotyping Project (KOMP2) produces mutant mice in a high-throughput pipeline using Cas9 for mutagenesis in the inbred C57BL/6N strain. This has enabled us to use whole genome sequencing to assess mutations in the genomes of 51 Cas9-derived founder mice representing 162 different gRNAs along with 25 inbred control mice. Illumina paired-end reads provided >35X coverage with ≥90% of bases with >25 reads. Variants (SNPs and indels) were identified using GATK4.0 and structural variants using an intersection of Lumpy, Manta, CNVkit, and Wham, followed by MetaSV. Variants were filtered out when they occurred in two or more mice, indicating the variant likely resulted from genetic drift rather than from Cas9 activity. We used CasOFFinder to identify predicted off-targets with up to 5 mismatches and one DNA or RNA bulge among the variants in the respective founder for each gRNA. There were 20 genes for which one or more Cas9-induced off-target mutations were identified (46 total with a range of 1-10 and average of 2.3 per founder). For 31 genes, no Cas9-induced off-target mutations were identified. Importantly, these analyses demonstrated that there was an average of ~3,500 variants unique to each animal – founder or untreated control. Two important conclusions can be drawn; (1) with appropriately designed Cas9 gRNAs off-target mutagenesis is rare; and (2) genetic drift within a carefully maintained inbred line of mice results in thousands of genetic variants between individuals within that line. These results have implications in the use of mice to model human disease, i.e. that backcrossing or outcrossing mice introduces significantly more variation than the use of Cas9 and the appropriate controls are littermate or line mate wild-type mice for most genetic experiments. These results also raise the question. What is “normal” genetic sequence in the context of model organisms and in humans - patients, controls, tissues and cell lines; for both assessing the specificity of Cas9 for genome editing and assessing the consequences of variants associated with disease?

A- A+

PgmNr 278: Using in vitro evolution to probe the genome-wide basis of chemotherapy resistance.

Authors:
M. Dow [1,2,6 *]; J.C. Rodriguez [3,6 *]; H. Carter [1,4,5]; E. Winzeler [3]

View Session Add to Schedule

Affiliations:
1) Division of Medical Genetics, Department of Medicine, University of California, San Diego, La Jolla, CA; 2) Health Science, Department of Biomedical Informatics, School of Medicine, University of California, San Diego, La Jolla, CA; 3) Division of Host-Microbe Systems & Therapeutics, Department of Pediatrics, University of California, San Diego, Gilman Dr., La Jolla, CA, USA; 4) Moores Cancer Center, University of California, San Diego, La Jolla, CA; 5) Cancer Cell Map Initiative, University of California, San Diego, La Jolla, CA; 6) *these authors contributed equally to this work

Advances in next generation sequencing (NGS) technologies has revolutionized our understanding of tumor biology and the mechanisms associated with the development of tumor resistance to treatments. However, the development of resistance remains difficult to study due to the genetic variability of patients, tumor heterogeneity, and access to material for hypothesis-testing. *In vitro* evolution and whole genome analysis (IVIEWGA) has proven to be a powerful tool for studying the emergence of resistance in microorganisms. Here we demonstrate that IVIEWGA can be used to identify alleles associated with drug resistance in a controlled and reproducible manner using a near-haploid chronic myelogenous leukemia (CML) cell line (HAP-1). We first performed *in vitro* directed evolution using five different anticancer drugs (Doxorubicin, Gemcitabine, Etoposide, Topotecan, and Paclitaxel). We next used whole genome sequencing (WGS) or whole exome sequencing (WES) of 56 independent resistant clones and their isogenic parents to detect ~220 unique gene amplification or deletion events and ~41,000 SNVs and indels, including 641 nonsynonymous variants. Through an optimized variant analysis pipeline that considers resistant clone allele frequency, coding or structural change, copy number variation, and gene background mutation rate we converged on several dozen candidate resistance mutations for further validation. In the set of validation targets we found multiple genes that have been previously associated with compound resistance or resistance to apoptosis including *EGFR*, Topoisomerases (*TOP1, TOP2A*), *MAP3K8, CDK5R2, DCK* as well as transporters (*GRM3, SLC13A4 and ABCB4*). Gene expression knockdown experiments (via both shRNA and CRISPR-*Cas*9) of several validation targets, including deoxycytidine kinase (DCK) and topoisomerase 2 alpha (TOP2A), confirm the role of these genes in drug resistance, demonstrating utility of IVIEWGA as an unbiased screen for mutations associated with drug resistance in tumors. The integration of IVIEWGA and our novel NGS analysis pipeline provides a general framework that can help generate mechanistic insight into drug resistance as well as facilitate the development of improved therapies.

A- A+

PgmNr 279: Driver fusions identified in 1,327 pediatric tumors and their implications in tumorigenesis and precision cancer care.

Authors:
F. Lin [1]; L. Surrey [1]; G. Wertheim [1]; M. Luo [1]; X. Zhao [1]; K. Cao [1]; R. Aplenc [2]; R. Bagatell [2]; A. Bauer [2]; T. Bhatti [1]; S. MacFarland [2]; J. Maris [2]; Y. Mosse [2]; A. Resnick [2]; M. Santi [1]; P. Storm [2]; S. Tasian [2]; A. Waanders [2]; S. Hunger [2]; M. Li [1,2]

View Session Add to Schedule

Affiliations:
1) Department of Pathology & Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA.; 2) Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA

Gene fusions represent one of the most important genomic alterations in cancers. Pediatric tumors have a relatively low mutation burden at diagnosis and higher rate of driver gene fusions when compared to adult tumors. We studied 1,327 consecutive pediatric tumors from patients at or younger than 21 years old using a large custom-designed RNA-Seq panel. The panel interrogates 110 known fusion partner genes and over 600 known fusion transcripts, and allows the detection of novel fusions. Gene fusions are detected in 454 (35%) cases (clinically significant in 437), including 101/376 (27%) CNS tumor, 125/443 (28%) non-CNS solid tumor, and 228/508 (45%) leukemia samples. The most common fusions are *ETV6-RUNX1* and *KMT2A*-fusions in leukemia, *EWSR1-FLI1* in non-CNS solid tumors, and *KIAA1549-BRAF* in CNS tumors. Total 159 unique fusions are identified and nearly half of them are previously unrecognized. Some fusions are pathognomonic of certain cancer, e.g., *PML-RARA* in acute promyelocytic leukemia and *DNAJB1–PRKACA* in fibrolamellar carcinoma. Other fusions are present in a specific group of diseases, such as *KMT2A*-fusions found exclusively in leukemia/lymphomas including B/T-ALL, AML and Burkitt Lymphoma; *EWSR1*-fusions mainly seen in Ewing sarcoma and other non-CNS soft tissue tumors, but also in a few B-ALL samples; *BRAF*-fusions highly enriched in low grade glioma, especially pilocytic astrocytoma (PA), but occasionally in some non-CNS tumors, for instance langerhans cell sarcoma; *RET*-fusions in papillary thyroid carcinoma (PTC) only. In contrast, NTRK-fusions are observed in a variety of tumors including PTC, PA, different soft tissue sarcomas, and both T-ALL and AML. Many driver fusions are mutually exclusive with other driver mutations, such as *BRAF*-fusions vs *BRAF* V600E mutation in low grade gliomas, and *RET*- or NTRK- fusions vs *BRAF* or *APC* mutations in PTC. Fusion genes are ideal therapeutic targets. In our cohort, 200 cases (15%) have a fusion or its downstream pathway targetable, including 83 *BRAF*-fusions, 26 NTRK-fusions, 23 *CRLF2*-fusions, 23 *RET*-fusions, and 23 ABL-fusions. Potentially targetable fusions, such as FGFR-fusions, are also frequently observed. These data demonstrate that this large RNA-Seq-based fusion panel can detect the vast majority of known and certain novel clinically relevant fusions in pediatric cancers accurately and efficiently for patient management. It also offers an excellent tool for research and new fusion gene discovery.

A- A+

PgmNr 280: Drug repositioning opportunities for breast cancer prevention and treatment.

Authors:
G. Chenevix-Trench [1]; J. Beesley [1]; K. McCue [1]; L. Fachal [2]; M. Miranda [1]; M. Spitzer [3,4]; I. Dunham [3,4]; A. Gaulton [5]; A. Peters [6]; I. Azani [6]; G. Monteith [6]; A. Antoniou [2]; D.F.E. Easton [2]; A. Dunning [2]; F. AlEjeh [1]; M. Ghoussaini [4,7]; Breast Cancer Association Consortium and Consortium for Investigators of Modifiers of BRCA1/2

View Session Add to Schedule

Affiliations:
1) QIMR Berghofer, Brisbane, Australia; 2) University of Cambridge, Cambridge, UK; 3) EMBL-EBI, Hinxton, UK; 4) Open Targets, Wellcome Genome Campus, Hinxton, UK; 5) CHEMBL, EMBL-EBI, Hinxton, UK; 6) University of Queensland, Brisbane, Australia; 7) Wellcome Sanger Institute, Hinxton, UK

Genome-wide association studies carried out by the Breast Cancer Association Consortium and the Consortium of Investigators of Modifiers of *BRCA1*and *BRCA2*, have identified approximately 200 loci associated with breast cancer risk. To predict the likely target genes at the GWAS loci, we have developed the INQUISIT (**in**tegrated expression **qu**antitative trait and ***i**n-**s**ilico* prediction of GWAS **t**argets) pipeline to the 7,652 candidate causal variants (CCVs) identified by fine mapping at 206 signals (P_{cond} values<1e-6). This identified 191 high confidence target genes along with a further 115 novel genes identified by transcriptome-wide association studies. To retrospectively validate existing drugs for breast cancer and identify drug repurposing opportunities, we used the Open Targets Platform, a resource that integrates genetic, genomic and drug data to identify and prioritise drug targets. We found 11 genes are targets for 44 drugs (both approved or in clinical trials) for breast cancer. These genes include *ESR1, CCND1, ALK1, FGFR2* and *TERT*. In addition 62 drugs targeting another eight high confidence genes and used for other diseases could potentially be re-positioned for primary or secondary prevention of breast cancer. These include Senicapoc, a small molecule inhibitor of KCNN4 which has shown a good safety profile in Phase 3 and 4 clinical trials for treating osteoarthritis and pain. *KCNN4* maps to 19q13.31, a locus which contains ten CCVs for breast cancer (*P*<3e-17 for estrogen-positive and *P*<7e-08 for estrogen-negative breast cancer respectively) for which INQUISIT prioritises three target genes: *KCNN4, PLAUR* and *SMG9*. We have not observed allele-specific differences for the putative *KCNN4* regulatory element using luciferase assays in breast cell lines. However, eQTL analyses strongly support *KCNN4* as the target gene in blood (rs56681946, *P*=5.23e-33 in GTEx, with the risk allele associated with increased expression) but not breast tissue, suggesting a potential role in immunosurveillance. We therefore investigated the use of Senicapoc for the treatment of triple negative breast cancer. Combining Senicapoc with chemotherapy significantly reduced growth of both the human MDA-MB-231 and murine MT1 cell lines in nude and immunocompetent mice (*P*<1e-04). Future experiments will explore the use of Senicapoc and the other compounds for primary prevention of mammary carcinoma in immunocompetent mouse models.

A- A+

PgmNr 281: Polygenic risk for skin autoimmunity impacts immune checkpoint blockade in bladder cancer.

Authors:
Z. Khan [1]; F. Di Nucci [1]; A. Kwan [1]; C. Hammer [1]; S. Mariathasan [1]; V. Rouilly [1]; J. Carroll [1]; M. Fontes [1]; S. Ley Acosta [1]; E. Guardino [1]; H. Chen-Harris [1]; T. Bhangale [1]; J. Rosenberg [2]; T. Powles [3]; J. Hunkapiller [1]; G.S. Chandler [1]; M.L. Albert [1,4]

View Session Add to Schedule

Affiliations:
1) Genentech, South San Francisco, CA 94080, USA; 2) Genitourinary Oncology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; 3) Barts Experimental Cancer Medicine Centre, Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK; 4) Current Address: insitro, South San Francisco, CA 94080, USA

PD-1 checkpoint inhibitors have made significant advances in the treatment of metastatic urothelial carcinoma (mUC), and tumor proximal measures of pre-existing immunity and neo-antigen load have been associated with response and longer survival. PD-1 and PD-L1 also play a role in immune regulation and peripheral tolerance to self-antigens. Consistent with this role, checkpoint inhibitors have been associated with toxicities called immune-related adverse events (irAEs), which are thought to arise due genetic and environmental factors. Analyses of dermatological irAEs indicate they are associated with improved overall survival (OS) following anti-PD-(L)1 therapy, however the relationship of these observations to immune activation or dermatological autoimmunity is unknown. We collected whole genome germline sequencing data from a subset of patients from IMvigor211, a phase 3 randomized controlled trial comparing atezolizumab (anti-PD-L1) monotherapy to chemotherapy for treatment of mUC (N=238 received atezolizumab; N=227 received chemotherapy) where pre-treatment tumor RNA-seq, tumor mutation burden, and PD-L1 immunohistochemistry measurements were also available. Using publicly available genome-wide association study (GWAS) summary statistics, we constructed patient level polygenic risk scores (PRSs) for dermatological autoimmune diseases and associated them with irAEs, OS, and tumor gene expression, adjusting for baseline covariates and genotype eigenvectors. We found that polygenic risk for psoriasis was associated with increased odds of skin irAEs (p=0.002; OR 1.79; 95% CI 1.24-2.40). High vitiligo (p=0.0016; HR 0.58; 95% CI 0.41-0.81), high psoriasis (p=5.5e-5; HR 0.50; 95% CI 0.36-0.70), and low atopic dermatitis (p=0.0008; HR 0.57; 95% CI 0.41-0.79) polygenic risk were predictive of longer OS under anti-PD-L1 monotherapy compared to chemotherapy, reflecting the Th17 polarization of these diseases. PRSs were uncorrelated with pre-treatment tumor proximal measures, establishing them as independent, predictive biomarkers. Our analysis suggests that shared genetic factors impact risk for dermatological autoimmunity and survival during anti-PD-L1 monotherapy in bladder cancer.

A- A+

PgmNr 282: Multi-trait analysis of skin cancer and related traits identifies dozens of novel loci for cutaneous melanoma, including at *CTLA-4* a key melanoma immunotherapy target.

Authors:
S. MacGregor [1]; J. Shi [2]; D.T. Bishop [3]; E. Nagore [4]; A.J. Stratigos [5]; M.C. Fargnoli [6]; P. Ghiorzo [7]; K. Peris [8]; A.E. Cust [9]; J. Han [10]; C. Olson [11]; D. Schadendorf [12]; G.J. Mann [13]; G.L. Radford-Smith [14]; N.G. Martin [15]; C. Hayward [16]; N.K. Hayward [17]; G.W. Montgomery [18]; S.V. Ward [19]; P.D.P. Pharoah [20]; C.I. Amos [21]; M. Zawistowski [22]; S. Puig [23]; D.L. Duffy [24]; F. Demenais [25]; K.M. Brown [2]; D.C. Whiteman [11]; M.T. Landi [2]; M.M. Iles [4]; M.H. Law [1]

View Session Add to Schedule

Affiliations:
1) Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Australia; 2) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; 3) Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK; 4) Department of Dermatology, Instituto Valenciano de Oncología, València, Spain; 5) 1st Department of Dermatology – Venereology, National and Kapodistrian University of Athens School of Medicine, Andreas Sygros Hospital, Athens, Greece; 6) Department of Dermatology, University of L’Aquila, L’Aquila, Italy; 7) Department of Internal Medicine and Medical Specialties, University of Genoa and Genetics of Rare Cancers, Ospedale Policlinico San Martino, Genoa, Italy; 8) Institute of Dermatology, Catholic University, Rome, Italy; 9) Cancer Epidemiology and Prevention Research, Sydney School of Public Health and Melanoma Institute Australia, the University of Sydney, Sydney, New South Wales, Australia.; 10) Department of Epidemiology, Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana, USA Melvin and Bren Simon Cancer Center, Indiana University, Indianapolis, Indiana, USA; 11) Population Health, QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 12) Department of Dermatology, University Hospital Essen, Hufelandstrasse 55, 45122 Essen, Germany.; 13) Centre for Cancer Research, Westmead Institute for Medical Research; Melanoma Institute Australia; University of Sydney; 14) Inflammatory Bowel Diseases, QIMR Berghofer Medical Research Institute, Brisbane, Australia Department of Gastroenterology and Hepatology, Royal Brisbane & Women’s Hospital, Brisbane, Australia University of Queensland, Brisbane, Australia.; 15) Genetic Epidemiology, The QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 16) MRC Human Genetics Unit, University of Edinburgh, Institute of Genetics and Molecular Medicine, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK.; 17) Oncogenomics, QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 18) Molecular Biology, the University of Queensland, Brisbane, Australia.; 19) Centre for Genetic Origins of Health and Disease (GOHaD), University of Western Australia. Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center; 20) Department of Oncology, University of Cambridge, Cambridge, United Kingdom; 21) Department of Medicine, Baylor College of Medicine, Houston, TX.; 22) Center for Statistical Genetics, University of Michigan, Ann Arbor MI, USA; 23) Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain. & Centro de Investigación Biomédica en Red en Enfermedades Raras (CIBERER), Valencia, Spain; 24) Genetic Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, Australia.; 25) Team of Genetic Epidemiology and Functional Genomics of Multifactorial Diseases, Inserm UMR-1124, Université de Paris, Paris, France

The skin cancers cutaneous melanoma (CM) and keratinocyte carcinoma (KC, comprising basal cell carcinoma and squamous cell carcinoma) have key risk factors in common. We exploit this for gene identification by leveraging large scale GWAS of CM, KC and their risk factors such as skin colour and burn type. The GWAS used included an unpublished clinically confirmed CM meta-analysis GWAS (30,134 cases and 81,415 controls), 3 KC GWAS datasets from Australia, United Kingdom and the United States (28,831 cases and 278,191 controls) and multiple GWAS of skin cancer related traits including skin colour, skin burn type, hair colour, freckling and facial ageing (up to 448,288 individuals from UK Biobank and Australian studies).

Using LD score regression we first demonstrate substantial correlations between these traits (e.g. 0.48 between CM and KC, 0.47 between CM and freckle count). Given these correlations we apply multi-trait analysis of GWAS (MTAG), a generalisation of inverse-variance-weighted meta-analysis that accounts for incomplete genetic correlation between traits. MTAG produces trait-specific SNP estimates for each input trait, and here we focus on CM. Functional Mapping and Annotation of GWAS (FUMA) and the OpenTargets platform were used to identify and annotate independent P<5e-8 SNPs.

We benchmarked our multi-trait analysis against the standard univariate CM analysis which identified 47 loci. A multi-trait analysis with CM and KC increased the number of genome-wide significant loci for CM to 57. Adding in the skin cancer related traits took the number of CM specific loci to 70. New CM hits included rs231724 which is close to and an eQTL for *CTLA-4*, a key immunotherapy target. rs231724 is associated with hypothyroidism in UK Biobank, as are two of the other novel CM loci from MTAG (rs2030519 near *LPP* and rs78456138 near *FAP*). Among the other novel loci are SNPs associated with breast cancer (rs4149909) and obesity (rs34517439, rs4714520, rs9574159, rs201475383).

Our analysis indicates genetic correlation can be leveraged to identify new risk genes for CM. At the meeting we will also discuss the impact of adding further traits, including vitamin D levels, risk propensity and autoimmune traits.

A- A+

PgmNr 283: *mantis-ml*: Stochastic semi-supervised learning to prioritise genes from high-throughput genomic screens.

Authors:
D. Vitsios; S. Petrovski

View Session Add to Schedule

Affiliation: Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK

Access to large-scale genomics datasets has increased the utility of hypothesis-free genome-wide analyses that result in candidate lists of genes. Often these analyses highlight several gene signals that might contribute to pathogenesis but are insufficiently powered to reach experiment-wide significance. This often triggers a process of laborious evaluation of highly-ranked genes through manual inspection of various public knowledge resources to triage those considered sufficiently interesting for deeper investigation. Here, we introduce a novel multi-dimensional, multi-step machine learning framework to objectively and more holistically assess biological relevance of genes to disease studies. To identify among candidate gene lists those that share important characteristics with disease-associated genes, we rely on a plethora of gene-associated annotations (including: Human Phenotype Ontology (HPO), Exome Aggregate Consortium Dataset (ExAC), Genotype-Tissue Expression (GTEx), model organism phenotype databases, genic-intolerance metrics, and many more). We developed *mantis-ml* to serve as an automated machine learning (AutoML) framework, following a stochastic semi-supervised learning approach to rank known and novel disease-associated genes through iterative training and prediction sessions of random balanced datasets across the protein-coding exome (n=18,626 genes). We applied this framework on three major disease groups: amyotrophic lateral sclerosis (ALS), chronic kidney disease (CKD) and epilepsy, achieving an average Area Under Curve (AUC) prediction performance of 0.85. We also created a Generic Mantis-ML Score (GMS; non disease-informed) that can distinguish OMIM disease-associated genes from the rest of the exome, outperforming published gene-level scores. Critically, to demonstrate applied utility on exome-wide association studies, we overlapped *mantis-ml* disease-specific predictions with data from published cohort-level association studies. We retrieved statistically significant enrichment of high *mantis-ml* predictions among the top-ranked genes from hypothesis-free cohort-level statistics ($p < 0.05$), suggesting the capture of true prioritisation signals. We believe that *mantis-ml* is a novel easy-to-use tool to support objectively triaging gene discovery and overall enhancing our understanding of complex genotype-phenotype associations.

A- A+

PgmNr 284: After F-measure: Deep learning marginal variants from largest-scale cohorts.

Authors:
W. Salerno [1]; X. Bai [1]; E. Maxwell [1]; P. Chang [2]; O. Krasheninina [1]; L. Habegger [1]; A. Carroll [2]; J.G. Reid [1]

View Session Add to Schedule

Affiliations:
1) Regeneron Genetics Center, Regeneron, New York, New York.; 2) Google AI, Google, Mountain View, CA

Genomic sequencing of hundreds of thousands of samples is now routine in programs such as UK Biobank, Geisinger Health System and TOPMed, with aggregate sizes rapidly approaching millions. The marginal value of sequencing at these ultra-large scales is the improved characterization of variation rare or absent at smaller scales. However, because common assessments of variant-calling performance are limited to few samples and select genomic regions, variant-calling methods optimized to these standards do not reflect the full spectrum of variation. While deep learning methods have proven successful at modeling at-scale complexity in other disciplines, they require extensive training on data sets of appropriate size and heterogeneity.

Here we present a large-scale curriculum for deep learning genomic variants. Drawing from more than 500,000 exomes sequenced via a highly automated NovaSeq pipeline, we minimize sample-to-sample variability as a potential confounding feature. Samples are mapped to multiple versions of the GRCh38 reference, and variants called with multiple compositional methods, allowing us to isolate training genotypes that are sensitive to primary analysis protocols. With 4,000 trios and 500 replicate pairs, we identify high-confidence genotypes in low-complexity and repetitive genomic regions. From these features we can define classes of marginal variants with sufficient representation for deep learning. With the DeepVariant framework these variant classes serve as the foundation for a comprehensive variant model optimized for highly-automated NovaSeq WES data at unprecedented scale.

A- A+

PgmNr 285: SV genotyping in large population cohorts utilizing sequence graphs.

Authors:
S. Chen [1]; E. Dolzhenko [1]; A.M. Gross [1]; B.R. Lajoie [1]; P. Krusche [2,3]; R. Petrovski [2]; R.M. Sherman [4]; F. Schlesinger [1]; D.R. Bentley [2]; M.C. Schatz [4,5]; F.J. Sedlazeck [6]; M.A. Eberle [1]

View Session Add to Schedule

Affiliations:
1) Illumina Inc., San Diego, CA, USA; 2) Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, UK; 3) Novartis Pharma AG, Basel, Switzerland; 4) Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA; 5) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA; 6) Baylor College of Medicine Human Genome Sequencing Center, Houston, TX, USA

Structural variations (SVs) represent the majority of person-to-person genomic variability (as measured by total differing base) and are associated with many human diseases. SVs can deviate significantly from the reference genome, often causing reads from these regions to be lost or misplaced during the standard alignment process. To address this, a graph representation has been proposed to explicitly incorporate deviations from the reference resulting in greater accuracy compared to a linear reference.

Here, we introduce our graph genotyper, Paragraph, that performs a local realignment of reads to a graph representation of the genome that includes known SVs. Paragraph can genotype both insertions and deletions and achieved a recall of 82% when tested against a high-quality SV truth set built from long-read sequence data. Furthermore, Paragraph is compatible with complex variant structure, facilitating a broader set of use-cases.

While a comprehensive SV catalog is currently not available, we have started to create a high-quality catalog using SVs identified in four genomes that were sequenced using the PacBio HiFi technology which generates highly-accurate reads with length longer than 10,000 base pairs. As new SVs are identified and added to this reference catalog, we genotype them across 2,504 samples of the 1000 Genome Project to perform additional annotation across multiple populations. For each SV, we report individual genotypes as well as population-level statistics. With this data, we calculate population measures such as HWE p-values, which allow us to identify genotyping errors and improve the performance of Paragraph. In our preliminary analysis of a subset of the data, we identified 155 SVs with potential impact on gene expression, and with the population allele frequencies, we observed a selection pressure on SVs from functional regions of the genome.

We demonstrate that Paragraph is well suited for genotyping SVs from the population, and the population-statistics can, in turn, provide insights into complicated regions of the genome. We expect that this expanding graph reference database will contribute to the characterization of the variant landscape in population, and ultimately help diagnose pathogenic SVs in clinical sequencing.

A- A+

PgmNr 286: Leveraging a cell-line cohort as a clinical reference panel: Analysis of high coverage 1000 genomes sequencing data.

Authors:
A.M. Gross [1]; B.R. Lajoie [1]; D.R. Bentley [2]; M.A. Eberle [1]; R.J. Taft [1]

View Session Add to Schedule

Affiliations:
1) Illumina Inc., 5200 Illumina Way, San Diego, CA, USA; 2) Illumina Cambridge Ltd., Illumina Centre 19 Granta Park, Great Abington, Cambridge, UK

Here we describe a comprehensive assessment of the recently released 1000 Genome Project sequencing data (WGS at ~30x coverage, n=2504 lymphoblastoid cell lines), and their use as a reference panel within a clinical WGS laboratory. Global analyses revealed genome-wide biases in these samples corresponding to genomic regions with early and late replication timing, resulting in lower uniformity of depth of coverage than DNA extracted from whole blood. To remedy this, we developed a method for quantification and adjustment of these biases, lowering systematic variation in the depth signal (2%, 18%, and 46% decrease in genome-wide variance for 1kb, 10kb, and 100kb depth bins, respectively), and creating a coverage background suitable for comparison with whole-blood based data. In addition, we conducted quality control of the samples and identified nearly two hundred large anomalies that represent cell line artefacts. Specifically, we identified aneuploidies on sex chromosomes (n=45) and autosomes (n=24), as well as large copy number variants (n=105 variants spanning more than one chromosomal band) and uniparental isodisomy (n=20). Using such anomalous samples in a population background may skew calculations of allele frequency when genotyping copy number variants. Furthermore, over-reporting of homozygous variants in such regions may alter interpretations of variant consequence in some cases.

Despite these limitations, we find that, with sample filtering, these data from the 1000 Genomes Project are a valuable reference resource, providing a diverse and publicly available dataset of presumably healthy individuals. Furthermore, the anomalous samples present a resource for testing genomes with germline and mosaic aneuploidy, uniparental disomy, and large CNVs. These samples mimic clinically relevant disorders that are not widely available to many laboratories, software or bioinformatic tool developers. We propose that wider use of such samples may yield improvements in variant calling and genotyping within regions of non-diploid copy-number or non-syndromic uniparental disomy. To illustrate this point, we highlight a recent case with a mosaic X-chromosome deletion indicative of Turner Syndrome detected by clinical WGS and similar to sex chromosome aneuploidies found in the 1000 Genomes Project data. In addition to the aneuploidy, the proband displayed a maternally inherited exon duplication on the DMD gene, resulting in a complex X-linked recessive inheritance.

A- A+

PgmNr 287: Prescription medication data improves genetic discoveries in EHR-based studies.

Authors:
T. Kiiskinen [1]; A.S. Havulinna [1,2]; J. Karjalainen [1,3,4]; M. Kurki [1,5,6]; S. Lemmelä [1,2]; N.J. Mars [1]; V. Salomaa [2]; H. Laivuori [1,7,8]; M. Daly [1,3]; A. Palotie [1,3,6]; S. Ripatti [1,3,9]; FinnGen

View Session Add to Schedule

Affiliations:
1) Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland; 2) National Institute for Health and Welfare, Helsinki, Finland; 3) The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 4) Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA; 5) Program in Medical and Population Genetics and Genetic Analysis Platform, Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 6) Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Analytic and Translational Genetics Unit, Department of Medicine, and the Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA; 7) Department of Obstetrics and Gynaecology, Tampere University Hospital, and Tampere University, Faculty of Medicine and Health Technology, Tampere, Finland; 8) Medical and Clinical Genetics, University of Helsinki and Helsinki University Hospital, Helsinki, Finland; 9) Department of Public Health, Clinicum, Faculty of Medicine, University of Helsinki, Helsinki, Finland

The FinnGen study aims to create a dataset of ~500,000 genotyped Finns (~10% of the population) connected to electronic health records (EHRs) by unique personal identification numbers. Using current data (n~140,000) we created algorithms for a high-resolution tree-structured hierarchical phenotype map of over 2,000 medical conditions for genome-wide and phenome-wide association studies.

The algorithms consider the unique properties of the nationwide Finnish EHRs, which record every healthcare visit over an individual's lifetime, but contain no symptom-level nor quantitative data. Harmonization not only over ICD revisions 8, 9 and 10 in hospital discharge and death registries, procedure codes, and cancer-registry data, but also the Social Insurance Institution drug purchase and reimbursement data, allows us to study the longitudinal phenotypes since 1952.

Medication data offers significant added value. First, medication data helps in defining conditions that do not cause extended hospitalization nor death. Including triptan purchases in the definition of migraine increased the case count from 5,123 to 7,520 and previous genome-wide hits from 0 to 4. In comparison, while migraine in UKBB with ICD-codes found only 2,870 cases (sample size=401,650) and no genome-wide hits.

Second, using validated reimbursement data on prescription medicine lowers the rate of false-positive cases: attempts to replicate previous GWAS-hits in inflammatory bowel disease were improved by including only reimbursed cases (rs1078650, $p=4.5*10^{-8}$, 11:76688605_C/A, $p=1.81*10^{-7}$, neither replicated without the reimbursement requirement).

Last, reimbursement data allows inclusion of validated true positives and efficient exclusion of false ones. Defining type 2 diabetes (T2D) without medication data resulted in 10,703 cases and 18 genome-wide top hits. However, one of these corresponded to the known type 1 diabetes (T1D)-associated HLA-DQA2 locus ($p=9.5*10^{-38}$). Inclusion of reimbursed T2D cases with exclusion of all T1D events (17,616 cases) resulted in disappearance of the T1D-associated HLA-DQA2 (association). With the refined T2D definition, we observed associations in 34 previously known loci and found 4 new loci driven by considerably Finland-enriched variants.

In conclusion, our phenotyping strategy with medication data allows more comprehensive and precise morbidity definitions in large datasets, enabling improved discovery in large-scale genomic studies.

A- A+

PgmNr 288: Genome-wide survey of parent-of-origin effects on EHR-derived quantitative traits in 90k DiscovEHR cohort.

Authors:
H. Kim; J. Staples; A. Marcketta; B. Ye; C. Gao; A. Shuldiner; C. Van Hout; RGC-Geisinger DiscovEHR Collaboration

View Session Add to Schedule

Affiliation: Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY

Parent-of-origin (PoO) effects refer to the differential effects of genes or variants on traits depending on their parental inheritance and can result from imprinting. While genetic variants in imprinted regions may affect complex traits in a PoO specific manner, these effects may not be captured by models that do not differentiate the PoO. The aim of our study was to screen for PoO effects in the DiscovEHR cohort comprised of >90k individuals with exome sequence, array genotypes, and medical traits extracted from the electronic health records (EHR).

Based on pairwise kinship estimates calculated from genetic data, we identified ~12,600 parent-offspring relationships. We merged exome and array genotypes and restricted the analysis to ~1.4M variants with minor allele count ≥ 3 among offspring. The PoO of variants was assigned using two methods. When possible, we assigned PoO based on evident Mendelian segregation of the parental and offspring genotypes. We also estimated PoO by comparing parental and offspring haplotypes around the variant, which had >98% concordance rate when compared to Mendelian method.

Next, we performed genome-wide PoO association analysis for 291 EHR-derived quantitative traits including lab test results and biometric measures. We tested the association of paternally and maternally inherited alleles first in 60k subjects for discovery and subsequently in 30k subjects for replication. We found four PoO-specific locus-trait associations that met genome-wide significance ($P < 5x10^{-8}$) that replicated with consistent effect directions. Two of these were at the known loci with previously unknown PoO effects: the paternal-specific association of a noncoding variant at the *APOB* locus with total cholesterol ($b_{PAT} = -0.22$, $P_{PAT} = 2.25x10^{-8}$ / $b_{MAT} = -0.04$, $P_{MAT} = 0.27$) and LDL-C, and the maternal-specific association of two noncoding variants near *HFE* with mean corpuscular hemoglobin ($b_{MAT} = 0.32$, $P_{MAT} = 7.28x10^{-10}$ / $b_{PAT} = 0.10$, $P_{PAT} = 0.06$). Two novel associations were the maternal-specific association of two noncoding variants in *ZFAND5* with percent neutrophils ($b_{MAT} = -1.65$, $P_{MAT} = 2.18x10^{-8}$ / $b_{PAT} = 0.17$, $P_{PAT} = 0.56$), and the maternal-specific association of a noncoding variant near *CD177* with albumin fraction ($b_{MAT} = -0.17$, $P_{MAT} = 4.88x10^{-8}$ / $b_{PAT} = -0.01$, $P_{PAT} = 0.68$). These results show that our approach has the potential to uncover previously unknown PoO effects of known associations and to identify novel associations with PoO effects.

A- A+

PgmNr 289: UK Biobank participants that live more than 20 km from their birthplace have higher socioeconomic and health correlates.

Authors:
I. Woods [1,2]; A. Williams [2]

View Session Add to Schedule

Affiliations:
1) Biology, Ithaca College, Ithaca, NY.; 2) Computational Biology, Cornell University, Ithaca, NY

The UKB Biobank (UKB) is a rich repository of genotype and phenotype data comprising nearly half a million people. The range of phenotypes assessed is especially diverse, and includes linked health records, physiological measurements, and demographic and lifestyle data. We sought to characterize demographic characteristics of the UKB samples and to examine phenotypic correlations, focusing especially on patterns of migration. Over 90% of the participants in the project lived within 25km of one of the 21 assessment centers at the time of sampling, indicating highly localized sampling of participants. However, only ~56% of the participants were born within 25km of their assessment center. Based on place of birth (POB) and place of residence (POR) at time of sampling, we partitioned the dataset into 'stayers' and 'movers', with a threshold POB-POR distance of 20km. We found that mover status is correlated with a range of phenotypes, including numerous positive physical health outcomes, higher educational attainment, and improved financial situation, whereas some measurements of social connectivity and environmental quality were lower among movers. These effects largely persisted in families: children of movers showed similar phenotypic patterns, even if they were not movers themselves. The degree of phenotypic correlation suggests that movers may be a sample of individuals that are not representative of others from their birthplace. This observation has practical importance for genome-wide association studies (GWAS). Because alleles vary with geography, GWAS may overestimate the effects of variants present at especially high or low frequencies within movers. Here, we test for potential confounding effects of migration in a GWAS for educational attainment. When including mover status as a covariate, we observe a decrease in the number of genome-wide significant SNPs, and effect estimates for significant SNPs decreased by ~20%. These results highlight the complicated nature of the UK Biobank cohort and stress the importance of careful analyses of demographic factors that influence localized sample collections.

A- A+

PgmNr 290: Large-scale identity-by-descent mapping in a biobank with more than 95,000 individuals identifies novel genome-wide significant regions associated with serum lipid levels.

Authors:
H.-H. Chen; L.E. Petty; Q.S. Wells; J.E. Below

View Session Add to Schedule

Affiliation: Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee.

Genomic segments shared due to a recent common ancestor can be detected using identity by descent (IBD), and represent an untapped opportunity to identify genes that harbor low frequency, large effect variants which were undetectable in previous genome-wide association studies (GWAS) due to low power at rare and heterogeneous loci. Abnormal lipids profile, abnormalities in one or more of total cholesterol (TC), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), and high-density lipoprotein cholesterol (HDL-C), has an estimated 53% prevalence in U.S. population and is an important risk for further cardiovascular disease. The high heritability of serum lipids, estimated from 0.35-0.64, has motivated many large-scale GWAS to identify genes impacting serum lipid levels. Although numerous lipids associated genes have been identified, much of heritability of serum lipids is still unexplained. The Vanderbilt biobank (BioVU) comprises over 280,000 DNA samples from participants with linked electronic medical record (EMR). In this study, we extracted serum lipids measures from participants' EMR and determined all pairwise IBD shared segments in more than 95,000 genotyped BioVU participants using GERMLINE on Illumina Multi-Ethnic Global Array data. Shared segments >3cM were used to identify regions of enriched genomic sharing associated with serum lipids. Our analysis approach assessed enrichment of both individual segments or any shared segments on lipids level as quantitative traits, genome-wide. In a preliminary subset (N=14,150), we found 536 IBD segments significantly associated with TC (p-value<5×10^{-8}), 5640 for TG, 405 for LDL, and 29 for HDL. Notably, one TC associated segment on chromosome 5 (p-value=1.2×10^{-8}, carrier=12) harbors 5 genes: *SNX13, LOC101927630, ELFN1, PRPS1L1*, and *ELFN1-AS1*, and *SNX13* has been well-known for its effect on HDL. Another IBD segment on chromosome 16 harbors only one gene, *POLR3K*, and is significantly correlated with TG (p-value=3.4×10^{-8}, carrier=29). *POLR3K* has not been previously reported in GWAS of serum lipids. Final analyses of all available BioVU genotyped data is on-going (N>95,000), and will include causal variant mapping via complete sequencing of risk-associated shared segments in carriers. This study leverages an underutilized characteristic of genomes- segments shared due to relatedness- to map novel genes impacting serum lipid, and preliminary results implicate both novel and known candidate lipid genes.

A- A+

PgmNr 291: Phenome-wide association study of loss of function variants in 128,382 UK Biobank whole exome sequences.

Authors:
M.M. Parker; L.D. Ward; G. Hinkle; P. Nioi

View Session Add to Schedule

Affiliation: Alnylam Pharmaceuticals, 300 3rd Street, Cambridge, Massachusetts.

Objective
To characterize the association of rare loss of function variants in UK Biobank with a broad spectrum of clinically relevant phenotypes.

Methods
We performed a large-scale phenome-wide association study of 8,361 predicted loss of function variants identified through exome sequencing of 128,382 white participants of UK Biobank. We tested a total of 658 phenotypes, including both measured quantitative traits and ICD10 diagnosis codes. Analyses were performed in PLINK using linear or logistic regression controlling for age, sex, and genetic ancestry.

Results
We identified a total of 8,361 predicted loss of function variants with a minor allele frequency (MAF) between 0.25% and 1%, including 2,544 stop-gained, 2,628 splice-altering and 3,189 frameshift mutations. Almost every gene has a least one loss of function variant (98%), and the average gene has 27 (range: 0 – 3,396). A total of 4,066 variants from 2,416 genes were homozygous loss of function (“knockouts”), distributed as 1,954 single observation, 1,368 rare (MAF < 1%), and 744 common variants (MAF > 1%). Association analysis revealed a total of 737 significant associations of loss of function variants with quantitative phenotypes or ICD10 diagnosis codes, of which 51 are rare variants (MAF between 0.025% and 1%). These include replication of many established genetic associations, in addition to novel large effect associations of loss of function variants with disease outcomes.

Conclusions
Using a phenome-wide association study of loss of function variants, we demonstrate the utility of focusing on rare protein-altering variants to identify novel disease associations and potential therapeutic targets. Results of this analysis will aid the advancement of genetically defined medicines like RNAi therapeutics.

A- A+

PgmNr 292: An eQTL analysis of RNA-seq data from 13,175 individuals using a personalized transcriptome.

Authors:
G.H. Halldorsson [1,2]; B. Gunnarsson [1]; R.L. Gudmundsson [1]; B.V. Halldorsson [1,4]; S.A. Gudjonsson [1]; D.F. Gudbjartsson [1,2]; P. Sulem [1]; O.TH. Magnusson [1]; U. Thorsteinsdottir [1,3]; K. Stefansson [1,3]; P. Melsted [1,2]

View Session Add to Schedule

Affiliations:
1) deCODE Genetics/Amgen, Reykjavik, Iceland; 2) School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland; 3) Faculty of Medicine, University of Iceland, Reykjavik, Iceland; 4) School of Science and Engineering, Reykjavik University, Reykjavik, Iceland

Understanding the effects of sequence variants on gene expression is crucial for understanding the role of variants, especially in the context of GWAS results. While GTEx has provided an invaluable resource for the community, the focus has been on covering a large number of tissues assayed. In this study we expand on the number of individuals sequenced as well and the number of variants tested.

We report on an eQTL pipeline to analyze RNA-sequencing data from whole blood of 13,175 individuals. Using familial imputation, the variants of each individual were fully phased into a maternal and paternal copy of each chromosome. Using this resource we constructed a personalized transcriptome for each individual, representing the two distinct copies of each gene inherited from each parent. RNA-seq data was quantified using this personalized transcriptome, reducing the reference bias, as well as enabling allele specific expression analysis.

The cis-eQTL association analysis used 33M variants and tested 19,256 genes expressed in blood using a 5Mb window flanking each gene. The large sample size yields significant power to detect minor effects of variants on expression. We found that 17.5M variants are a cis-eQTL for some gene. We find that 95% of the 10,618 genes with expression of TPM greater than 1.0, have at least one significant eQTL.

To aid the interpretability of the results we implemented an iterative conditional analysis scheme. This method cleans up associations that are not causal and most likely due to linkage disequilibrium between markers. On average each gene had 4.3 independently significant eQTL markers.

For trans-eQTL analysis we tested for association using a reduced set of 1.8M markers, by taking the LD structure into account. We find 348,279 trans-eQTLs in 10,222 genes (p-value threshold 2.5e-8) with an average number of 36.1 trans-eQTLs per gene.

Finally we demonstrate the advantage of having fully phased haplotypes and haplotype specific quantification of RNA-seq data by examining cases of eQTL markers where only heterozygous carriers are available. In these cases the allele specific expression can be estimated for markers that are not directly observed in the RNA-seq data which gives an increase in power to detect eQTLs compared to standard eQTL analysis.

A- A+

PgmNr 293: Phenome-wide association study using the EHR-linked BioVU biobank links GREM2 variant Q76E to high risk of anemia.

Authors:
A.K. Hatzopoulos; D.H. Wu; E. Farber-Eger; Q.S. Wells

View Session Add to Schedule

Affiliation: Medicine, Vanderbilt University Medical Center, Nashville, TN.

We have previously shown that the human Q76E variant (rs142343894) of the Bone Morphogenetic Protein (BMP) signaling antagonist GREMLIN 2 (GREM2), with minor allele frequency (MAF) of ~0.004, is present with higher frequency in probands of families with lone atrial fibrillation. The substitution of glutamine (Q) with glutamic acid (E) inserts a negatively charged moiety within a positively charged interface that is critical for heparin binding and BMP signaling regulation, thus interfering with the inhibitory capacity of GREM2. BMP signaling has fundamental roles in many developmental processes and diseases, raising the possibility that the Q76E variant has broad pathogenic potential. To test this hypothesis, we performed a Phenome-Wide (PheWAS) association study using *BioVU,* the Vanderbilt University Medical Center *biobank* of de-identified DNA samples and genetic data linked to electronic health records (EHR). Our results show that Q76E carriers have both higher incidence and increased severity of inflammatory anemias than non-carrier, age- and gender-matched controls. The risk is particularly striking in cancer patients, with carriers having approximately 2.1 times higher odds ratio to develop anemia compared to cancer patients with normal GREM2. Q76E cancer patients have low neutrophil values, suggesting susceptibility to infections. Analysis of *Grem2*$^{-/-}$mice shows that GREM2 regulates the growth and differentiation of bone marrow hematopoietic stem and progenitor cells (HSPCs). Our findings indicate that GREM2 has an important role in the regulation of bone marrow hematopoiesis and that aberrant GREM2 activity caused by the Q76E variant perturbs generation of blood cells, increasing the risk of anemia. Our findings might guide the development of personalized therapeutic strategies for patient populations with compromised GREM2 activity and abnormal regulation of BMP signaling.

PgmNr 294: Using the history of biochemistry to illuminate the genetic architecture of metabolite GWAS and its implications for complex human phenotypes.

Authors:
E.B. Fauman [1]; P. Surendran [2]; I.D. Stewart [3]; L.A. Lotta [3]; K. Suhre [4]; G. Kastenmuller [5]; J. Danesh [2]; N.J. Wareham [3]; A. Butterworth [2]; C. Langenberg [3]

View Session Add to Schedule

Affiliations:
1) Integrative Biology, Internal Medicine Research Unit, Pfizer, Cambridge, Massachusetts.; 2) Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; 3) MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom; 4) Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Qatar-Foundation; 5) Helmholtz Centre Munich, Germany

---

Introduction: The utility of GWAS is frequently limited by the need for additional functional genetics studies to identify the likely causal genes. However, where the phenotype reflects the level of a specific metabolite, the follow-up experiment one might envision has often already been performed under the auspices of classic biochemistry.

Methods: We included metabolite lead SNPs from the two largest ongoing untargeted and cross-platform metabolite GWAS meta-analyses as well as 189 published studies, together constituting over 5000 associations for over 600 metabolites. We combined a systematic search of structured prior knowledge as contained within databases such as HMDB, OMIM, Uniprot, Entrez Gene and the GO ontology with expert manual curation to provide definitive causal gene annotations for as many loci as possible.

Results: This systematic curation has produced a “gold standard” set of 350 causal genes each with prior experimental validation. The experimental evidence precedes the genetic study by an average of 20 years. At 275 loci the deduced causal gene is the gene closest to the lead SNP. Genes mediating metabolite phenotypes include enzymes, transporters, solute carriers and transcription factors. With the clarity of assigned causal genes we can identify allelic series within a single gene affecting non-overlapping and distinct subsets of functions. For example, within the TTR gene (Transporter of Thyroxine and Retinol) we identify one SNP influencing levels of retinol but not thyroxine and a conditionally independent SNP influencing level of thyroxine but not retinol. This has implications for the use and interpretation of PheWAS analyses in general in that variants acting on a specific gene may not uniformally impact all the functions of that gene.

Conclusion: The legacy of classic biochemistry provides a robust framework for the interpretation of genetic effects on metabolites, with lessons for the interpretation of the genetic architecture of human phenotypes more generally.

A- A+

PgmNr 295: Universal LD blocks in the human genome.

Authors:
S. Christensen [1,2]; J.N.J. McManus [1]

View Session Add to Schedule

Affiliations:
1) Kallyope, Inc., 430 East 29th Street, New York, NY; 2) Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL

Genetic variants in the human genome have a well-known statistical structure, whereby linkage disequilibrium (LD) blocks delineate statistical correlations between polymorphisms. The correlations circumscribed by LD blocks, and their differences across populations, are fundamental to the statistical fine-mapping of genome-wide association studies (GWAS). Previous work has described haplotype blocks in the genome, which are crucial for imputation but less relevant for fine-mapping, or used a heuristic to find approximate LD blocks of an assumed, predetermined size. Here, we undertake the first unbiased, statistically rigorous partition of the human genome into LD blocks, without making any prior assumptions on the LD structure. We apply a novel graph clustering algorithm, which finds the global maximum of an optimization criterion, to delineate all of the LD blocks in the genome, separately for each of the populations sequenced by the 1000 Genomes Project (1KGP). We identify distinct patterns of LD in different ancestries, as expected, but we uncover a universal statistical structure shared across ancestries, over which the ancestral differences are overlaid. Among the 1KGP populations with a sample size of at least 400, the number of LD blocks on the autosomes ranges from 6-8 x 10^3. In each population, the widths of the LD blocks approximately follow an exponential distribution (implying a Poisson distribution of recombination hotspots), with a median width in the range 200-290 Kb. Previous studies of recombination rates have suggested that populations of diverse ancestry may share recombination hotspots. We verify this conjecture with the discovery of common LD boundaries shared across populations at high resolution, constituting 2 x 10^3 universal LD blocks that span the autosomes, with a median (mean) size of 700 Kb (1.5 Mb). Our results have important implications for statistical genetics. The universal LD blocks, in particular, enable genome-wide (rather than region-specific) statistical fine-mapping of trans-ancestry GWAS data, a computation that to our knowledge has never been carried out. We make the population-specific and universal LD blocks available as an open resource to the academic community, to facilitate these and other applications in statistical genetics.

A- A+

PgmNr 296: Public programmatic access to GWAS summary statistics and analytical methods.

Authors:
M. von Grotthuss [1]; J. Massung [1]; B. Alexander [1]; L. Caulkins [1]; M. Costanzo [1]; M. Duby [1]; C. Gilbert [1]; D.K. Jang [1]; R. Koesterer [1]; P. Singh [1]; O. Ruebenacker [1]; A. Boughton [2]; R. Welch [2]; M. Boehnke [2]; N. Burtt [1]; J. Flannick [1,3]

View Session Add to Schedule

Affiliations:
1) The Broad Institute of MIT and Harvard, Cambridge, MA; 2) University of Michigan, School of Public Health, Ann Arbor, MI; 3) Boston Children's Hospital / Harvard Medical School, Boston, MA

The central challenge in human disease genetics is now biological translation: thousands of common variant associations with hundreds of traits offer as-yet-unrealized opportunities to better understand disease pathogenesis. Consequently, most human genetic method development now seeks to discern insights from genome-wide association study (GWAS) summary statistics (p-values and effect sizes): e.g., inferring disease-relevant cell types or pathways from global statistic distributions, or predicting causal genes from integrated genomic datasets. Nearly all such 'post-processing' methods require a full set of GWAS summary statistics and, in turn, produce results often used by yet more methods downstream. The mission of the KnowledgePortal Network (KPN) is to provide a resource of association statistics and methods for every complex disease, alongside epigenomic and transcriptomic datasets for every disease-relevant cell type. Here we describe an application programming interface (API) for (a) accessing the full set of GWAS summary statistics for 100 datasets across 234 traits, (b) conducting on-demand association analysis for 23 traits in >60,000 exomes, and (c) accessing the results of eight GWAS post-processing methods. The API is publicly accessible via the SmartAPI registry, and the underlying data can be further analyzed via custom workflow design language (WDL) scripts within the Terra platform. Because the canonical GWAS results for most traits include non-public data, we present an approach to balance public access with patient privacy restrictions. We demonstrate these APIs through three GWAS post-processing methods applied to >150 traits and >1M samples: (a) an overlap-aware 'bottom-line' meta-analysis for every trait (which increases sample size up to 5-fold over the largest published analysis); (b) stratified LD-score regression on these bottom-line results for >100 epigenomic annotations (which catalogs disease-relevant cell types); and (c) three eQTL co-localization methods at each bottom-line disease-associated locus in each disease-relevant cell type (offering the highest-powered such analysis to date). These results are integrated alongside those of other methods and GWAS datasets within a genetics 'knowledge graph', encoding links among diseases, variants, genes, and cell types. These resources, accessible via web portals at KPN4CD.org, could inform countless approaches or experiments to understand the biology captured by GWAS.

A- A+

PgmNr 297: The unbiased length spectrum of human de novo mutations in 4,330 children.

Authors:
A. Farrell [1]; W. Richards [1]; A. Docherty [2]; H. Coon [3, 4]; G. Marth [1]

View Session Add to Schedule

Affiliations:
1) Biology, University of Utah, Salt Lake City, Utah.; 2) University of Utah, Department of Psychiatry and Human Genetics, Salt Lake City, UT; 3) University of Utah, Department of Psychiatry, Salt Lake City, UT; 4) University of Utah, Department of Biomedical Informatics, Salt Lake City, UT

De novo genetic variation introduces new variations into the population and can cause genetic diseases. Despite its importance, the true spectrum of de novo human mutations has been elusive primarily for technical reasons: accurately detecting the roughly 70 de novo mutations in 6 billion base pairs poses a formidable challenge; compounded by the biases of mapping-based variant detection methods. Moreover, distinct classes of tools are currently used to reliably detect single nucleotide variations (SNVs) and short insertions-deletions (INDELs) of up to ~50bp; and 500bp or larger structural variation events (SVs); leaving a “detection gap” for medium-sized (50-500bp) events.

We developed RUFUS, a k-mer based reference-free detection method for all types and sizes of de novo mutations. With RUFUS, mutations are identified and assembled before reads are compared to the reference, removing all reference bias and variant size limitation, resulting in completely even sensitivity and specificity across variations of all sizes. Experimental validations of RUFUS mutation calls in multiple datasets indicate extremely high accuracy. We applied this method to detect de novo mutations in 4,330 children in the Simons Foundation Simons Simplex Collection dataset, the largest dataset to date in which to study de novo genetic variation. Here we present the first comprehensive de novo somatic mutation dataset where variants of all sizes were detected by a single algorithm. Across both autistic probands and siblings, we saw an average of 72.6 denovo mutations per sample; 86.14% of these are SNVs, 13.58% are 1-50bp INDELs, 0.20% are medium-sized (50-500bp) INDELs, and 0.08% are >500bp SV events (i.e. on average 62.54 SNVs, 9.86 short INDELs, 0.145 medium INDELs, and 0.058 SVs per individual). When accounting for the size of the events, de novo SVs alter the genome of the child by far the most, on average by 3,804bp, and are present in approximately 1 of 17 births. Medium-sized INDELs, the rate of which until this study could not be fully ascertained, alter on average 20.69bp per birth, an effect comparable to that of SNVs (62.54bp per birth), and occur in 1 of 7 births. The overall distribution of de novo event size is indistinguishable between probands and siblings. However, de novo variation within autism-associated genes in the probands is markedly higher higher with RUFUS (i.e. more than double) as compared to the the siblings, a phenomenon part of ongoing investigation.

A- A+

PgmNr 298: Estimating assortative mating and its changes over time in samples of unrelated individuals.

Authors:
P. Turley [1,2]; R. Li [3]; L. Yengo [4]; D. Cesarini [5]; D. Benjamin [3,6]; B. Neale [1,2]; P. Visscher [4]; M. Kimball [3,7]

View Session Add to Schedule

Affiliations:
1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; 2) Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 3) Behavioral and Health Genomics Center, Center for Economic and Social Research, University of Southern California, 635 Downey Way, Los Angeles, CA 90089, USA; 4) Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, 4072, Australia; 5) Department of Economics and Center for Experimental Social Science, New York University, New York, NY, USA; 6) Department of Economics, University of Southern California, 635 Downey Way, Los Angeles, CA 90089, USA; 7) Department of Economics, University of Colorado - Boulder, Boulder, CO 80309

Introduction: Assortative mating (AM) is when parents are more similar to each other than would be expected if they were paired at random. Here, we consider genetic AM for single traits and pairs of traits, where parents have correlated genetic risk within or across traits. Genetic AM leads to more people with extreme values of genetic risk, leading to greater prevalence and severity of disease in the population and increased economic inequality. AM also biases genetic associations and induces genetic correlation between traits even when the traits share no common biology. The effects of AM are persistent, affecting traits in the population for several generations.

Objective: We develop and implement a method to estimate single-trait and cross-trait AM and its changes over time for a broad range of health, behavioral, and geographic traits.

Methods: We propose a novel method to estimate the correlation (which we denote as *M*) of parental polygenic scores for single traits and pairs of traits using only data on unrelated individuals. Because AM inflates the variance or covariance of polygenic scores we can use this excess (co)variance to infer AM for the corresponding traits. Our method estimates AM beyond population stratification or geographic factors, and it is robust to selected samples under weak assumptions. Using data from the UK Biobank, we are able to obtain estimates with much more precision than using traditional direct methods in other data sets.

Results: We see substantial increases in AM for education--from *M*=.09 (*SE*=.08) in 1940 to *M*=.21 (*SE*=.02) in 1965--and for height--from *M*=.06 (*SE*=.04) in 1940 to *M*=.12 (*SE*=.02) in 1965. We also estimate a constant, low level of AM for BMI over all years (*M*=.02, *SE*=.007). In contrast to previous literature, we do not observe any evidence of AM on Autism Spectrum Disorder for any year (*M*=.004, *SE*=.007). For our cross-trait analyses, we see that a moderate amount of the genetic correlation between education and a number of other traits is due to AM. For example, we estimate that 24% (*SE*=2.0%) of the genetic correlation between education and depression is driven by AM rather than by biological factors. For height, we estimate this value to be 22% (*SE*=1.8%), and for smoking behavior, 16% (*SE*=1.6%).

Conclusion: AM can be precisely estimated using data on unrelated individuals. These estimates suggest that care must be taken in interpreting genetic associations and genetic correlation estimates for many traits.

A- A+

PgmNr 299: X-chromosome dosage compensation dynamics in human early embryos.

Authors:
G. Fan [1]; K. Huang [1]; Q. Zeng [2]; Y. Feng [2]; Y. Hu [1]; L. Qin [3]; Q. An [1]; B. Lv [2]; J. Liu [3]; Z. Xue [2]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics, UCLA, Los Angeles, CA USA.; 2) Translational Center for Stem Cell Research, Tongji Hospital, Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, 200065, China; 3) State Key Laboratory of Reproductive Medicine, Center of Clinical Reproductive Medicine, First Affiliated Hospital, Nanjing Medical University, Nanjing 210029, China

---

In mammals, female cells are obliged to inactivate one of two X chromosomes to achieve dosage parity with the single X chromosome in male cells,and it is also thought that the single active X chromosome is increased 2-fold to achieve dosage balance with two sets of autosomes (X:A ratio = 1, or Ohno's hypothesis). However, the ontogeny of X-chromosome inactivation and augmentation of the single active X remains unclear during human embryogenesis. Here, we perform single-cell RNA-seq analysis to examine the timing of X:A balancing and X-inactivation (XCI) in pre- and peri-implantation human embryos up to fourteen days in culture. We find that X-chromosome gene expression in both male and female preimplantation embryos is approximately balanced with autosomes (X:A ratio = 1) after embryonic genome activation (EGA) and persists through fourteen days *in vitro*. Cross-species analysis of preimplantation embryo also show balanced X:A ratio within the first few days of development. By single-cell mRNA SNP profiling, we find XCI beginning in day 6-7 blastocyst embryos, but does not affect X:A dosage balance. XCI is most evident in trophoectoderm (TE) cells, but can also be observed in a small number of inner cell mass (ICM)-derived cells including primitive endoderm (PE) and epiblast (EPI) cells. Analysis between individual XaXa and XaXi sister cells from the same embryo reveals random XCI and persistently balanced X:A ratio, including sister cells transitioning between XaXa and XaXi states. We therefore conclude that the male X-chromosome undergoes X chromosome augmentation prior to the simultaneous X-chromosome inactivation and augmentation in females. Together, our data demonstrate an evolutionally conserved model of X chromosome dosage compensation in humans and other mammalian species.

A- A+

PgmNr 300: Detection of uniparental disomy in a population of 32,000 clinical exome trios.

Authors:
J. Scuffins [1]; J. Keller-Ramey [1]; L. Havens [1]; G. Douglas [1]; N. Robin [2]; N. Rudy [2]; K. Retterer [1]

View Session Add to Schedule

Affiliations:
1) GeneDx, Gaithersburg, MD; 2) Department of Genetics, University of Alabama at Birmingham, Birmingham, AL

Uniparental Disomy (UPD) has been implicated in disease through imprinting, recessive variant unmasking, and underlying trisomy mosaicism. However, data on the prevalence and spectrum of UPD remains limited. Trio Exome Sequencing (ES) presents a comprehensive method for detection of UPD alongside variant analysis.

We reviewed 32,067 ES trios referred for diagnostic testing due to suspected Mendelian disease in order to create a profile of UPD observations and their disease association. Possible UPD results were identified by recurrent Mendelian errors; results were overlapped with read-depth based coverage data to distinguish UPD from large deletions. Events were categorized as whole chromosome or segmental UPD, and whole chromosome results as isodisomy, heterodisomy, or mixed iso/heterodisomy by degree of homozygosity.

99 whole chromosome and 13 segmental UPD results were identified along with 37 deletions ≥5Mb in length. The incidence of whole chromosome UPD was 1:347. Most commonly observed were chromosomes 1 and 15 (17x each), 16, 22, 2, and 14, with four instances of double aneuploidy with chromosome X/Y. Mendelian errors made up a higher percentage of informative SNPs in isodisomy (19.9%) than heterodisomy (7.3%). Isodisomy was more commonly observed in large chromosomes along with a higher rate of homozygous pathogenic variants. Heterodisomy was more commonly observed in chromosomes associated with imprinting or trisomy mosaicism (14, 15, 16, 20, 22). Whole chromosome UPD was associated with a positive test outcome in 48/99 cases (48.5%), 28 by imprinting and 20 recessive variant unmasking. Only three cases reported a positive result unrelated to the UPD. The parent of origin skewed maternal (69/99 cases, p=1.1E-04) and average parental ages were higher in the UPD than non-UPD group (maternal 32.7, p=1.9E-05; paternal 35.1, p=5.5E-04). Segmental UPD regions were non-recurrent and ranged in size from 5-93Mb. They were uniformly consistent with isodisomy and associated with a terminal end of the chromosome, while deletions were a mix of terminal and interstitial location.

To illustrate the clinical utility of our method, we present a patient with congenital anomalies and growth delay for whom prior karyotyping and aCGH had been negative. ES revealed maternal heterodisomy of chromosome 22, and follow-up FISH confirmed Trisomy 22 mosaicism. As UPD was overall relevant to 1:688 diagnoses, interrogating UPD is a valuable addition to diagnostic ES.

A- A+

PgmNr 301: Comprehensive estimation of the tissue of origin of circulating cell-free DNA.

Authors:
C. Caggiano [1, 4]; B. Celona [2]; F. Garton [3]; B. Black [2]; N. Wray [3]; A. Dahl [4]; N. Zaitlen [4]

View Session Add to Schedule

Affiliations:
1) Bioinformatics Program, University of California Los Angeles, Los Angeles, CA; 2) Department of Cardiology, University of California San Francisco, San Francisco, CA; 3) Institute for Molecular Bioscience, University of Queensland, Queensland, Australia; 4) Department of Neurology, University of California Los Angeles, Los Angeles, CA

Cell-free DNA (cfDNA) in the bloodstream originates from dying tissues. The analysis of cfDNA provides a non-invasive biomarker for diseases characterized by tissue-specific cell death. The purpose of this work is to create a statistical model that can accurately estimate which tissues are contributing to the presence of cfDNA in the blood. To do this, we leverage the distinct DNA methylation profile of each tissue type throughout the body, and use this information to estimate the contribution of each of these tissues to the cfDNA mixture. Decomposing these mixtures, however, is difficult, as cfDNA of a disease relevant tissue may only be present in the blood only in small amounts. Futhermore, many DNA methylation datasets, such as those from the ENCODE or ROADMAP projects, are whole genome bisulfite sequencing (WGBS), which are generally of low read depth (~10x). This low read depth means that methylation count data may be missing at a given DNA methylation site (CpG), or observed so infrequently as to be unreliable. Finally, accurately decomposing cfDNA mixtures requires a robust understanding of all possible tissue types that could potentially contribute to the mixture. This robust reference, however, is nearly impossible to assemble, as there are hundreds of distinct tissue types, and because the methylation state for a CpG in a tissue can vary. We developed an EM algorithm that estimates tissue type proportion from both WGBS cfDNA input and tissue reference data. Notably, our algorithm can handle missing and low count data, and does not rely on CpG site curation. Our EM algorithm can also estimate an arbitrary number of ‘unknown’ tissue type categories. We show in simulations that our algorithm can accurately estimate tissue of origin of cfDNA mixtures. Simulations also demonstrate that we can effectively estimate cfDNA originating from ‘rare’ cell types. We also apply our EM algorithm to cfDNA from ALS patients. We can detect differences between the tissue of origin of cfDNA from ALS patients and from healthy controls using cfDNA methylation, illustrating that this method can potentially identify clinical biomarkers for complex human diseases.

A- A+

PgmNr 302: Non-invasive monitoring of kidney transplant conditions via donor-derived cell-free DNA.

Authors:
P. Nguyen [1,2]; H. Nakaoka [1]; K. Saigo [3]; T. Hayano [4]; I. Inoue [1]

View Session Add to Schedule

Affiliations:
1) Laboratory of Human Genetics, National Institute of Genetics, Mishima, Shizuoka, Japan; 2) Department of Genetics, The Graduate University for Advanced Studies (SOKENDAI), Mishima, Shizuoka, Japan; 3) Department of Surgery, National Hospital Organization Chiba-East Hospital, Chiba, Japan; 4) Department of Systems Bioinformatics, Yamaguchi University Graduate School of Medicine, Ube, Yamaguchi, Japan

---

Monitoring graft conditions requires intensive follow-up medical examinations. In kidney transplantation, graft integrity is conventionally assessed by histopathology and serum markers. Host-versus-graft immune reactions can cause graft injury, leading to the release of donor-derived DNA into circulation. Based on inherited genetical difference between donor and recipient, novel non-invasive methods employing detection of donor-derived cell-free DNA (ddcfDNA) are widely studied recently.
With the aim to explore the dynamics of ddcfDNA in kidney transplant patients and to investigate its potential use as a novel biomarker for monitoring graft integrity, we collected ~300 plasma samples from 39 patients with extensive follow-up intervals, from the imminent days post-surgery to the timepoints over 2 years. We established capture-based sequencing procedure on extracted cell-free DNA which targets 1000 selected single nucleotide polymorphism sites. Our pooled-indexing protocol can be performed within 2 days and well-suited for analyzing a large number of samples.
Our analyses described typical baseline dynamic of ddcfDNA level in transplant recipients annexed with detail clinical measurements. Levels of ddcfDNA on the first day post-surgery (n=34) ranged from 0.45% to 28.9% (median = 25.3%) and remained highly variable during the first week, then exponentially decreased to below 1% after 2 weeks in more than 50% cases. Compared to routine serum markers (i.e Creatinine, Cystatin-C and blood urea nitrogen), ddcfDNA level showed higher sensitivity in reporting rejection episodes. Notably, in 80% of cases having signs of rejection episodes and in all confirmed rejection cases, ddcfDNA level were remarkably elevated. We reported that early-days high level of ddcfDNA was linked with bleeding amount of donor surgery (Pearson’s r=0.52, p=0.0117). Additionally, ischemic durations of transplanted kidney could significantly affect ddcfDNA stabilizing rate in recipients. On the other hand, we found out that ddcfDNA level sensitively corresponded with immunosuppressant dosages which were administered on recipients. Significant correlation of ddcfDNA levels with circulating trough level (C) of Tacrolimus (Pearson’s r=0.49, $p<10^{-5}$) was also reported, implying the indicative capability of ddcfDNA level on drug metabolism and dosing effectiveness. Our study demonstrated that ddcfDNA level is a preferable non-invasive marker for monitoring multiple aspects of graft health.

A- A+

PgmNr 303: Massively parallel characterization of regulatory dynamics during neural induction.

Authors:
A. Kreimer [1,2]; F. Inoue [2]; T. Ashuach [1]; N. Ahituv [2]; N. Yosef [1]

View Session Add to Schedule

Affiliations:
1) University of California Berkeley, CA.; 2) University of California San Francisco, CA

The temporal interplay between gene regulation and gene expression during cell differentiation remains largely unknown. Using neural induction as a model, we set out to decipher these dynamics. We performed RNA-seq, ChIP-seq (H3K27ac, H3K27me3) and ATAC-seq on human embryonic stem cells at seven early neural differentiation time points (0-72 hours). We found that DNA accessibility precedes H3K27ac, which is then followed by gene expression changes. Using massively parallel reporter assays (~2,500 sequences tested at all seven time-points), we further show that temporal enhancers correlate with H3K27ac. Development of a prioritization method that incorporated all genomic data identified key transcription factors (TFs) involved in temporal function, several of which were functionally validated to be important and novel neural induction regulators. Combined, our results provide a resource of genes and regulatory elements that orchestrate neural induction and illuminate the temporal framework that is needed to obtain this differentiation.

A- A+

PgmNr 304: The ATP-dependent chromatin remodeler CHD7 is critical for neuronal lineage differentiation by changing chromatin accessibility and nascent RNA.

Authors:
D.F. Hannum [1,2]; H. Yao [2]; S.F. Hill [3]; R.D. Albanus [1]; W. Lou [2]; J.M. Skidmore [2]; G.J. Sanchez [2]; A. Saiakhova [6]; S.L. Bielas [5]; P.C. Scacheri [6]; M. Ljungman [4]; S.C.J. Parker [1,5]; D.M. Martin [2,5]

View Session Add to Schedule

Affiliations:
1) Dept Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI; 2) Dept of Pediatrics, University of Michigan, Ann Arbor, MI; 3) College of Literature, Science and Arts, University of Michigan, Ann Arbor, MI; 4) Dept of Radiation Oncology, University of Michigan, Ann Arbor, MI; 5) Dept of Human Genetics, University of Michigan, Ann Arbor, MI; 6) Dept of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH

Background: CHARGE syndrome, a rare congenital multiple anomaly condition, is caused mainly by haploinsufficiency of the ATP-dependent chromatin remodeling enzyme Chromodomain Helicase DNA binding protein 7 (*Chd7*). Brain abnormalities and intellectual disability are commonly observed in CHARGE patients. In addition, neuronal differentiation is reduced in CHARGE patient-derived iPSCs and in conditional knockout mouse brains. However, the underlying mechanisms of *Chd7* function in nervous system development is not well understood.
Methods: *Chd7*$^{+/+}$ and *Chd7*$^{Gt/Gt}$ (null) embryonic stem cells (ESCs) were derived from sibling blastocysts. Using a previously described protocol, ESCs were differentiated to neural progenitor cells (NPCs) and then further differentiated to neurons and glial cells. Quantitative RT-PCR, cell growth assays, immunostaining, ATAC-sequencing and Bru-sequencing were performed on ESCs and NPCs.
Results: *Chd7* expression increased during ESC to NPC and neuronal differentiation, suggesting important roles for *Chd7* in NPCs and neuronal differentiation. Interestingly, loss of *Chd7* did not affect ESC or NPC identity or proliferation, but significantly reduced TUJ1^{+} neurons (P = 9.7 x 10^{-4}) and GFAP^{+} glial cells (P = 5.8 x 10^{-9}). ATAC- and Bru-seq experiments identified transcription factors (*Pax3*, *Tbx3*, *Nkx6-1*, and *Zic5*) with differentially accessible promoter regions and expression in *Chd7*$^{Gt/Gt}$ vs wildtype NPCs. These transcription factors are known to play roles in central nervous system development and neural/glial differentiation.
Conclusion: *Chd7* loss does not adversely affect ESCs and NPCs but results in a decrease of differentiated neurons and glial cells. CHD7 has a profound effect on the chromatin landscape in NPCs and genome-wide studies indicate mis-regulated transcription factors in *Chd7*$^{Gt/Gt}$ NPCs. These studies suggest that *Chd7* acts preferentially during the transition of NPCs to neurons and glial cells to promote one or more aspects of differentiation and lineage differentiation.

A- A+

PgmNr 305: Identifying critical downstream gene targets of transcription factors associated with disease.

Authors:
A.M. Barbeau; A. Hoang; K. Hazel; O. Corradin

View Session Add to Schedule

Affiliation: Whitehead Institute for Biomedical Research, Cambridge, MA

Transcription factors (TFs) play a critical role in regulating gene expression and defining cellular identity. Results from GWAS frequently associate dysregulation of TF expression with disease. However, these TFs regulate many genes which can vary by cell type. Thus, deconvoluting the mechanism by which TF dysregulation contributes to disease is a substantial challenge. Here, we present a two-pronged approach to delineate downstream targets and pathways that are critical to disease. First, we identify the pathogenic cell type for the TF, i.e. the cell type in which dysregulation of the TF contributes to disease. Second, we evaluate TF target genes and assess SNPs within TF binding sites (TFBS) for association with disease. This enables us to identify the subset of TF targets that contribute to disease pathogenesis. We applied this approach to a multiple sclerosis (MS) risk allele which physically interacts with TF *IRF8* in 9 immune cell types. We identified all SNPs in putative transcriptional enhancers that also interact with *IRF8* using ChIP and Hi-C. We utilized a previously published approach to assess the role these additional regulatory variants play in defining risk. This enables us to create a genetic risk ‘barcode’ that distinguishes local regulatory regions that contribute to MS risk from those that do not. We compared this barcode to active chromatin regions to identify likely pathogenic cell types. We predicted that IRF8 contributes to MS risk through its aberrant expression in monocytes, and not the expected T cells. We next performed IRF8 ChIP-seq and analyzed DNA variants within promoter proximal TFBS for contribution to MS risk. IRF8 TFBS that contribute to MS risk were significantly enriched for H3K27ac specifically in the monocytes, supporting our pathogenic cell type prediction. We performed a pathway analysis of the promoters harboring IRF8 TFBS that alter MS risk and found an enrichment of genes involved in the Notch signaling pathway. Interestingly, IRF8 dysregulation is also associated with monocyte counts. Collectively, these results point to IRF8 dysregulation in monocytes disrupting myeloid maturation as a key contributor to MS. This approach integrates genetic risk with chromatin interaction data to generate novel hypotheses about the mechanisms by which TFs contribute to disease and can be readily applied to other disease associated TFs.

A- A+

PgmNr 306: Spatially-resolved single-cell chromatin accessibility in the adult mouse brain.

Authors:
C.A. Thornton [1]; A. Mishra [2]; A.P. Barnes [2]; B.J. O'Roak [1]; A.C. Adey [1,2,3]

View Session Add to Schedule

Affiliations:
1) Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR.; 2) Knight Cardiovascular Institute, Portland, OR; 3) Knight Cancer Institute, Portland, OR

The adult brain is a complex network of neuronal and non-neuronal cell type interactions with spatially oriented regions that accomplish a near infinite variety of computational tasks. The mammalian cerebral cortex in particular carries out higher cognitive processes such as vision, motor function and hearing, and is characterized by six spatially distinct layers comprised of unique populations of neuronal and non-neuronal cell types. The advent of single-cell omics technologies has allowed for the characterization of individual cell epigenomic states in heterogeneous tissues. However, our ability to characterize how cells vary by spatial orientation is limited. In this study we introduce a novel technique for the characterization of single-cell epigenomic landscapes while maintaining the spatially-resolved origin of the profiled cells. We focus on the somatosensory cortex of adult mice, in order to demonstrate the gradient of epigenetic states of cells from lower to upper cortical layers. Spatially-resolved Single-cell Combinatorial Indexed Assay-for-Transposase-Accessible-Chromatin using sequencing (sci-ATAC-seq) was used to generate 19,896 single-cell chromatin accessibility profiles from spatially oriented microbiopsies from upper cortex, lower cortex and whole mouse brains. We utilized topic-based clustering to identify major neuronal and non-neuronal cell types, and further resolved epigenetic gradients based on upper and lower cortical layer spatial identity. Additionally, we have applied this novel technique to stroke model mice and demonstrate the ability of spatially resolved sci-ATAC-seq to reveal chromatin remodeling that occurs in a spatial gradient extending from the infarct. We believe that this novel method can be utilized across many tissues in order to fully characterize the spatial epigenomic landscapes of complex tissues.

A- A+

PgmNr 307: AT-007, a novel CNS penetrant aldose reductase inhibitor prevents the metabolic and tissue specific abnormalities of Galactosemia, in a GALT deficient rat model of disease.

Authors:
R. Perfetti; S. Shendelman

View Session Add to Schedule

Affiliation: Applied Therapeutics, New York, New York.

Galactosemia is a rare, devastating metabolic disease that affects how the body processes galactose, and for which there are no approved therapies. It is caused by deficiency in any of the enzymes that metabolize galactose - GALK, GALT or GALE. Galactose is a sugar produced endogenously by the body, and is also a metabolite of lactose. While prompt identification of infants via newborn screening and immediate implementation of a lactose-restricted diet prevents many fatalities, the long-term consequences of the disease persist, due to endogenous generation and accumulation of galactose and galactose-metabolites in various tissues. The clinical presentation of Galactosemia is characterized by cognitive and intellectual deficiencies, speech and motor pathologies, pre-senile cataracts, tremors, and ovarian insufficiency in females. The development of the clinical phenotype has been linked to conversion of galactose to galactitol, an aberrant metabolite of galactose, catalyzed by the enzyme Aldose Reductase (AR). AT-007 is a novel CNS penetrant and retinally penetrant AR inhibitor (ARI). The normal function of AR is to convert glucose to sorbitol; however, in Galactosemia patients, the enzyme can break down galactose to galactitol. Galactitol is not a metabolite produced in healthy individuals, as the affinity of the enzyme for the substrate requires abnormally high levels of galactose. We used a genetic rat model of Galactosemia (GALT null), which recapitulates critical features of Galactosemia in humans to test the study hypothesis. The accumulation of galactitol and the onset of cataracts and CNS deficiencies (cognitive, memory and motor deficits measurable by rotarod and water maze testing) in this animal model serve as biochemical and clinical markers of disease progression. Daily administration of AT-007 in GALT null rats reduced galactitol, prevented the onset of cataracts, and ameliorated the CNS dysfunction, as measured by rotarod testing. A dose response was demonstrated, showing that higher doses of AT-007 produce a greater effect on galactitol reduction and more favorable disease outcomes. Conclusions: AR inhibition with a potent CNS penetrant ARI reduces accumulation of the toxic metabolite galactitol, and ameliorates the outcome of the disease in an animal model of Galactosemia. These findings support further investigation of the therapeutic potential of AT-007 in Galactosemia.

PgmNr 308: Taurine supplementation treatment for progressive retinal degeneration and cardiomyopathy caused by a novel recessive gene *SLC6A6*.

Authors:
M. Ansar [1]; E. Ranza [1,2,3]; M. Shetty [4]; S.A. Paracha [5]; M. Azam [6]; M.T. Sarwar [5]; I. Kern [7]; O. Farooq [8]; C.J. Pournaras [9]; A. Malcles [10]; L. Ali [6]; F.A. Santoni [1,11]; P. Makrythanasis [1,12]; R. Qamar [6]; J. Ahmed [5]; K. Henry [4]; S.E. Antonarakis [1,2,13]

View Session Add to Schedule

Affiliations:
1) Department of Genetic Medicine and Development, University of Geneva, Geneva, Geneva, Switzerland; 2) Service of Genetic Medicine, University Hospitals of Geneva, Geneva, Switzerland; 3) Current address, Medigenome, Swiss Institute of Genomic Medicine, Geneva, Switzerland; 4) Dept. of Biomedical Sciences, School of Medicine and Health Sciences, University of North Dakota, Grand Forks, ND, United States; 5) Institute of Basic Medical Sciences, Khyber Medical University, Peshawar, Pakistan; 6) Department of Biosciences, Faculty of Science, COMSATS University, Islamabad, Pakistan; 7) Pediatric Nephrology and Metabolism Unit, Pediatric Subspecialties Service, Children's Hospital, Geneva University Hospitals, Geneva, Switzerland; 8) Bahria University Medical and Dental College, Karachi, Pakistan; 9) Hirslanden Clinique La Colline, Geneva, Switzerland; 10) Department of Ophthalmology, University Hospitals of Geneva, Geneva, Switzerland; 11) Current address, Department of Endocrinology Diabetes and Metabolism, University Hospital of Lausanne, Lausanne, Switzerland; 12) Current address, Biomedical Research Foundation of the Academy of Athens, Athens, Greece; 13) iGE3 Institute of Genetics and Genomics of Geneva, Geneva, Switzerland

---

We have studied a Pakistani consanguineous family with two children with progressive visual impairment and mild cardiomyopathy, and identified a pathogenic homozygous missense variant Gly399Val in the 8th transmembrane domain of *SLC6A6*. 3D modeling of this variant in the novel causative gene has indicated that it likely causes displacement of the Tyr138 (TM3) side chain, important for transport of taurine. The two affected children had very low blood taurine levels (6 and 7 μmol/l), heterozygous parents had intermediate levels (24 and 34 μmol/l), and a non-carrier sibling had normal levels (71 μmol/l) (normal values: 37-127μmol/l). Taurine uptake of HEK-293 SLC6A6 mutant cells (Gly399Val) was between 11-18% compared to normal. Experiments in fibroblasts of affected individuals gave similar results. Detailed clinical evaluation showed that the affected 15 y.o. boy had complete visual loss, while the affected 5 y.o. girl, had retained some visual function. Both affected siblings had mild hypokinetic cardiomyopathy with systolic dysfunction. Slc6a6 knockout in a previously reported mouse model caused progressive retinal degeneration, cardiomyopathy and very low taurine levels in blood and other tissues (PMID: 17875433). Following ethics approval from the Swiss authorities, we have initiated an oral taurine supplementation (100mg/kg/day, in 3 doses) to both affected individuals. Clinical evaluations performed after two years of treatment showed normal taurine levels in blood, no further progression of the retinal degeneration, and rather slight improvement in vision of the affected girl (documented by multifocal ERG); very strikingly echocardiography was normal in both affected individuals compared to the cardiomyopathy with systolic dysfunction which was noted before starting the treatment.

We conclude that taurine supplementation could arrest the progressive retina degeneration and could

improve the symptoms of cardiomyopathy caused by hypomorphic functional defects in the SLC6A6 transporter in humans.

A- A+

PgmNr 309: A de novo *CLCN7* variant alters lysosomal pH and leads to lysosomal storage and albinism.

Authors:
M.C. Malicdan [1, 5, 6]; E.R. Nicoli [1]; M. Weston [2]; M. Hackbarth [1]; A. Becerril [2]; A. Larson [4]; W.M. Zein [3]; P.R. Baker II [4]; J.D. Burke [5]; H. Dorward [5]; M. Davids [1]; Y. Huang [1]; D.R. Adams [1, 5, 6]; P.M. Zerfas [7]; D. Chen [8]; T.C. Markello [1, 6]; C. Toro [1, 6]; T. Wood [9]; G. Elliott [10]; M. Vu [11]; U.D.N. Undiagnosed Diseases Network [12]; W. Zheng [11]; L. Garrett [10]; C.J. Tifft [1, 6]; W.A. Gahl [1, 5, 6]; D.L. Day-Salvatore [2]; J.A. Mindell [13]

View Session Add to Schedule

Affiliations:
1) UDP, NHGRI, National Institutes of Health, Bethesda, Maryland.; 2) Membrane Transport Biophysics Section, National Institute of Neurological Disorders and Stroke, NIH, Bethesda, Maryland 20892, USA; 3) Ophthalmic Genetics and Visual Function Branch, National Eye Institute, NIH, Bethesda, Maryland 20892, USA; 4) Department of Pediatrics, Section of Genetics, University of Colorado School of Medicine, Aurora, Colorado 80045, USA; 5) Human Biochemical Genetics Section, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892, USA; 6) Office of the Clinical Director, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892, USA; 7) Diagnostic and Research Services Branch, Office of Research Services, NIH, Bethesda, Maryland 20892, USA; 8) Division of Hemapathology, Mayo Clinic, Rochester, Minnesota 55905, USA; 9) Metabolic Laboratory, Greenwood Genetic Center, Greenwood, South Carolina 29646, USA; 10) Embryonic Stem Cell and Transgenic Mouse Core, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892, USA; 11) National Center for Translational Science, NIH, Rockville, Maryland 20850, USA; 12) Undiagnosed Diseases Network, Common Fund, Office of the Director, NIH, Bethesda, Maryland 20892, USA; 13) Department of Medical Genetics and Genomic Medicine, Saint PeterÂ's University Hospital, New Brunswick, New Jersey 08901, USA.

Introduction: Optimal lysosome function requires maintenance of an acidic pH, which is tightly regulated by proton pumps and counterion transporters such as the Cl^-/H^+ exchanger, CLC-7, encoded by *CLCN7*. The role of CLC-7 in maintaining lysosomal pH has been controversial.
Methods: Clinical and genetic evaluations were performed independently in two clinical centers. Functional validations were carried out using cells (dermal fibroblasts from patients and controls) and model organisms (Xenopus oocytes and mice).
Results: Two children of different ethnicities with developmental delay presented with lysosomal storage and hypopigmentation with a *de novo* Y715C variant in *CLCN7.* Neither proband had osteopetrosis, unlike patients and animal models with loss-of-function variants in *CLCN7*. Probands' biopsy material and primary fibroblasts exhibited enlarged intracellular vacuoles, findings that were also documented upon overexpression of human Y715C ClC-7 in control fibroblasts. In *Xenopus* oocytes, heterologous expression of the Y715C resulted in increased outward currents compared with the WT protein. Our results suggest that the Y715C is either a hypermorph or a neomorph. The pathogenicity of the *CLCN7* variant was further supported by mice genetically engineered to harbor the Y715C variant in one allele. These *Clcn7* knock-in mice exhibited hypopigmentation, hepatomegaly with lysosomal storage, and enlarged vacuoles in cultured fibroblasts, recapitulating the major clinical phenotypes observed in our probands.
Measurent of lysosomal pH in probands' fibroblasts revealed ph of approximately 0.2 units lower than control cells; the effects of this hyperacidified lysosomal pH can explain many of the disease

properties. Notably, treatment of the fibroblasts from our probands with chloroquine, an a drug that alkalinizes lysosomes, normalized the lysosomal pH and diminished the number of enlarged vacuoles .

Conclusion: *CLCN7* Y715C is a *de novo*, gain-of-function variant that increased ion transport, resulting to lysosomal hyperacidity; this in turn leads to abnormal storage, hypopigmentation, and enlarged intracellular vacuoles. The findings from this novel disease and the response to a potential, mechanism-based, therapeutic intervention are consistent with models where the CLC-7 antiporter plays a critical role in maintaining lysosomal pH. Our work exemplifies precision medicine and provides a proof-of-principle for drug repurposing.

A- A+

PgmNr 310: Results of an open-label phase 2 study of ManNAc in subjects with GNE myopathy.

Authors:
N. Carrillo [1]; M.C. Malicdan [1,3]; K. Bradley [1]; C. Slota [2,6]; J.A. Shrader [5]; P. Leoyklang [1]; B. Class [2]; J. Perreault [2]; C. Ciccone [1]; R. Parks [5]; M. Quintana [4]; J. Galen [5]; J. Heiss [7]; S. Van Wart [8]; C.T. Driscoll [9]; S.M. Berry [4]; M. Huizing [1]; W.A. W.A. [1,3]

View Session Add to Schedule

Affiliations:
1) Medical Genetics Branch, NHGRI, NIH, Bethesda, MD; 2) Therapeutics for Rare and Neglected Diseases, NIH, Bethesda, MD; 3) Undiagnosed Diseases Program, NHGRI, NIH, Bethesda, MD; 4) Berry Consultants LLC, Austin, TX; 5) Department of Rehabilitation Medicine, NIH Clinical Center, Bethesda, MD; 6) RTI Health Solutions, Research Triangle Park, NC; 7) National Institute of Neurological Disorders and Stroke, NIH, Bethesda, MD; 8) Enhanced Pharmacodynamics LLC, Buffalo, NY; 9) Technology Transfer Office, NHGRI, NIH, Bethesda, MD

Sialylation of glycans is crucial for several biological processes, including cell interactions and regulation of cell surface functions. GNE myopathy is a rare genetic muscle disease caused by mutations in UDP-GlcNAc 2-epimerase/ManNAc kinase (*GNE*), the enzyme that initiates and regulates N-acetylneuraminic acid (Neu5Ac) biosynthesis and glycan sialylation. GNE myopathy is characterized by progressive skeletal muscle weakness and atrophy in young adults. Several studies provide evidence that hyposialylated muscle membrane glycans play an essential role in the pathophysiology of GNE myopathy. There is no approved therapy for the disease. N-acetylmannosamine (ManNAc), an uncharged monosaccharide and first committed precursor in Neu5Ac biosynthesis, is an oral therapeutic candidate that restores sialylation and prevents muscle weakness in a mouse model of GNE myopathy. A first-in-human, randomized, placebo-controlled, double-blind study (NCT01634750) showed that a single dose of ManNAc is safe and restored Neu5Ac production in GNE myopathy patients. A Phase 2, open-label, single-center study (NCT02346461) was conducted from 2015-2018 in twelve subjects with genetically-confirmed GNE myopathy. Primary objectives were to assess long-term safety and tolerability, pharmacokinetics, and biochemical effect of ManNAc as measured by increased sialylation of muscle membrane glycans. Secondary objectives were to identify clinical endpoints suitable for subsequent trials. Subjects received oral ManNAc at 6 grams twice daily (12 grams/day) and were evaluated at baseline, 3, 6, 12, 18, 24, and 30 months. Long-term administration of ManNAc at 6 grams twice daily was safe. Biochemical efficacy was shown by a statistically significant increase in muscle membrane glycoprotein sialylation after 3 months of daily ManNAc administration. Measures of muscle strength (including quantitative muscle assessment), function, and patient-reported outcomes were evaluated for suitability to test clinical efficacy in future trials. The findings of this study support further development of ManNAc as a potential therapy for GNE myopathy. A multi-center, randomized, placebo-controlled, double-blind trial of ManNAc in subjects with GNE myopathy is planned.

A- A+

PgmNr 311: Long-read transcriptome sequencing in over 60 human tissue samples reveals isoform diversity.

Authors:
B. Cummings [1,2]; G. Garborcauskas [1]; M. Micorescu [3]; T. Bowers [1,4]; M.T. Costello [4]; F. Aguet [1]; K. Ardlie [1]; D.G. MacArthur [1,2]

View Session Add to Schedule

Affiliations:
1) Medical and Population Genetics, Broad Institute, Cambridge , MA, USA; 2) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA. USA. ?; 3) Oxford Nanopore Technologies, New York Genome Center, New York, NYC, USA; 4) Genomics Platform, Broad Institute, Cambridge, MA, USA

Short-read RNA sequencing (RNA-seq) has transformed our understanding of the transcriptional landscape of human tissues. However, a fundamental drawback of this approach is that it cannot assess full-length isoforms, which represent the true biological unit of the transcriptome. Instead, isoform measures from short read data are probabilistic and not directly quantified. This probabilistic nature of isoform quantification via RNA-seq is likely to have resulted in an underestimation of the level of tissue-specific splicing in humans and reduced power of analyses of the tissue-specific impact of genetic variants.

Long-read RNA-seq is an emerging technology that allows capture of full length isoforms, thus bypassing the flaws associated with short read RNA-seq. This technology has recently reached the point of scalable application to larger sample sets and has not been applied at scale to human tissues.

Here, we present the generation and analysis of long read RNA-seq data for over 65 human tissue samples from the Genotype Tissue Expression (GTEx) project. We first present benchmarking between the Oxford Nanopore Technologies (ONT) and Pacbio IsoSeq and within ONT protocols using human cardiac and GTEx fibroblast samples. We show that while current PacBio protocols allow for improved error correction, the substantially lower cost of ONT cDNA-PCR protocol allows for deeper isoform characterization. We next describe the generation and analysis of long read data on GTEx skeletal muscle, liver, lung, two heart regions (left ventricle and atrial appendage), and three brain subregions (cerebellar hemisphere, frontal cortex, and putamen). We generate sequencing runs of between 2-15 million reads, of which 30-60% represent complete end-to-end sequenced fragments, depending on a variety of factors including RNA quality, and experimental variables such as buffer. Results show that 30% of identified fragments represent unannotated isoforms. However, benchmarking against GTEx short-read RNA-seq reveals important error modes that can result in false isoform annotations. Overall this work represents the first large-scale application of long read RNA-seq to biobanked human tissue samples, and highlights key challenges that will guide future human isoform characterization studies.

A- A+

PgmNr 312: Refining polymorphic retrotransposon insertions in human genomes.

Authors:
W. Zhou [1]; R. Mills [1,2]

View Session Add to Schedule

Affiliations:
1) Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA; 2) Department of Human Genetics, University of Michigan Medical School, 1241 East Catherine Street, Ann Arbor, MI 48109, USA

Mobile element insertions (MEIs), including Long interspersed element-1 (L1), *Alu*, and SVA (SINE-VNTR-*Alu*) retrotransposons, comprise approximately 46% of the human genome and have been shown to play an important role in human development and disease. Various strategies have been developed to identify candidate polymorphic MEIs from short-read whole genome sequencing data, though they struggle in regions where reads can map equally to multiple alternative genomic positions. Long-read sequencing technology provides a better resolution in such regions by directly sequencing long stretches of contiguous DNA that enable the discovery of potential overlooked MEIs. Here, we present a comprehensive analysis of retrotransposon insertions across a diverse set of samples using an enhanced version of PALMER, an approach that uses a pre-masking strategy to consider endogenous reference repeats and then searches against a library of mobile element sequences to detect non-reference insertions within the remaining unmasked sequences. PALMER was designed to specifically identify characteristic features of retrotransposon insertions (e.g. target-site duplications, 3’ poly(A) tract, 5’ inverted sequence, and 3’ transduction). We applied PALMER to recently generated data from fifteen high coverage (>50x) PacBio whole genome sequences and identified a set of polymorphic MEIs (902 L1s, 5958 *Alu*s, and 358 SVAs), 53.8% (3880/7218) of which were detected in multiple samples. We observed 42.2% (381/902) of L1s, 33.8% (2011/5958) of *Alu*s, and 39.4% (141/358) of SVAs were absent in recent PacBio assembly-based studies, where an over-estimation may exist for SVAs (755 vs 358) and an under-estimation for human-specific L1s (615 vs 902) due to their omission or mis-annotation. We showed that these missing MEIs were predominantly found in endogenous reference LINE/SINE regions ($P < .0001$), suggesting that such regions hinder typical discovery approaches. An analysis of unique breakpoint junction sequences in short-read high coverage 1000 Genomes Project samples (n=2504, ~30x) further revealed that the 92.6% (6683/7218) of our detected MEIs were found in at least one sample and 71.8% (5185/7218) with an allele frequency of >5%. Together, we present a more holistic view of retrotransposon insertions in human populations and provide additional utility for studies using both short and long-read sequencing technology. The PALMER Software is available at https://github.com/mills-lab/palmer.

A- A+

PgmNr 313: Incorporating long transcriptomic data into GENCODE.

Authors:
J.E. Loveland [1]; J.M. Mudge [1]; J.M. Gonzalez [1]; T.J. Hunt [1]; A. Frankish [1, 2]; P. Flicek [1,2]; GENCODE Consortium

View Session Add to Schedule

Affiliations:
1) European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom; 2) Wellcome Trust Sanger Institute, Wellcome Trust Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

The accurate identification and structural and functional annotation of genes and transcripts in the human genome is fundamental for high quality analysis for genome biology and clinical genomics. Gene annotation that is incorrect or incomplete impacts downstream analysis and introduces potentially significant false positive and false negative errors. As part of the GENCODE project, we are responsible for producing detailed reference annotation of all human and mouse protein-coding genes, pseudogenes, long non-coding RNAs and small RNAs.

Historically gene annotation was supported predominantly by ESTs and mRNAs from INSDC databases but the emergence of long transcriptomic sequencing methods such as PacBio ISOseq and ONT replaces these data types but at massively greater volumes currently and even higher volumes in future. Fully automated approaches with these new data generate gene annotation rapidly which allows for scalability. However, for the annotation of the human genome, where there is mature reference annotation and genes and proteins not already agreed upon are almost always complex and difficult to interpret, it is not of sufficient quality. Similarly, generation of gene annotation based on short read RNAseq data identifies individual features such as novel exons, introns and splice sites in the human genome, but also fails to meet the standards of expert manual annotation. It is essential to develop methods to maintain the very high quality provided by manual annotation with the ability to scale provided by automated pipelines.

We have developed a manually supervised platform agnistic workflow for long transcriptomic data. In a pilot project PacBio Isoseq reads were aligned with two different methods; STAR and Gmap, SLRseq reads also with Gmap, then quality filtered and merged, introns confirmed with RNAseq. The resulting transcripts were clustered into models and filtered for novelty, then confirmed by manual assessment so as to maintain the quality of the GENCODE geneset. Novel lncRNA loci generated by this method (TAGENE) have been added to the GENCODE geneset (available in e97, GENCODE 31) and its utility for adding alternatively spliced transcripts to existing protein coding genes is being assessed. We are will also present a new annotation interface in order to more efficiently manually assess TAGENE transcripts.

PgmNr 314: What comes next? Long-read whole genome sequencing (WGS) to solve non-diagnostic whole exome sequencing and short-read WGS.

Authors:
N. Sobreira [1]; S. Aganezov [2]; R. Sherman [2]; E. Wholer [1]; E. Martin [1]; C. Bradburne [3]; S. Raskin [4]; D. Valle [1]; M. Schatz [2]

View Session Add to Schedule

Affiliations:
1) Institute of Genetic Medicine, Johns Hopkins Univ, Baltimore, Maryland.; 2) Department of Computer Science, Johns Hopkins University, Baltimore, Maryland; 3) Applied Physics Laboratory, Johns Hopkins University, Baltimore, Maryland; 4) Department of Genetics, Universidade Federal do Parana, Curitiba, Brazil.

It is well documented that WES has low sensitivity for detecting structural variants (SVs). CNVs longer than 300Kb are routinely identified by FISH, array-CGH, or whole genome SNP genotyping. However, deletions and duplications of intermediate size are not easily identified by these technologies or by WES. Current technologies also struggle to detect inversions and translocations. Recently, Oxford Nanopore introduced a new long-read nanopore sequencing platform that has the potential to address this important class of genomic variation. Here we describe 1 of 3 unsolved families from the Baylor-Hopkins Center for Mendelian Genomics with non-diagnostic WES (+/- short-read WGS) that we selected for long-read WGS. A non-consanguineous family in which the proband and 3 other males over 3 generations have an apparently X-linked congenital progressive myopathy. We performed WES in 4 affected males but did not identify any rare functional coding variant segregating with the phenotype. We performed the long-read sequencing on the proband and an affected maternal uncle. Using a single PromethION flow cell each, 22X and 25X coverage was generated for each patient in reads averaging ~10Kbp. SVs analysis on the long reads was performed using two orthogonal analysis pipelines, Sniffles and PBSV, which identified 21,808 and 23,007 SVs in the two patients. We focused on the variants that were common to both patients and were not present in 15 control human genomes also sequenced by long-read technologies which yielded 5,757 variants. We then examined which variants were within or near to known protein-coding genes and identified 73 candidate genes. Next, we focused on genes on chromosome X and identified only one gene: Apolipoprotein O Like (*APOOL)*, disrupted by a 240bp duplication on exon 9. Interestingly, *APOOL* was not sequenced by the WES performed in these 2 patients and other 2 affected males of this family. *APOOL* encodes a cardiolipin-binding protein that functions in a mitochondrial inner membrane protein complex and plays a critical role in determining mitochondrial cristae morphology. Consequently, the SV in *APOOL* togehter with its known function make it a strong candidate for the myopathy phenotype. We are currently evaluating the segregation of this variant in other family members. Thus, our results support long-read WGS as a new and powerful strategy for identifying the genes responsible for Mendelian diseases unexplained by WES and short-read WGS.

A- A+

PgmNr 315: Impact of pharmacogenetic testing on statin outcomes: Primary results from the Integrating Pharmacogenetics in Clinical Care (I-PICC) Study randomized trial.

Authors:
C.A. Brunette [1]; S. Advani [1]; A. Hage [1,2]; S.-J. Seo [1,2]; S.J. Miller [1]; N. Majahalme [1]; A.J. Zimolzak [1, 3]; J.L. Vassy [1,4,5]

View Session Add to Schedule

Affiliations:
1) VA Boston Healthcare System, Boston, MA; 2) Massachusetts College of Pharmacy and Health Sciences, Boston, MA; 3) Baylor College of Medicine, Houston, TX; 4) Harvard Medical School, Boston, MA; 5) Brigham and Women's Hospital, Boston, MA

Background: The association between the *SLCO1B1* rs4149056 C allele and simvastatin myopathy is well validated. The impact of *SLCO1B1* genotyping on patient care is largely unknown. The I-PICC Study is a randomized trial measuring the impact of preemptive point-of-care *SLCO1B1* genotyping on LDL cholesterol and guideline-concordant cardiovascular disease (CVD) care.

Methods: Primary care patients in the VA Boston Healthcare System are eligible if they are statin-naïve but have elevated CVD risk by American College of Cardiology/American Heart Association (ACC/AHA) guidelines (aged 40-75 with ≥1 of the following criteria: 1) preexisting CVD, 2) diabetes, 3) LDL ≥190mg/dL, or 4) 10-year CVD risk ≥7.5%). Consented patients are enrolled during clinical care if and when their primary care provider (PCP) signs an order for *SLCO1B1* genotyping of an extant clinical blood sample. Enrolled patients are randomized to have their PCPs receive the results at baseline (PGx+) or after 12 months (PGx-). The primary outcome is 12-month change in LDL. Secondary outcomes include concordance with ACC/AHA guidelines for statin therapy (CVD prevention) and Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for simvastatin therapy (drug safety) at 12 months. Exploratory outcomes include PCP documentation of communicating PGx results to patients and offering statin therapy.

Results: Enrollment of 408 patients (cared for by 39 PCPs) was completed in July 2018. Genotypes were balanced between the arms: 77% vs 65% TT, 21% vs 33% TC, and 2.6% vs 2.3% CC in the PGx+ vs PGx- arms, respectively. Of the first 350 patients, 59/165 (36%) and 60/185 (32%) were offered statin therapy and 26/165 (16%) and 22/185 (12%) initiated statin therapy during the study period in the PGx+ and PGx- arms, respectively. The mean (SE) change in LDL cholesterol was -1.5 mg/dL (1.3) in the PGx+ arm and -2.0 mg/dl (1.5) in the PGx- arm. 12 (7.3%) and 12 (6.5%) of patients in the PGx+ and PGx- arms, respectively, were receiving ACC/AHA guideline-concordant care at 12 months; all patients in both arms were receiving CPIC-concordant care at 12 months. PCPs documented communicating PGx results to 26/165 (16%) patients in the PGx+ arm.

Discussion: Final outcomes for all 408 enrollees will be available for hypothesis-testing in July 2019. Results of the I-PICC Study will inform the clinical utility of preemptive *SLCO1B1* testing in the routine practice of medicine.

A- A+

PgmNr 316: Pharmacy-supported return of over 10,000 pharmacogenomics test results: Semi-urgent eConsult data from the Mayo-Baylor RIGHT10K Pharmacogenomics Consortium Study.

Authors:
J. Wright [1]; E. Matey [1]; J. Giri [2]; R. El Melik [1]; L. Oyen [1]; J. Anderson [1]; W. Nicholson [2]

View Session Add to Schedule

Affiliations:
1) Pharmacy Department, Mayo Clinic, Rochester, MN; 2) Center for Individualized Medicine, Mayo Clinic, Rochester, MN

Background: Mayo Clinic conducted pre-emptive pharmacogenomics (PGx) testing for over 10,000 biobank participants as part of a research study. Test results for 13 genes have been returned to the electronic health record (EHR). Pharmacy leadership advocated for pharmacist involvement in providing an eConsult (electronic review of PGx results for each patient's current medications with reference to drug-gene interactions). Primary care providers (PCPs) may need guidance to manage current medications based on drug-gene interactions. Hence, this eConsult process was established to support and encourage adoption by PCPs who did not order this test.

Methods: In addition to the PGx results for the 13 genes, a supplemental report for over 300 medications was added to each patient's EHR. These medications were categorized as follows:
- Semi-urgent: Potential medications to cause serious harm based on PGx results. Expected time for completion of a semi-urgent eConsult is 48 hours from time of assignment.
- Clinically actionable: Potential medications to cause an adverse drug reaction or have reduced efficacy. Expected time for completion of a clinically actionable eConsult is 5 business days from time of assignment.

A pharmacist eConsult was provided only when patients had current medications that were identified as either semi-urgent or clinically actionable.

Results: Out of the 10,083 results returned to the EHR, 2,843 patients (28.2%) had an eConsult completed. Of these eConsults, 61 were semi-urgent (2.1%) and 2,782 (97.9%) were clinically actionable.
The medications in order of most to least common in the semi-urgent eConsults were clopidogrel 41/61 (67%), 9/61 (15%) citalopram, 7/61 (11%) escitalopram, 2/61 (3%) tramadol, 1/61 (2%) 5-fluorouracil, and 1/61 (2%) allopurinol. PCPs accepted 54% of pharmacists' semi-urgent eConsults recommendations.

Future Direction: Assess future adoption of PGx by clinicians across Mayo Enterprise and provide necessary tools such as education to support successful ongoing implementation efforts.

A- A+

PgmNr 317: Genome-wide association study of alanine transaminase and aspartate transaminase identifies novel variants with divergent effects on metabolic traits.

Authors:
V.L. Chen [1, 2]; X. Du [1]; Y. Chen [1]; S.K. Handelman [1, 2]; E.K. Speliotes [1, 2]

View Session Add to Schedule

Affiliations:
1) Division of Gastroenterology and Hepatology, University of Michigan, Ann Arbor, MI.; 2) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI.

Background: Alanine transaminase (ALT) and aspartate transaminase (AST) are commonly-used indicators of hepatocellular liver disease. While AST and ALT levels are partially genetically-determined, the specific genetic variants and biological etiology underlying this variation remain incompletely understood. **Methods**: We used Scalable and Accurate Implementation of GEneralized mixed model to carry out GWAS of inverse normally-transformed ALT or AST, controlling for age, sex, and principal components in >390,000 white individuals from the UK BioBank (>20,717,468 imputed SNPs), followed by meta-analysis with ALT and AST GWAS results from >130,000 individuals of Japanese ancestry from BioBank Japan (5,951,600 imputed SNPs) using METAL. We examined effects of genome-wide significant ($P < 5x10^{-8}$) SNPs on quantitatively measured nonalcoholic hepatic steatosis (GOLD consortium) and alcohol-related cirrhosis (PMID 26482880) ($P < 0.05$ reported), and 9 UK BioBank traits/diagnoses (FDR < 0.05 reported). DEPICT analyses were carried out with SNPs with $P < 1x10^{-5}$ and tissues with FDR enrichment < 0.05 are reported. **Results**: 229 and 291 SNPs associated ($P < 5x10^{-8}$) with ALT and AST respectively consisting of 331 independent loci, 314 of which are novel. Many ALT-increasing SNPs increased nonalcoholic hepatic steatosis, including one near *HECTD4* not previously associated with liver disease in the population. Other SNPs increased alcohol-related cirrhosis, including a newly-identified SNP near *ZNF777*. Other novel variants were in or near genes involved in liver drug metabolism (*CYP2A6*), lamin biology (*MLIP*), sterol metabolism (*ABCG8*), and phospholipid metabolism (*ABCB4*). ALT-increasing SNPs had divergent effects on cardiometabolic traits: some promoted pan-deleterious abnormalities with high serum triglycerides, serum LDL-cholesterol, and waist-to-hip ratio (PPARG), while others produced a favorable plasma lipid profile (*TM6SF2*) or decreased waist-to-hip ratio (*C5orf67*). DEPICT analyses revealed enrichment for liver, small/large intestine, and adipose tissue among ALT-increasing SNPs. **Conclusions**: We identified 314 novel SNPs that reproducibly elevate ALT/AST in UK BioBank and BioBank Japan. These variants identify genes with diverse metabolic effects and mechanisms of liver disease, which may guide more precise diagnosis and treatment of known and new liver diseases.

A- A+

PgmNr 318: Massively parallel functional profiling of *CYP2C9* variants using a yeast activity assay.

Authors:
C.J. Amorosi [1]; L.H. Wong [1,3]; K.A. Sitko [1,4]; M.G. McDonald [2]; A.E. Rettie [2]; D.M. Fowler [1]; M.J. Dunham [1]

View Session Add to Schedule

Affiliations:
1) Department of Genome Sciences, University of Washington, Seattle, WA.; 2) Department of Medicinal Chemistry, University of Washington, Seattle, WA.; 3) LabGenius, London, UK.; 4) Juno Therapeutics, Seattle, WA.

The field of pharmacogenomics is currently overwhelmed by the huge amount of genetic variation being discovered by new sequencing efforts. One key limitation is the lack of corresponding functional annotation of these gene variants that would allow the field to link them to clinically actionable drug responses. We are addressing this problem in a particularly important pharmacogene: *CYP2C9*. *CYP2C9* encodes an enzyme responsible for metabolizing many different drugs including warfarin, phenytoin, and flurbiprofen, and genetic variants in this gene are known to affect the efficacy of these and other drugs. We have developed a yeast-based activity assay to test thousands of variants in a pooled fashion using a deep mutational scanning approach. Yeast has been used as a model system for recombinant P450 expression for over 30 years and can be engineered to express highly active human P450 enzyme. Our yeast assay, which uses activity-based protein profiling, is able to recapitulate the activity of known variants in both individual and pooled tests. Briefly, humanized yeast cells expressing a single *CYP2C9* variant are bound in an activity-dependent manner by a modified CYP2C9 inhibitor that is then labeled via click chemistry with a fluorophore for cell sorting and sequencing. This is done in a massively parallel manner, such that the entire library of variants is sorted into bins based on activity level, and each bin is deep sequenced to determine variant frequency. We have created a barcoded library of CYP2C9 single amino acid variants and have determined variant activity scores using our yeast-based assay and deep sequencing. So far, we have generated activity scores for 70% of the 9,800 possible CYP2C9 single amino acid variants. Preliminary analysis shows that roughly 60% of missense variants tested have significantly decreased activity and may have altered drug metabolism. We are in the process of validating individual variants using gold standard *in vitro* metabolic assays with clinically relevant substrates to determine the reliability of our data. Our approach will lead to advances in adverse drug response prevention by providing *CYP2C9* clinical guidance for patients carrying both currently known and yet-to-be discovered alleles.

A- A+

PgmNr 319: Targeted in-frame deletion of the F8 B domain to restore FVIII function in patient-iPSCs derived ECs.

Authors:
D. Liang; Z. Hu; M. Zhou; L. Wu

View Session Add to Schedule

Affiliation: Center for Medical Genetics, School of Life Sciences, Central South University, 110 Xiangya Road, Changsha, Hunan, China

FVIII protein includes three A domains, one B domain and two C domains. Given that the *F8* cDNA is too large to be packaged into AAV capsids, addition of B domain (*F8*-B) deleted *F8* has shown therapeutic effects in animal models and clinical trials. To date 475 mutations have been recorded within *F8*-B, and more than 90% of the *F8*-B mutations result in premature termination and cause severe haemophilia A (HA). Here we described a genetic correction strategy via an in-frame deletion with CRISPR/Cas9 to restore the function of FVIII in HA patient-iPSCs.We firstly reframed a frameshift mutation in *F8*-B to restore the *F8* transcript and FVIII function via a targeted deletion of a few base pairs using CRISPR-Cas. And then, to expand this strategy to cover all mutations within *F8*-B, the entire *F8*-B was deleted. The *F8*-corrected iPSCs were differentiated into endothelial progenitor cells (EPCs) and endothelial cells (ECs), in which the clotting activity of the corrected FVIII was investigated *in vitro* and *in vivo*. The bleeding phenotype was rescued in HA mice after transplantation of the *F8*-corrected EPCs. Our results demonstrate an efficient approach for targeted gene correction via introduction of in-frame deletion in *F8*-B to restore the *F8* transcript and FVIII function in patient-iPSC derived ECs with potential clinical impact in HA gene therapy. And for the first time, we demonstrated *in vitro* and *in vivo* the FVIII function that is encoded by the endogenous B domain-deleted *F8*.

A- A+

PgmNr 320: A viable mouse model of *cblC* deficiency displays growth failure and reduced survival which are rescued by hydroxocobalamin and AAV gene therapy.

Authors:
J.L. Sloan [1]; K.C. Murphy [1]; M. Arnold [1]; N.P. Achilly [1]; G. Elliot [2]; P. Zerfas [3]; V. Hoffmann [3]; B.P. Brooks [4]; D. Watkins [5]; D.S. Rosenblatt [5]; C.P. Venditti [1]

View Session Add to Schedule

Affiliations:
1) MGMGB, NHGRI, Bethesda, Maryland.; 2) Mouse and ES Core Facility, NHGRI, Bethesda, MD; 3) Office of Research Services Branch, OD, NIH, Bethesda, MD; 4) Genetics and Visual Function Branch, NEI, NIH, Bethesda, MD; 5) Department of Human Genetics, McGill University and Research Institute McGill University Health Centre, Montreal, Quebec

---

Combined methylmalonic acidemia and homocysteinemia *cblC* type (*cblC*) is the most common inborn error of cobalamin metabolism. Disease manifestations include poor survival, growth failure, anemia, neurocognitive impairment and a progressive maculopathy and retinopathy. Early identification by newborn screening and lifelong daily injections of hydroxocobalamin (OHCbl) improve survival and some disease related complications but neurological and visual impairment persist despite therapy. To explore disease pathophysiology and develop novel therapies, we created a mouse model of *cblC* using TALENs targeting exon 2 of *Mmachc* focusing our studies on two alleles: c.165_166del p.Pro56Cysfs*4 (del2) and c.162_164del p.Ser54_Thr55delinsArg (del3). We observed a decreased number of homozygous mutant pups at birth ($p<0.05$) with both alleles, but mutant embryos were present in expected ratios at E18.5 and manifested IUGR. The median survival was 5 days with complete lethality by 1 month (del2/del2 n=14; del3/del3 n=66; $p<0.0001$). At 2 weeks, *Mmachc*$^{del3/\mathrm{del3}}$mutants weighed 35% less than littermates (n=12; $p<0.0001$) and displayed the characteristic cellular and biochemical features of *cblC*, including impaired synthesis of AdoCbl and MeCbl, significantly elevated plasma methylmalonic acid and homocysteine, and decreased methionine compared to controls (n=5-10, $p<0.05$). Pathological examination of *cblC* mice at 1 month revealed thinning of the corpus callosum, dilated ventricles, hepatic lipidosis and testicular hypoplasia. To explore systemic gene therapy as a treatment for *cblC*, we generated two AAV vectors: rAAVrh10-CBA-m*Mmachc* and rAAV9-CBA-h*MMACHC* that were delivered by a single intrahepatic neonatal injection (1 x 10^{11}GC/pup) and compared to treatment with weekly prenatal and prenatal+postnatal OHCbl injections. Dramatically improved clinical appearance and increased survival was observed in mutants treated with AAVrh10 (n=9), AAV9 (n=11), prenatal+postnatal OHCbl (n=16), and prenatal OHCbl+AAV9 (n=11) ($p<0.0001$), with treated mutants living beyond 1 year. [MMA] was reduced at 6-12 months with prenatal OHCbl+AAVrh10 (182 µM; n=3) and prenatal OHCbl+AAV9 (137 µM; n=3), but not with prenatal+postnatal OHCbl (>1500 µM; n=4) treatment. In summary, we developed the first viable animal model of *cblC* which recapitulates several disease manifestations and describe successful treatment with both OHCbl and AAV gene transfer, a novel therapeutic approach for this disorder.

A- A+

PgmNr 321: The evolutionary history of 2,658 cancers.

Authors:
P. Van Loo [1]; C. Jolly [1]; I. Leshchiner [2]; S.C. Dentro [3]; S. Gonzalez [4]; P.T. Spellman [5]; D.C. Wedge [6]; M. Gerstung [1]; the PCAWG Evolution and Heterogeneity Working Group and the PCAWG network

View Session Add to Schedule

Affiliations:
1) The Francis Crick Institute, London, United Kingdom; 2) Broad Institute of MIT and Harvard, Cambridge, USA; 3) Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 4) European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom; 5) Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA; 6) Big Data Institute, University of Oxford, Oxford, United Kingdom

---

Over the course of a lifetime, our cells accumulate DNA damage in the form of somatic mutations. The varied fitness effects of these mutations expose cells to natural selection; whilst many are neutral, some may be advantageous and result in clonal expansion. Recent work has shown that this process of mutation, selection and clonal expansions is already happening in phenotypically normal somatic tissues. When this process is left unchecked, this can result in the development of cancer. Little is known, however, about this transition from normal to cancerous tissue, either in terms of timescale, or in the sequence of genomic changes.

Although whole genome sequencing provides a snapshot of the cancer genome at diagnosis, it is possible to use the relationships between mutations to infer partial orderings of events during a tumour's evolutionary past. Furthermore, as the mutations themselves bear signatures of their underlying causes, one can examine the changing activity of mutational processes over time, and use clock-like signatures to approximate real time. Applying and integrating such analyses across 2,658 cancer genomes from 39 cancer types, we create typical timelines of tumour evolution, reminiscent of the mutational ordering proposed by Fearon and Vogelstein for colorectal adenocarcinoma, but with considerable additional detail, and related in real time to the point of diagnosis.

Our real-time analyses show that major events, such as whole genome duplication (WGD), typically occur years, if not decades, before diagnosis. Pan-cancer, events preceding WGD include biallelic inactivation of canonical driver genes, signatures of exogenous mutagens (smoking, UV light), and distinctive copy number changes in certain cancer types. Post-WGD, and in subclonal evolution, we observe increased chromosomal instability, defective repair of DNA, and a greater diversity of drivers. Taken together, these analyses give substantial insight into the biology of tumourigenesis, and highlight the potential for earlier detection and therapeutic intervention.

A- A+

PgmNr 322: Unexpectedly high rates of recurrent somatic mutations in cancer genomes reveal mutational processes, the structure and behavior of the genome to stress and allow clinically relevant classification.

Authors:
I.G. Gut [1,2]; M.D. Stobbe [1]; G.A. Thun [1]; J.P. Whalley [1]; M. Oliva [1]; E. Raineri [1]

View Session Add to Schedule

Affiliations:
1) Centro Nacional de Analisis Genomico (CNAG), Center for Genomic Regulation, Barcelona Instiitute of Technology (BIST), Barcelona, SPAIN; 2) Universitat Pompeu Fabra, Barcelona, SPAIN

By analyzing somatic mutations in 2,583 cancer genomes, of the Pan-Cancer Analysis of Whole Genomes project of the International Cancer Genome Consortium (ICGC), from patients with 39 different tumor types from many different tissues, we detected a far higher rate of exactly the same shared somatic mutations being present in different tumors, within and across tumor types, than expected by chance. These recurrent somatic mutations are 1,057,935 (2.4%) of the ~43.4 million Somatic Single-base Mutations (SSMs) in the cohort, and 186,576 (8.7%) out of the ~2.1 million Somatic Insertion/deletion Mutations (SIMs). This suggests that non-random mutational processes are acting upon specific parts of the genome. To prove this hypothesis, the characteristics of the somatic mutations were captured with 42 features, nine related to recurrence. These features served as input for principal components analysis followed by hierarchical clustering on the resulting principal components. We obtained 16 statistically sufficiently powered clusters. The eight main clusters capture different mutational processes with varying levels of different types of recurrent mutations reflected in them. One of the clusters captures microsatellite instable tumors and is characterized by a high percentage of recurrent SIMs and of 1 bp C/G deletions in the context of a 5-10 bp C/G homopolymer. A cluster dominated by esophagus adenocarcinoma samples showed high levels of recurrent T>G and T>C SSMs, which potentially is linked to DNA damage caused by gastric reflux. A cluster of mostly lung cancer samples is characterized by a high level of C>A SSMs, a high percentage of 1 bp C/G deletions and a low level of recurrence. Using these results we recapitulate several of the known cancer genome processes, but importantly have unearthed novel mutational manifestations that we are able to marry with new processes and potential mechanisms to explain them. The clusters also allow us to attribute new tumors to clinically relevant classes using a single analytical test – whole-genome sequencing.

A- A+

PgmNr 323: Integrated analysis of NGS and optical mapping resolves the complex structure of highly rearranged focal amplifications in cancer.

Authors:
J. Luebeck [1]; V. Deshpande [2]; C. Coruh [3]; S. Raisi [2]; D. Pai [4]; S. Wu [5]; C.S. Zhang [1]; K.M. Turner [5]; J.A. Law [2,6]; P.S. Mischel [5,7]; V. Bafna [2]

View Session Add to Schedule

Affiliations:
1) Bioinformatics & Systems Biology, UC San Diego, La Jolla, California.; 2) Department of Computer Science & Engineering, UC San Diego, La Jolla, California.; 3) Salk Institute for Biological Studies, La Jolla, California; 4) BioNano Genomics, San Diego, California; 5) Ludwig Cancer Institute, UC San Diego, La Jolla, California.; 6) Division of Biological Sciences, University of California, San Diego, La Jolla, California; 7) Department of Pathology, University of California at San Diego, La Jolla, California

Copy number amplifications (CNA) are a hallmark of the cancer genome. The increase in copy number of oncogenes on focally amplified regions imparts positive selective pressure that mediates the rapid proliferation of cells. Presence of such focal amplifications has also been associated with genome instability and increased pathogenicity. Despite their importance, the mechanisms causing focal CNAs are incompletely understood. Proposed mechanisms include chromosomal translocation with duplication, chromothripsis and others. New reports suggest circular extrachromosomal DNA (ecDNA) exists in up to 40% of cancer types, and are an important driver for focal CNA. Thus, methods which reconstruct focal CNAs will enable a greater understanding of cancer biology.

An earlier tool to analyze focal CNAs (Deshpande 2019), used next-generation sequencing (NGS) data to create a graph encoding the rearrangement breakpoints, as a prelude to identifying the full structure. Paths and cycles extracted from breakpoint graphs provide the signatures of the rearrangement events, but are complex and rarely admit an unambiguous structure due to the complexity of focal CNAs. Here, we present a method, Amplicon Reconstructor (AR), that integrates NGS data with optical mapping (OM) data or long-read data. OM data provides physical maps of DNA which can be assembled into ultra-long OM assemblies (N50 ~50 Mbp), providing larger scaffolds than other long-read strategies. AR then employs a graph-based method to identify long paths and cycles in a breakpoint graph. Extensive simulations validate AR reconstructed amplicons, demonstrating the high fidelity of our approach.

We applied AR to NGS & OM data from seven patient-derived and immortalized cell lines, and used multiple cytogenetic approaches to validate our findings. Our method reconstructed complete circular ecDNA structures, each larger than 1 Mbp, in three glioblastoma cell lines, and identified locations where the ecDNA had reintegrated into chromosomes. In K562 cells, AR identified a complex rearrangement including the chr9-chr22 BCR-ABL fusion, and also genomic regions from chr13. In the HCC827 cell line, AR enabled reconstruction of a breakage fusion bridge structure. Finally, AR can suggest alternate reconstructions indicative of cell to cell heterogeneity. Taken together, our results provide complete reconstructions giving new insight into focal CNA structure unavailable from measurements of breakpoints and copy numbers alone.

A- A+

PgmNr 324: Detailed modeling of positive selection significantly improves detection of cancer driver genes.

Authors:
S. Zhao; X. He; M. Stephens

View Session Add to Schedule

Affiliation: Human Genetics, Univ Chicago, Chicago, Illinois.

Identifying driver genes is a central problem in cancer biology, and many methods have been developed to identify driver genes from somatic mutation data. However, existing methods either lack explicit statistical models, or rely on very simple models. Here, we present driverMAPS (Model-based Analysis of Positive Selection), a more comprehensive model-based approach to driver gene identification. This new method explicitly models, at the single-base level, the effects of selection in driver genes, as well as highly heterogeneous background mutational processes. The selection model captures elevated mutation rates in functionally important sites using multiple external annotations, as well as spatial clustering of mutations. The background mutation model accounts for variation in mutation rates due to both known covariates and unknown factors. Simulations under realistic evolutionary models demonstrate that driverMAPS greatly improves power to detect driver genes compared with current state-of-the-art approaches. Applying driverMAPS to TCGA data from 20 tumor types identifies 159 new potential driver genes. Cross-referencing this list with data from external sources provides further support for these findings. The novel genes include the mRNA methyltransferase METTL3-METTL14, and we experimentally validated METTL3 as a potential tumor suppressor gene in bladder cancer. These results provide strong support for the emerging hypothesis that mRNA modification is an important biological process underlying tumorigenesis.

A- A+

PgmNr 325: Detecting cancer vulnerabilities through gene networks under purifying selection.

Authors:
H. Horn [1]; A. Gupta [1, 3]; C. Fagre [1]; P. Razaz [1,4]; A. Kim [1]; M. Lawrence [1,5]; G. Getz [1,5]; J.T. Neal [1]; K. Lage [1,2]

View Session Add to Schedule

Affiliations:
1) Broad Institute, Cambridge, Massachusetts.; 2) Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston; 3) Departments of Biology and Computer Science, Massachusetts Institute of Technology, Cambridge; 4) Center for Genomic Medicine and Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston; 5) Department of Pathology and MGH Cancer Center, Massachusetts General Hospital, Boston

---

High throughput gene knockout studies have significantly advanced over the last years, to the extent of being applied to hundreds of distinct cell lines. The generated data constitutes a remarkable resource to gain novel insights into gene essentiality and the underlying molecular mechanisms.

One of the major constraints with current high throughput screening approaches is the need for viable cell lines. This leads to a lack of data for a) cell lines that cannot be cultivated in vitro and b) patient specific cell lines, as isolating and propagating cells from patient samples is still a non-trivial task.

Existing computational methods that are routinely used to calculate the mutational burden (e.g. MutSig) are underpowered to detect genes under purifying selection (genes that are significantly depleted for mutations), given currently available sample sizes.
We have recently shown, that aggregating the mutational signal over a gene’s neighborhood is a viable way to improve the statistical power to identify novel candidate cancer genes (NetSig).

Based on these results, we have extended the method to detect neighborhoods depleted for mutations and thereby, genes under purifying selection. We integrate a protein-protein network of ~17.000 genes with mutational data from ~4500 tumor samples to nominate 86 high confidence candidate genes. These 86 candidate genes are strongly enriched in known essentiality gene sets (1.5x in ExAC pLi >= 0.9, 3x enrichment in high throughput CRISPR-based gen knock-out studies and Cyclops genes) and known drug targets.
To further validate our candidate genes experimentally, we are screening a subset of 8 cell lines (not yet part of the publicly available high throughput screenings) using both shRNA gene knock-down and CRISPR based gene knock-outs.

A- A+

PgmNr 326: Inferring clonal lineages and mutational chronologies in triple negative breast cancer using high-throughput single cell DNA sequencing.

Authors:
J. Leighton [1,2]; M. Hu [2]; A. Davis [1,2]; E. Sei [2]; Y. Wong [3]; N. Navin [1,2,4]

View Session Add to Schedule

Affiliations:
1) MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX; 2) Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX; 3) Department of Biomolecular Chemistry, University of Wisconsin, Madison, WI; 4) Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX

Triple negative breast cancer (TNBC) is an aggressive subtype of breast cancer with high rates of metastasis and recurrence, where TNBC patients have poor 5-year survival and ~50% are non-responsive to chemotherapy. Aneuploidy is a cancer hallmark that is pervasive in over 90% of breast cancer patients and is indicative of complex genomic rearrangements that are acquired during tumor initiation. Although copy number aberrations have been extensively studied in relation to aneuploidy and TNBC initiation, little is currently known regarding the timing and impact of single nucleotide variants (SNVs) contributing to these early transformative genomic events. Paramount to novel treatment options is understanding the underlying biology of initiation in the early stages of TNBC development, where inferring clonal lineages and mutational chronologies can help characterize the order, timing, and potential impact of single point mutations, such as *TP53*, in association with early mechanisms of genomic instability, genome doubling, and aneuploid transformation. We developed high-throughput DNA sequencing methods to delineate clonal lineages and order point mutations utilizing bulk deep-sequencing to first create a patient specific gene panel covering ~330 genes over 5 invasive ductal carcinoma TNBC patient samples collected at MD Anderson Cancer Center. After isolating the respective diploid and aneuploid populations of each tumor by flow cytometry, single cell sequencing was applied by employing droplet-based microfluidics (Mission Bio) and our novel patient specific gene panel to profile point mutations in ~23,500 cells across the 5 TNBC patients. Computational clustering analyses and phylogenetic tumor tree modeling of massive-scale high dimensional data revealed extensive tumor heterogeneity, primarily linear clonal lineages of point mutations acquired in successive “blocks”, and truncal heterozygous *TP53* mutations across all patients, thereby offering compelling insights into the timing, correlation, and clonal distribution of these SNVs. Resolving the order, timing, and evolutionary relationships of early point mutations, in association with *TP53*, provides new insights into the compelling biology behind transformative genomic events in TNBC initiation and progression, thereby inspiring dynamic basic and translational genomics research to further develop early detection approaches, preventative treatments, and precision medicine therapies in breast cancer.

A- A+

PgmNr 327: Extreme methylation variation from whole-genome bisulphite sequencing among patients with negative clinical exomes.

Authors:
T. Pastinen; W. Cheung; N. Miller; S. Herd; J. Johnston; I. Thiffault; E. Farrow; A. Walter; M. Gibson; S. Younger; C. Berrios; E. Grundberg; D. Zhou; C. Lawson; L. Grote; S. Hughes; C. Schwager; E. Fleming; K. Engleman; L. Cross; J. Kussman; H. Welsh; J. Jenkins; R. Gadea; M. Strenk; J. Gannon; S. Amudhavali; E. Rush; B. Heese; C. Saunders

View Session Add to Schedule

Affiliation: Center for Pediatric Genomic Medicine (CPGM), Children's Mercy Kansas City, Kansas City, MO

Molecular diagnostic yields by symptom driven exome sequencing (ES) in patients with rare disease are only 25-35%. Additional technologies to enhance detection of disease variants genes include whole genome sequencing (WGS), and structural variation (SNV) analysis using new sequencing methods. Recently, microarray-based methylation analysis to explore perturbation of gene regulation in rare disease has been suggested. We are deploying a new molecular testing framework in pediatric rare disease. Whereby patients undiagnosed after clinical ES are recruited in a diverse rare disease cohort observed in clinical genetic service. We expand capacity for variant detection using WGS and 10X Linked Read sequencing for SNV and CNV detection along with whole genome bisulphite sequencing (WGBS) to access the functional effects of non-coding rare variants. Methylation outlier detection was first validated using known disease changes (eg. *FMR1* repeat expansion). Subsequently, with pediatric control and the first 100 rare disease proband samples with non-diagnostic clinical exomes we initiated comparative WGBS analyses. We applied three parallel analytical approaches: allele-specific methylation (ASM), extreme methylation level (z-score) and long-range methylation correlation changes (mCor) and explored general characteristics of methylation outliers. Among ~100 extreme ASM or z-score methylation variants observed per patient genome there is up to five-fold enrichment of very rare DNA variants close-by (gnomad MAF<0.005, <100bp distance). Furthermore, rare DNA variant linked CpG methylation is more commonly observed in DNaseI hypersensitive (DHS) sites (enrichment P<0.001) and the rare DNA changes proximal to methylation outliers are significantly enriched both at conserved transcription factor binding sites as well as at DHSs. Similarly, when long-range (10kb-3Mb) *cis*-correlation in methylation (r>0.8) between regulatory elements observed in the control population is lost (mCor outliers) the median distance to CNV is significantly shorter (340kb vs. 11Mb). Altogether, a subset of rare DNA variants and CNVs generate unusual local patterns in blood DNA methylation detected by WGBS, which impacts regulatory elements with a range of tissue specificity. Early overlap of phenotypically relevant disease genes (OMIM) shows hemizygous hypermethylation of conserved regulatory element in up to 10% of patients, providing novel prioritization of rare non-coding disease variants.

A- A+

PgmNr 328: Clonal hematopoiesis of indeterminate potential and epigenetic age acceleration.

Authors:
D. Nachun [1]; A. Lu [2]; A. Bick [3]; P. Natarajan [3]; D. Levy [4]; A. Reiner [5]; J. Wilson [6]; S. Horvath [2]; S. Jaiswal [1]; NHLBI Trans-Omics for Precision Medicine

View Session Add to Schedule

Affiliations:
1) Department of Pathology, Stanford University, Palo Alto, California.; 2) Department of Human Genetics, UCLA, Los Angeles, CA; 3) Massachusetts General Hospital, Boston, MA; 4) Population Sciences Branch, NHLBI, Framingham, MA; 5) Fred Hutchinson Cancer Research Center, University of Washington, Seattle, WA; 6) Mississippi Center for Clinical and Translational Research, University of Mississippi, Jackson, MS

Epigenetic clocks have shown that patterns of DNA methylation from blood cells are strongly correlated with chronological age. Those with accelerated methylation age (methylation age that is greater than expected for chronological age) are at higher risk for several diseases of aging and death, but the biological processes underlying such advanced epigenetic age are incompletely understood. Clonal hematopoiesis of indeterminate potential (CHIP) results from somatic mutations in blood stem cells and may be found in ~20% of the elderly. CHIP most commonly arises due to mutations in the DNA methylation altering enzymes, TET2 and DNMT3A, and also associates with increased risk of death, cancer, and cardiovascular disease. Whether CHIP associates with accelerated methylation age is unknown. We used methylation and whole genome sequencing data from several cohorts in TOPMed together comprising thousands of persons to show that CHIP is strongly associated with increased epigenetic age acceleration. The most consistent association is observed for intrinsic aging (2.8 ± 0.36 years, $p < 2.5 \times 10^{-14}$), which measures epigenetic aging that is independent of changes in cell composition, while a more variable association was seen with extrinsic aging (2.5 ± 0.46 years, $p < 4.6 \times 10^{-7}$), which captures epigenetic aging that is driven by changes in cell composition. We also analyzed the gene-specific effects of CHIP mutations on epigenetic aging, dividing our CHIP carriers into DNMT3A, TET2, and all other CHIP mutations. We found that the increase in intrinsic age acceleration seen in CHIP was very consistent across different genes, while the increase in extrinsic aging was lower with DNMT3A mutations and higher in TET2 mutations. The epigenetic clock software we used can also predict cell type composition and leukocyte telomere length (LTL) from methylation. We observed an increased predicted proportion of CD8+/CD28-/CD45RA- T-cells and decreased predicted LTL. Future experiments should seek to determine whether there is a causal relationship between CHIP and epigenetic age acceleration and whether intrinsic or extrinsic age acceleration is predictive of health outcomes in those with CHIP.

A- A+

PgmNr 329: Exploiting omics to measure genome-by-environment interactions.

Authors:
C. Amador [1]; Y. Zeng [1,2]; R. Walker [3,4]; A. McIntosh [3,5]; K. Evans [3,4]; D. Porteous [3,4]; A. Campbell [4]; C. Hayward [1]; P. Navarro [1]; C. Haley [1,6]

View Session Add to Schedule

Affiliations:
1) MRC Human Genetics Unit, University of Edinburgh, Edinburgh, United Kingdom; 2) Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-Sen University, China; 3) Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, United Kingdom; 4) Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, United Kingdom; 5) Division of Psychiatry, University of Edinburgh, Edinburgh, United Kingdom; 6) The Roslin Institute and Royal (Dick) School of Veterinary Sciences, University of Edinburgh, Edinburgh, United Kingdom

Most health-related outcomes and phenotypes are complex traits influenced by both genetic and environmental variation. Many past studies have utilised genomics to estimate the amount of disease variation driven by genetic variation, and to identify causal polymorphisms that contribute to individuals’ disease risk. Some of the sources of environmental variation are also known, e.g., depression risk is affected by stressful life events, and obesity-related traits are affected by diet and other lifestyle components. The ability to measure these environmental variables is crucial to achieve accurate predictions and to implement personalised medicine, however, in most studies, measures for environmental variables are obtained from questionnaires and are sometimes inaccurate. Variation in methylation at CpG sites has been associated with a range of environmental variables and has the potential to be used as a proxy for environmental measures. Here we use methylation variation as a proxy for tobacco usage. We used a subset of ~600 methylation CpG sites associated with smoking status, previously identified in independent studies, in a mixed linear model framework to measure the amount of variation in BMI explained by genomic and environmental information, either self-reported (from questionnaires) or via smoking-associated methylation variation. We also used these to calculate the amount of variation in BMI explained by genome-by-smoking interactions. Using data from 5,000 individuals from the Generation Scotland cohort, we estimated the heritability of BMI to be ~50% and an effect of smoking status (self-reported) explaining 2% of BMI variation. However 22% of BMI variation (on top of the 50% explained by genomic variation) was estimated to be driven by the effect of smoking-associated methylation. Genome-by-smoking interactions explained an extra 10% of BMI variation (in addition to the independent direct effects of genetics and the environment, i.e., including smoking). These results were consistent when modelled based on self-reported status or via smoking-associated methylation. Using methylation, we have shown that omics data can be used as a proxy for environmental variation and can be extremely useful when modelling environmental effects or genome-by-environment interactions.

A- A+

PgmNr 330: Cell-type specific DNA methylation QTL of primary melanocytes with multi-QTL integration enhances melanoma GWAS annotation beyond eQTL.

Authors:
J. Choi [1]; T. Zhang [1]; M. Kovacs [1]; M. Xu [1]; A. Vu [1]; S. Loftus [2]; W. Pavan [2]; D.T. Bishop [3]; A.J. Stratigos [4]; P. Ghiorzo [5]; K. Peris [6]; G.J. Mann [7]; G.L. Radford-Smith [8]; N.G. Martin [9]; S.V. Ward [10]; S. Puig [11]; D.L. Duffy [9]; F. Demenais [12]; E. Nagore [13]; D.C. Whiteman [14]; S. MacGregor [15]; M.T. Landi [1]; M.H. Law [15]; M.M. Iles [16]; J. Shi [1]; B. Pasaniuc [17]; K.M. Brown [1]; Melanoma Meta-Analysis Consortium

View Session Add to Schedule

Affiliations:
1) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland; 2) National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland; 3) Division of Haematology and Immunology, Leeds Institute of Medical Research, University of Leeds, Leeds, UK; 4) 1st Department of Dermatology – Venereology, National and Kapodistrian University of Athens School of Medicine, Andreas Sygros Hospital, Athens, Greece; 5) Department of Internal Medicine and Medical Specialties, University of Genoa and Genetics of Rare Cancers, Ospedale Policlinico San Martino, Genoa, Italy; 6) Institute of Dermatology, Catholic University, Rome, Italy; 7) Centre for Cancer Research, Westmead Institute for Medical Research; Melanoma Institute Australia; University of Sydney; 8) Inflammatory Bowel Diseases, QIMR Berghofer Medical Research Institute, Brisbane, Australia; 9) Genetic Epidemiology, The QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 10) Centre for Genetic Origins of Health and Disease (GOHaD), University of Western Australia; Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center; 11) Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain. & Centro de Investigación Biomédica en Red en Enfermedades Raras (CIBERER), Valencia, Spain; 12) Team of Genetic Epidemiology and Functional Genomics of Multifactorial Diseases, UMR-1124, INSERM, Université de Paris, Paris, France; 13) Department of Dermatology, Instituto Valenciano de Oncología, València, Spain; 14) Population Health, QIMR Berghofer Medical Research Institute, Brisbane, Herston, Australia; 15) Statistical Genetics, The QIMR Berghofer Medical Research Institute; 16) Division of Pathology and Data Analytics, Leeds Institute Medical Research, University of Leeds, Leeds, UK; 17) Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles

---

Expression quantitative trait loci (eQTL) analysis has been useful for annotating trait-associated loci from genome-wide association studies (GWAS). While tissue-level eQTL datasets are abundantly available, eQTLs alone usually explain a small minority of GWAS loci for most traits with typical colocalization rates well under 25%. Moreover, in the context of melanoma predisposition we have previously shown that a homogenous eQTL dataset focusing on the cell type of disease origin (melanocyte) considerably outperforms heterogeneous tissue-based eQTL datasets of larger sample sizes (skin tissue) in annotating risk-associated loci. To further improve QTL-based GWAS annotation and explore cell-type specific alternative QTLs, we performed multiple QTL analyses of primary melanocytes, focusing here on DNA methylation QTL (meQTL). We established a meQTL dataset using Illumina Human Methylation 450K BeadChips and the same set of cultured primary melanocytes (n=106) from which we generated an eQTL dataset. From this, we identified 13,263 significant eProbes for *cis*-meQTL (QTLtools, FDR < 0.05). To further facilitate integration of multiple types of QTLs, we also performed splicing (sQTL) and microRNA QTL (miQTL) analyses from the same samples. To identify potential susceptibility genes and *cis*-regulatory mechanisms underlying melanoma risk, we utilized data from a new melanoma GWAS meta-analysis of 36,760 cases and 375,188 controls which identified 54 loci (3X more loci than previous studies). Using these data, we performed colocalization analyses (HyPrColoc) incorporating multi-QTL datasets. With meQTL alone, we observed colocalization for 24 of 54 melanoma GWAS loci. When combined with eQTL, sQTL, and miQTL, 30 loci displayed colocalization with ≥ 1 QTL dataset, with 24 of them showing posterior probability > 0.8. For 12 of those 30 loci, colocalization was observed for both meQTL and eQTL. Thus, by adopting a cell-type specific multi-QTL approach, we vastly improved on annotation compared to eQTLs alone, finding colocalizing QTLs for a majority (56%) of loci from the largest melanoma GWAS performed to date. We are presently conducting additional analyses including methylome-wide association study and mediation analyses between DNA methylation and mRNA expression, which we anticipate will further enhance our understanding of the genes and regulatory mechanisms underlying melanoma susceptibility, as well as melanocyte-specific gene expression regulation patterns.

A- A+

PgmNr 331: Estimation of total mediation effect for multiple types of high-dimensional omics mediators in over 3500 individuals provides novel insight into aging-related variation in blood pressure.

Authors:
Y. Zhao [1]; T. Yang [2]; J. Zou [1]; Z. Wang [1]; J. Niu [3]; H. Chen [4]; P. Wei [5]

View Session Add to Schedule

Affiliations:
1) Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center UTHealth, Houston, Texas.; 2) Division of Biostatistics, The University of Minnesota, Minneapolis, MN; 3) Section of Nephrology, Baylor College of Medicine, Houston, TX; 4) Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, TX; 5) Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX

Environmental exposures can regulate intermediate molecular phenotypes, such as the methylome, transcriptome and metabolome, by different mechanisms and thereby lead to different health outcomes. It is of significant scientific interest to unravel the role of potentially high-dimensional intermediate phenotypes in the relationship between environmental exposure and traits. Mediation analysis is an important tool for investigating such relationships. However, it has mainly focused on low-dimensional settings and there is a lack of a good measure of the total mediation effect. Here, we extend an R-squared (Rsq) effect size measure, originally proposed in the single-mediator setting, to the moderate- and high-dimensional mediator settings in the mixed model framework. Using extensive simulations we demonstrate appealing operating characteristics of the proposed Rsq measure of total mediation effect and the estimation procedure. By applying the proposed estimation procedure to the Framingham Heart Study (FHS) of 1655 individuals, we found that 82% (95% confidence interval (CI) = [54%, 100%]), 55% ([33%, 90%]) and 41% ([10%, 83%]) of the aging-related variation in systolic blood pressure can be explained by the methylomics, mRNA expression and metabolomics profile, respectively, whereas the microRNA expression profile was not found to be a significant mediation mechanism between aging and blood pressure. Furthermore, these findings in the FHS were replicated in the Women's Health Initiative (WHI) study of 1867 individuals. Finally, we have developed an R package "RsqMed" to implement the proposed novel mediation measure and estimation procedure.

A- A+

PgmNr 332: *TET3* deficiency: Delineation of the first human Mendelian disorder of the DNA demethylation machinery.

Authors:
J.A. Fahrner [1]; A. Petracovici [2,3]; C. He [2,3,4]; H.W. Moore [5]; R. Louie [5]; M. Ansar [6]; R.L.P. Santos-Cortez [7]; E.J. Prijoles [5]; R. Bend [5]; B. Keren [8]; C. Mignot [8,9]; M.C. Nougues [8]; K. Õunap [10]; T. Reimand [10]; S. Pajusalu [10]; J. Buratti [8]; E.G. Seaby [11,12]; K. McWalter [13]; A. Telegrafi [13]; T. Cottrell [14]; S. Sithambaram [14]; S. Douzgou [14,15]; D. Baldridge [16]; M. Shinawi [16]; S.M. Leal [17]; G.B. Schaefer [18]; R. Stevenson [5]; S. Banka [14,15]; R. Bonasio [2,3]; D.B. Beck [19]; Deciphering Developmental Disorders study

View Session Add to Schedule

Affiliations:
1) Department of Pediatrics, Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA; 2) Department of Cell and Developmental Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; 3) Epigenetics Institute, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; 4) Hunan Key Laboratory of Plant Functional Genomics and Developmental Regulation, Hunan University, Changsha, Hunan, P.R. China; 5) Greenwood Genetics Center, Greenwood, SC, USA; 6) Department of Biochemistry, Faculty of Biological Sciences, Quaid-I-Azam University, Islamabad, Pakistan; 7) Department of Otolaryngology, University of Colorado School of Medicine, Aurora, CO, USA; 8) Assistance Publique-Hôpitaux de Paris, Groupe Hospitalier Pitié-Salpêtrière, Département de Génétique, Paris, France; 9) Centre de Référence Déficiences Intellectuelles de Causes Rares, Paris, France; 10) Department of Clinical Genetics, United Laboratories, Tartu University Hospital, Tartu, Estonia; 11) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 12) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA; 13) GeneDx, Gaithersburg, Maryland, USA; 14) Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University National Health Service Foundation Trust, Health Innovation Manchester, Manchester, UK.; 15) Division of Evolution & Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK; 16) Department of Pediatrics, Division of Genetics and Genomic Medicine, Washington University School of Medicine, St. Louis, MO, USA; 17) Center for Statistical Genetics, Gertrude H. Sergievsky Center, Taub Institute for Alzheimer’s Disease and the Aging Brain, Department of Neurology, Columbia University Medical Center, New York, NY, USA; 18) University of Arkansas for Medical Sciences, Lowell, Arkansas, USA; 19) National Human Genetics Research Institute, National Institutes of Health, Bethesda, MD, USA

DNA methylation and post-translational modifications of histone tails play essential roles in development by dynamically regulating chromatin structure and gene expression. Mendelian disorders of the epigenetic machinery, or chromatin modifying disorders, are inherited conditions that disrupt these processes and account for a substantial proportion of neurodevelopmental and growth abnormalities in children. Most known disorders in this class are caused by pathogenic variants in histone-modifying enzymes and chromatin remodelers. Far fewer have been linked to deficiencies in the DNA methylation machinery. The latter include disorders caused by defects in DNA methyltransferase “writers” that place 5-methylcytosine and “readers” that interpret it. Cytosine methylation of DNA is the quintessential epigenetic mark, yet no human disorder of DNA

demethylation has been delineated. Here, we describe in detail the first Mendelian disorder caused by disruption of DNA demethylation, *TET3* deficiency. TET3 is a methylcytosine deoxygenase that initiates DNA demethylation during early zygote formation, embryogenesis, and neuronal differentiation and is intolerant to haploinsufficiency in mice and humans. We identify and characterize 11 cases of human *TET3* deficiency in 8 families with the common phenotype of intellectual disability (ID)/ global developmental delay, hypotonia, autistic features, movement disorders, growth abnormalities, and facial dysmorphism. Mono-allelic frameshift variants in *TET3* occur throughout the coding region, whereas most mono-allelic and bi-allelic missense variants localize to conserved residues within the catalytic domain and in most cases display hypomorphic function in a catalytic activity assay. *TET3* deficiency shows substantial phenotypic overlap with other Mendelian disorders of the epigenetic machinery, including ID, other neurobehavioral findings, and growth abnormalities, underscoring shared disease mechanisms. By describing in detail for the first time a deficiency in the DNA demethylation pathway, our work defines a novel biochemical category of epigenetic machinery disorders and expands our knowledge of this important group of diseases. Ongoing work to further characterize *TET3*-deficient individuals, their causative variants, and resulting molecular perturbations, including genome-wide DNA methylation analysis, will lead to a deeper understanding of the role of DNA methylation and demethylation in human development and disease.

A- A+

PgmNr 333: Transfer learning significantly improved clustering accuracy in single-cell RNA-seq analysis.

Authors:
J. Hu [1]; X. Li [2]; G. Hu [3]; M. Li [1]

View Session Add to Schedule

Affiliations:
1) Philadelphia, Pennsylvania.Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.; 2) Center for Applied Statistics, School of Statistics, Renmin University, Beijing, China.; 3) School of Mathematical Sciences, Nankai University, Tianjin, China.

Recent development of single-cell RNA-seq (scRNA-seq) technologies has led to enormous biological discoveries, yet also introduced computational challenges. An important step in scRNA-seq analysis is to cluster cells into different types. Existing scRNA-seq clustering methods often suffer from low accuracy for data with few cells or low sequencing depth. To overcome this limitation, it is appealing to borrow cell type knowledge learned from a well-studied source dataset to help cluster a new target dataset generated in a different study. Transfer learning, a machine learning method that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem, suits perfectly for this purpose. Here we describe t-DESC, a transfer learning algorithm for scRNA-seq clustering by deep embedding. t-DESC starts from building a training neural network to extract gene-expression signatures from a well-labeled source dataset. This step enables initializing the target network with parameters initialized by information learned from the training network. The target network then leverages information in the target dataset to fine-tune its parameters in an unsupervised manner, so that the target-data-specific gene-expression signatures can also be captured. Once fine-tuning is finished, the target network is then used to cluster cells in the target data. To experimentally showcase the strengths of t-DESC, we analyzed multiple scRNA-seq datasets. We show that with transferred cell type-specific gene expression information from the source data, the clustering accuracy for the target data was significantly improved, and this is true even for target data generated from a different platform, a different tissue, or a different species from the source data. For instance, transferring the knowledge gained from a human kidney non-immune cell dataset to a mouse kidney non-immune cell dataset increased the adjusted Rand Index (ARI) from 0.331 to 0.848. We further compared t-DESC with four other recently developed scRNA-seq analysis methods, including Moana, scmap, SAVER-X and Seurat 3.0. Comprehensive benchmark evaluations show that t-DESC is robust across different scenarios and shows consistently better performance than all other methods. With the increasing popularity of scRNA-seq in biomedical research, we expect t-DESC will substantially improve the accuracy in single-cell clustering, especially for small or low-quality datasets.

A- A+

PgmNr 334: SMNN: Batch effect correction for single-cell RNA-seq data via supervised mutual nearest neighbor detection.

Authors:
Y. Yang [1]; G. Li [2]; H. Qian [2]; Y. Li [1,3,4]

View Session Add to Schedule

Affiliations:
1) Genetics, Univ North Carolina at Chapel Hill, Chapel Hill, North Carolina.; 2) Statistics and Operations Research, Univ. North Carolina, Chapel Hill, North Carolina; 3) Biostatistics, Univ. North Carolina, Chapel Hill, North Carolina; 4) Computer Science, Univ. North Carolina, Chapel Hill, North Carolina

An ever-increasing deluge of single cell RNA-sequencing (scRNA-seq) data has been generated, often involving different time points, laboratories or sequencing protocols. Batch effect correction has been recognized to be indispensable when integrating scRNA-seq data from multiple batches. A recent study proposed an effective batch effect correction method by detecting mutual nearest neighbors (MNN) across batches and correcting batch effects accordingly. MNN has demonstrated superior performance over alternative methods. However, the original MNN method is unsupervised in that it ignores cluster label information of single cells, which can further improve effectiveness of batch effect correction, particularly under realistic scenarios where true biological differences are not orthogonal to batch effect. In this work, we propose SMNN: batch effect correction via supervised mutual nearest neighbor detection. SMNN either takes cluster/cell-type label information as input or infers cell types using scRNA-seq clustering. It then detects mutual nearest neighbors within matched cell types and corrects batch effect accordingly. To assess the performance of SMNN in real data, we applied it to two hematopoietic scRNA-seq datasets, generated using different sequencing platforms: MARs-seq and SMART-seq2 respectively. Compared to MNN, SMNN provides improved merging within the corresponding cell types across batches. Cells from the same cell type across the two batches are closer (2.8 - 8.2% reduction when measured by Euclidean distance) after SMNN correction than MNN. Furthermore, we also examined the ability to differentiate cell types after SMNN and MNN correction in three real datasets. In all the three datasets, clustering after SMNN correction shows improved (by 7.6 - 42.3%) Adjusted Rand Index (ARI) than MNN, suggesting that SMNN correction more accurately recovers cell-type specific features.

A- A+

PgmNr 335: A machine learning approach to enhance resolution of single-nucleus RNA-seq data by removing debris contamination.

Authors:
M. Alvarez [1]; E. Rahmani [2]; B. Jew [3]; K.M. Garske [1]; Z. Miao [1,3]; J.N. Benhammou [1,5]; C.J. Ye [4]; J.R. Pisegna [1,5]; K.H. Pietiläinen [6,7]; E. Halperin [1,2,3]; P. Pajukanta [1,3,8]

View Session Add to Schedule

Affiliations:
1) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; 2) Computer Science Department in the School of Engineering, UCLA, Los Angeles, CA, USA; 3) Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA; 4) Institute for Human Genetics, Department of Epidemiology and Biostatistics, Department of Bioengineering and Therapeutic Sciences, UCSF, San Francisco, USA; 5) Vache and Tamar Manoukian Division of Digestive Diseases, UCLA, Los Angeles, CA, USA; 6) Obesity Research Unit, Research Programs Unit, Diabetes and Obesity, University of Helsinki, Biomedicum Helsinki, Helsinki, Finland; 7) Obesity Center, Endocrinology, Abdominal Center, Helsinki University Central Hospital and University of Helsinki, Helsinki, Finland; 8) Institute for Precision Health, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

INTRODUCTION: Single-nucleus RNA-seq (snRNA-seq) measures gene expression in individual nuclei instead of cells, allowing for unbiased cell type characterization in solid tissues. We show that snRNA-seq is, however, commonly subject to contamination by extra-nuclear RNA, which leads to spurious clustering in downstream analyses if left uncorrected.
OBJECTIVE: Our goal was to develop a method to remove debris-contaminated droplets from snRNA-seq data to yield high quality clustering of cell types.
METHODS: We present a novel approach to remove debris-contaminated droplets in snRNA-seq experiments, called Debris Identification using Expectation Maximization (DIEM). Our method models background-contaminated droplets using a Dirichlet-multinomial distribution, and calculates a background score and principal components from gene expression. DIEM then runs semi-supervised EM on these features to classify candidate droplets as either debris or nuclei.
RESULTS: We evaluated DIEM using 3 snRNA-seq data sets from fresh human maturing adipocytes *in vitro*, fresh mouse brain tissue, and frozen human adipose tissue from 6 individuals. All 3 data sets showed a high prevalence of extra-nuclear RNA contamination leading to invalid cell types. We then compared DIEM with the existing methods SoupX and EmptyDrops (designed for single-cell RNA-seq). While correction with SoupX or EmptyDrops led to identification of true cell types, both methods still produced debris-contaminated clusters, similar to those seen without correction. Filtering using DIEM removed droplets containing high expression of background-enriched and mitochondrial genes. This led to removal of all false positive clusters marked by contamination and to more distinct cell types. We then analyzed the cell types identified after DIEM filtering in the adipocytes and adipose tissue. The maturing adipocyte nuclei consisted of 4 preadipocyte and 1 adipocyte clusters, while adipose tissue consisted of 1 adipocyte, 2 stromal, 2 vascular, and 5 immune cell types. We further integrated the *in vitro* maturing adipocyte data with *in vivo* adipose tissue cell types to map the identity of adipocyte progenitor cells. Consistent with a fibroblast-like gene expression profile, the preadipocytes

confidently mapped to a single stromal cell type.

CONCLUSIONS: Our new method DIEM effectively removed debris-contaminated droplets from snRNA-seq data, and enabled discovery of 10 cell-types from frozen adipose tissue.

A- A+

PgmNr 336: Cell-ID enables the identification of rare individual cells and their reproducible gene signatures across independent single-cell RNA-seq datasets.

Authors:
A. Rausell [1,2]; A. Cortal [1]

View Session Add to Schedule

Affiliations:
1) Paris Descartes University-Sorbonne Paris Cité, Imagine Institute, Clinical Bioinformatics Lab, Paris, France; 2) INSERM UMR 1163, Institut Imagine, Paris, France

Single-cell transcriptome profiling of patient's biological samples may help identifying abnormal rare cell fractions potentially associated to disease. Nonetheless, the computational identification of bona-fide rare cells, representing <2%, is challenged by high levels of biological and technical noise. Current computational methods for single-cell data analysis often rely on a low-dimensional representation of cells. The characterization of cell types is then typically carried out through a clustering step followed by differential gene expression analysis among groups. However, such approach is sensitive to batch effects that may compromise the replication of findings across independent donors. Moreover, an exhaustive exploration of cellular heterogeneity requires a per-cell gene signature assessment rather than a group-based analysis. Such possibility was lacking in the scientific literature.

Here we present Cell-ID, a robust statistical method that performs gene signature extraction and functional annotation for each individual cell in a single-cell RNA-seq dataset. Cell-ID is based on Multiple Correspondence Analysis and produces a simultaneous representation of cells and genes in a low dimension space. Genes are then ranked by their distance to each individual cell providing unbiased per-cell gene signatures. Such signatures proved valuable to estimate cell similarities across independent datasets and overcomed batch effects arising from different technologies, tissues-of-origin and donors. We evaluated Cell-ID on a diverse collection of single-cell RNA-seq libraries including blood cells, airway epithelial cells, as well as pancreatic islet cells from both healthy donors and Type 2 diabetes patients. Cell-ID correctly predicted well-established rare cell types at individual cell resolution. We then showed that the unbiased per-cell gene signatures obtained by Cell-ID allowed the identification of the corresponding cells of the same type both within and across independent datasets. Finally, we applied Cell-ID to pancreatic islet single-cell RNA-seq data where we illustrated the ability of our per-cell signature analysis to uncover functionally relevant cell heterogeneity that would have been missed by a clustering-based approach. Overall, we demonstrate that Cell-ID enables the robust identification of rare or even unique cells with a potential role in human disease. Cell-ID is freely distributed as an R package with a user-friendly shiny interface.

A- A+

PgmNr 337: Model selection-based scRNA-seq quantitation algorithm that controls overfitting and unwarranted imputation of technical zeros.

Authors:
K. Choi; D.A. Skelly; M.J. Vincent; G.A. Churchill

View Session Add to Schedule

Affiliation: The Jackson Laboratory, Bar Harbor, Maine.

Single-cell RNA sequencing is a powerful tool for characterizing cell heterogeneity and gene expression dynamics. But increased sampling bias and technical variability in current droplet-based single-cell RNA sequencing (dscRNA-seq) data poses analytical challenges. For example, high proportion of zero counts has been one of the major issues in modeling gene expression. The difficulty lies in delineating technical dropouts (which should be recovered if any) from biological zeros due to severe undersampling or stochastic gene expression. The current trend is to impute zeros using either data smoothing algorithms or mixture (zero-inflated) models. These approaches assumes that clustering has perfectly divided cells into homogeneous subtypes, and therefore, there is no more marker genes that further distinguishes subtypes within each cell type. On the other hand, more recent perspective argues that dscRNA-seq data is *not* zero-inflated since the numbers of zeros are not beyond expectation of negative binomial or multinomial models. There is still no strong consensus on whether to use zero-inflated models and, if necessary, how to best resolve dropouts and sampling zeros.

We developed a new scRNA-Seq quantitation method, PoolClass, in which we assess the predictive accuracy of fitted Poisson, negative binomial, and zero-inflated negative binomial models. These models are applicable to UMI counts directly and therefore no need of arbitrary preprocessing of counts, e.g., normalization across cells or log-transformation with added pseudocounts. Our model comparison enables to compute denoised *rates* of gene expression using the best model which each gene data is conforming to. We found, although minority, a considerable number of genes in droplet scRNA-seq data fit best with zero-inflated model, as opposed to multiple recent findings in which zero-inflation is not supported. The proportion of zero-inflated genes drops to ~10% within each cell type after clustering. We show our model selection approach controls high variability within each cell type and yet maintains cell-to-cell heterogeneity or gene expression stochasticity. We demonstrate that our approach improves the performance on downstream analyses such as identifying subpopulation of cells or detecting differentially expressed genes. We implemented our algorithm as an open-source Python package, PoolClass, that is available at https://github.com/churchill-lab/poolclass.

A- A+

PgmNr 338: Accurate estimation of cell composition in bulk expression through robust integration of single-cell information.

Authors:
B. Jew [1]; M. Alvarez [2]; E. Rahmani [3]; Z. Miao [1,2]; A. Ko [2]; J.H. Sul [1,4]; K.H. Pietiläinen [5,6]; P. Pajukanta [1,2,7]; E. Halperin [1,2,3]

View Session Add to Schedule

Affiliations:
1) Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA.; 2) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA.; 3) Computer Science Department in the School of Engineering, UCLA, Los Angeles, CA.; 4) Department of Psychiatry and Biobehavioral Sciences, UCLA, Los Angeles, CA.; 5) Obesity Research Unit, Research Program for Clinical and Molecular Metabolism, University of Helsinki, Helsinki, Finland.; 6) Obesity Center, Endocrinology, Abdominal Center, Helsinki University Central Hospital and University of Helsinki, Helsinki, Finland.; 7) Institute for Precision Health, David Geffen School of Medicine at UCLA, Los Angeles, CA.

INTRODUCTION: Transcriptomic analyses of tissues are an essential and now widely used method to study biology and disease pathophysiology. A limitation of these studies is the fact that most tissues are composed of a mixture of cell types, each contributing RNA to the total expression estimate. Early attempts have been made to estimate cell type proportions from bulk expression by leveraging information from reference datasets that include single-cell (or single-nucleus) RNA-seq on the same tissue. However, existing methods are limited in terms of their treatment of the inherent technological biases that differ between single-cell/single-nucleus RNA-seq and bulk RNA-seq.
METHODS: We present Bisque, a tool for estimating cell type composition from bulk RNA-seq data. Bisque includes two modes of operation: reference-based and marker-based. The reference-based approach utilizes single-cell/single-nucleus-based expression profiles and cell composition estimates to accurately decompose bulk expression, accounting for biases between bulk and single-cell/single-nucleus sequencing technologies. When a reference expression profile is not available, the marker-based approach provides estimates of cell type abundances using only known marker genes.
RESULTS: We applied our method to two bulk-tissue RNA-seq cohorts of human subcutaneous adipose (n=106) and dorsolateral prefrontal cortex (n=636) with single-nucleus RNA-seq data available for a subset of the samples (n=6 and n=8). In each dataset, Bisque outperformed existing methods (MuSiC, CIBERSORTx and BSEQ-sc) in accuracy and efficiency. Our reference-based method was able to estimate cell type proportions that matched expected physiological distributions; furthermore, these estimates recovered previously reported associations between cell type proportions and measured phenotypes. Our marker-based approach recovered these associations as well. Moreover, Bisque processed each dataset efficiently compared to existing methods (2.5x, 68.2x, and 1641.5x faster than MuSiC, BSEQ-sc, and CIBERSORTx on the cortex dataset, respectively).
CONCLUSIONS: Bisque provided a fast and accurate means of estimating cell type composition from bulk RNA-seq data using single-nucleus RNA-seq in both subcutaneous adipose and dorsolateral prefrontal cortex tissue.

A- A+

PgmNr 339: The genetic architecture of hematological traits within and between populations.

Authors:
G. Lettre [1,2]; M.H. Chen [3,4]; L. Raffield [5]; A. Mousas [1,2]; T. Jiang [6]; P. Akbari [6]; S. Sakaue [7]; E.L. Bao [8,9]; C.A. Lareau [8,9]; M. Chen [10]; C.W.K. Chiang [10]; Y. Okada [7]; V.G. Sankaran [8,9]; N. Soranzo [11]; A. Reiner [12]; A.D. Johnson [3,4]; P.L. Auer [13]; on behalf of the Blood-Cell Consortium

View Session Add to Schedule

Affiliations:
1) Montreal Heart Institute, Montreal, Quebec, Canada; 2) Universite de Montreal, Montreal. Quebec, Canada; 3) National Heart, Lung and Blood Institute, Bethesda, MD, USA; 4) The Framingham Heart Study, Framingham, MA, USA; 5) University of North Carolina, Chapel Hill, NC, USA; 6) University of Cambridge, Cambridge, UK; 7) Osaka University Graduate School of Medicine, Suita, Japan; 8) Broad Institute, Cambridge, MA, USA; 9) Children's Hospital Boston, Boston, MA, USA; 10) University of Southern California, Los Angeles, USA; 11) Sanger Institute, Hinxton, UK; 12) University of Washington, Seattle, WA, USA; 13) University of Wisconsin-Milwaukee, Milwaukee, WI, USA

The success in GWAS has been most impressive in studying phenotypic variation in populations of European ancestry (EA), leaving many open questions regarding the genetics of human traits and diseases in other populations. To address these questions, we undertook large trans-ethnic and ancestry-specific meta-analyses for blood traits in 746,667 participants, including 184,535 non-EA individuals. Because blood traits vary widely between populations, they represent ideal models to explore genetic differences across populations.

Our trans-ethnic meta-analyses identified 5,552 loci at genome-wide significance (P<5x10^{-9}), including 71 novel loci not found in EA populations. Bayesian fine-mapping found that trans-ethnic meta-analyses produced 95% credible sets (cs) that were 30% smaller than EA-only meta-analyses (Wilcoxon's P=3x10^{-4}). To gain functional insights, we used g-chromVAR to integrate the trans-ethnic 95% cs with ATACseq profiles from 18 human hematopoietic cell populations. The trans-ethnic variants were highly enriched for chromatin accessible regions (P=4x10^{-3}). These results will guide future functional experiments.

Next, we compared heritabilities and cross-trait genetic correlations between the 2 largest ancestry-specific meta-analyses: EA and East Asian-ancestry (EAS). Despite similarities, we measured significant heterogeneity for blood traits between these 2 populations. Because natural selection could account for such heterogeneity, we explored overlaps between ancestry-specific blood trait loci and selective sweeps identified in populations from the 1000 Genomes Project. We found significant overlaps for white blood cells (EA, EAS, African (AFR)), monocytes (EA, AFR), eosinophils (EA), neutrophils (AFR), lymphocytes (LYM) (EAS), and platelets (PLT)(EA, EAS). These loci included several genes known to be under positive selection (e.g. *DARC*, *SH2B3*), but also new candidates such as *IL6* strongly associated with PLT in EAS. Finally, we identified in South Asians a missense variant in *IL7* (rs201412253, MAF=2.6%, <0.4% in other populations) strongly associated with increased LYM. We showed in vitro that this variant increases IL7 secretion (+83%, P=2.7x10^{-5}).

In conclusion, our results for hematological traits highlight the potential genetic, clinical, and biological value of a more global representation of populations in genetic studies.

PgmNr 340: Large scale GWAS identifies clinically relevant rare variation for blood cell traits.

Authors:
L.M. Raffield [1]; D. Vuckovic [2,3]; E.L. Bao [4,5,6]; C.A. Lareau [4,5,6]; P. Akbari [7]; M. Chen [8]; T. Jiang [7]; A. Mousas [9]; A. Reiner [10]; W.J. Astle [11,12,13]; A.S. Butterworth [14,15]; A.D. Johnson [8]; P. Auer [16]; G. Lettre [9]; V.G. Sankaran [4,5,6]; N. Soranzo [3]; Blood Cell Traits Consortium (BCX)

View Session Add to Schedule

Affiliations:
1) Department of Genetics, University of North Carolina, Chapel Hill, NC, USA; 2) Department of Haematology, University of Cambridge, Cambridge, UK; 3) Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK; 4) Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA; 5) Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA; 6) Broad Institute of MIT and Harvard, Cambridge, MA, USA; 7) Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, UK; 8) Population Sciences Branch, National Heart, Lung and Blood Institute's The Framingham Heart Study, Framingham, MA, USA; 9) Department of Medicine, Montreal Heart Institute and Université de Montréal, Montréal, Quebec, Canada; 10) Department of Epidemiology, University of Washington, Seattle, WA, USA; 11) Department of Haematology, Cambridge Biomedical Campus, University of Cambridge, Cambridge, UK; 12) National Health Service Blood and Transplant, Cambridge Biomedical Campus, University of Cambridge, Cambridge, UK; 13) Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK; 14) NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 15) MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 16) Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Red blood cells, white blood cells, and platelets are essential for oxygen transport, immune response, and thrombosis respectively. These hematological indices are highly heritable and associated with risk of diseases. Here, we present results from a collaborative GWAS of 29 blood cell phenotypes in 563,085 European ancestry participants. This effort more than doubles the GWAS sample size for these traits, and approximately quadruples the number of associations identified in Europeans, with 10,723 conditionally distinct associations. This includes 510 rare (<1% minor allele frequency (MAF)) and 820 low-frequency variants (1-5% MAF), including known or suspected pathogenic variants from studies of Mendelian diseases. Integration of clinical annotations from ClinVar and HGMD demonstrated a ~2.4x larger effect size for high confidence pathogenic variants compared to benign. This difference was used to re-assign pathogenicity for variants with uncertain annotations which have effect sizes in the high confidence pathogenic range (e.g., *TUBB1* missense variant rs139473150, ~1 standard deviation higher platelet distribution width per minor allele). Rare variants were included in a phenome-wide association study (PheWAS) for 529 binary phenotypes from the UK Biobank. We identified 95 significant variant-phenotype associations for 24 unique variants (Bonferroni corrected $p < 9.5x10^{-5}$). Biologically plausible associations with novel hematological rare variants include lymphocyte variants with autoimmune conditions such as type 1 diabetes (e.g., rs115730672), red cell trait associated *PIEZO1* missense variant rs202127176 with varicose veins, and an *APOA5* 5’ UTR variant with hypercholesterolemia and mean corpuscular hemoglobin concentration (rs45611741). Next, we utilized SpliceAI, a state-of-the-art neural net classifier, to identify nine fine-mapped variants with strongly predicted splice-altering consequences, such as mean platelet volume associated *PEAR1* variant rs149254521. Finally, we integrated GWAS variants with ATAC-seq from 18 human cell types to identify novel cell type enrichments for the regulation of blood traits. We also found significantly higher chromatin accessibility in progenitor versus committed cell types for pleiotropic variants. These results show the power of large-scale blood cell trait GWAS to interrogate clinically meaningful variants across the allele frequency spectrum and to integrate Mendelian and complex trait genetics.

A- A+

PgmNr 341: Age, sex, and genetics influence the abundance of infiltrating immune cells in human tissues.

Authors:
A.R. Marderstein [1,2,3,4]; M. Uppal [2,3]; A. Verma [1,2,3]; B. Bhinder [2,3]; J. Mezey [1,2,4]; A.G. Clark [1,4]; O. Elemento [1,2,3]

View Session Add to Schedule

Affiliations:
1) Tri-Institutional Program in Computational Biology & Medicine, Weill Cornell Medicine, New York, NY, USA; 2) Institute of Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA; 3) Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA; 4) Department of Computational Biology, Cornell University, Ithaca, NY, USA

Introduction: Human immune systems vary dramatically across individuals, yet the environmental and genetic determinants of this variability remain poorly characterized. Many studies have focused on studying immune cell composition in peripheral blood, yet normal tissues and organs also consist of infiltrating immune cells. Despite a crucial role in disease biology such as cancer progression and autoimmune disease, the variability in tissue infiltration across individuals and between tissues has not been documented, and the mechanisms enabling such variation in baseline infiltration have not been elucidated.

Methods: Using bulk RNA-seq data from 53 distinct GTEx tissue types, we applied cell-type deconvolution algorithms to estimate immune content. We performed unsupervised clustering to evaluate heterogeneity in immune cell composition between samples of a single tissue type and between different tissues. We correlated age, sex, and genetic polymorphisms with specific infiltration patterns by utilizing a novel analytical framework that leverages information across deconvolution methods, allowing incorporation of multiple algorithms simultaneously rather than reliance on a single one. Methods were validated with in silico admixtures.

Results: We observed tissue-specificity of immune-rich infiltration patterns, with a mode of 1 immune-rich tissue per individual. We found 21 of 73 infiltration-related phenotypes to be associated with either age or sex ($FDR < 0.1$), including a significant increase of CD8+ T cell content in female breast tissue compared to male ($P = 5.4 \times 10^{-34}$). Through our genetic analysis, we discovered 13 infiltration-related phenotypes have genome-wide significant SNP associations (iQTLs) ($P < 5.0 \times 10^{-8}$), with a significant enrichment of tissue-specific expression quantitative trait loci (eQTLs) in suggested iQTLs ($P < 10^{-5}$). We highlight an association between neutrophil content in lung tissue and a regulatory variant near the *CUX1* transcription factor gene ($P = 9.7 \times 10^{-11}$), which is a tumor suppressor previously linked to neutrophil infiltration and the regulation of several immune response genes.

Conclusion: We describe the first evaluation of immune infiltration across healthy tissues in the human body. Together, our results identify key factors influencing inter-individual variability of specific tissue infiltration patterns, which could provide insights on therapeutic targets for shifting infiltration profiles to a more favorable one.

A- A+

PgmNr 342: Multimodal single-cell analysis of 70,000 human memory T cells characterizes genetic associations with immune cell states and gene expression in a Peruvian tuberculosis progression cohort.

Authors:
A. Nathan [1,2,3,4,5]; J.I. Beynor [2,3,4,5]; S. Suliman [3]; Y. Baglaenko [2,3,4,5]; K. Ishigaki [2,3,4,5]; Y. Luo [2,3,4,5]; I. Van Rhijn [3]; M.B. Murray [6]; D.B. Moody [3]; S. Raychaudhuri [1,2,3,4,5]

View Session Add to Schedule

Affiliations:
1) Department of Biomedical Informatics, Harvard Medical School, Boston, MA; 2) Center for Data Sciences, Brigham and Women's Hospital, Boston, MA; 3) Division of Rheumatology, Immunology and Allergy, Department of Medicine, Brigham and Women's Hospital, Boston, MA; 4) Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA; 5) Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA; 6) Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA

Genetic variants influence disease-associated T cell states and their gene expression. For example, memory T cell phenotypes have been associated with progression of latent *Mycobacterium tuberculosis* infection to active tuberculosis (TB), but host genetic factors driving this progression are unknown. To understand the effect of genetic variants on T cell states, we take a single-cell approach to 1) robustly define cell states, and 2) measure the effects of variants on cell state abundance and gene expression.

We are conducting a large multimodal single-cell sequencing study of 259 genotyped donors (132 cases who progressed to active TB and 127 household contacts with latent TB). We are using CITE-seq to simultaneously measure single-cell expression of over 30,000 genes and 31 surface markers to phenotype more than 300,000 memory T cells isolated from peripheral blood.

Here we analyzed pilot data from over 70,000 cells from 47 donors—the largest single-cell T cell data set with both RNA and surface markers to our knowledge. We used an integrative approach to jointly define 16 T cell states based on single-cell gene and protein expression. Cells segregated along major axes of variation defined by functional pathways (like cytokine signaling), and resolving some states required protein data (e.g., Th17 and gamma delta T cells). We identified novel states with unknown functions, such as a cytotoxic CD26+CCR5+CD4+ population. Using MASC (Fonseka et al Sci Trans Med 2018) we observed that this state decreased with age (OR = 0.69, 95% CI: 0.58-0.83, $p = 0.002$). An activated state was weakly associated with case status ($p = 0.007$) after adjusting for age and sex. In an initial analysis of these 70,000 memory T cells, 915 genes are regulated by a significant eQTL in at least one single-cell cluster ($FDR < 0.05$), and ~10% of these were detected in 6 or fewer clusters. We are now jointly modeling all clusters in interaction analyses to confirm that these are state-specific eQTLs.

Our findings underscore the importance of single-cell association studies to connect cell states, genetic variation, and gene expression. With multimodal single-cell data, we defined memory T cell states and cis-eQTLs from just a fraction of our growing data set. Our full data set of over 300,000 cells will enable precise characterization of T cell state-specific eQTLs and variants associated with state abundance—important steps toward understanding TB and other immune-mediated diseases.

A- A+

PgmNr 343: Monogenic and polygenic inheritance become instruments for clonal selection.

Authors:
P. Loh [1,2]; G. Genovese [2,3,4]; S.A. McCarroll [2,3,4]

View Session Add to Schedule

Affiliations:
1) Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; 3) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA; 4) Department of Genetics, Harvard Medical School, Boston, MA

Clonally expanded blood cells with somatic mutations (clonal hematopoiesis, CH) are commonly acquired with age and increase blood cancer risk. To identify genes and mutations that give selective advantage to mutant clones, we identified among 482,789 UK Biobank participants some 19,632 autosomal mosaic chromosomal alterations (mCAs), including deletions, duplications, and copy number-neutral loss of heterozygosity (CNN-LOH). Analysis of these acquired mutations along with inherited genetic variation, revealed 52 inherited, rare, large-effect coding or splice variants (in seven genes) that greatly (OR 11–758) increased vulnerability to CH with specific acquired CNN-LOH mutations. Acquired mutations systematically replaced the inherited risk alleles (at *MPL*) or duplicated them to the homologous chromosome (at *FH*, *NBN*, *MRE11*, *ATM*, *SH2B3*, and *TM2D3*). Three of the seven genes (*MRE11*, *NBN*, *ATM*) encode components of the MRN-ATM pathway, which limits cell division after DNA damage and telomere attrition; another two (*MPL*, *SH2B3*) encode proteins that regulate stem cell self-renewal. In addition to these monogenic inherited forms of CH, we found a common and surprisingly polygenic form: CNN-LOH mutations across the genome tended to cause chromosomal segments with alleles that promote hematopoietic cell proliferation to replace their homologous (allelic) counterparts. This dynamic reveals a challenge for lifelong cytopoiesis in any genetically diverse species: individuals inherit unequal proliferative genetic potentials on paternally and maternally derived chromosome-pairs, and readily-acquired mutations that replace chromosomal segments with their homologous counterparts give selective advantage to mutant cells.

To estimate the fraction of CNN-LOH clones attributable to protein-altering variants at risk loci, we examined exome sequence data available for 49,960 participants. Collectively, *MPL*-altering CH risk variants or ultra-rare (MAF<0.0001) variants were present in 39 of 71 individuals with 1p CNN-LOH events spanning *MPL* (vs. 0.5 expected), indicating that ~54% of acquired 1p CNN-LOH events are driven by inherited coding or splice variants at *MPL*. Similarly, inherited variants at *ATM*, *NBN*, *SH2B3*, and *TM2D3* appeared to drive ~17–33% of CNN-LOH events spanning these loci. Additionally, at the frequently-mutated *DNMT3A*, *TET2*, and *JAK2* loci, ~24–60% of CNN-LOH mutations appeared to provide second hits to somatic point mutations detectable from exome sequencing reads.

A- A+

PgmNr 344: Inherited causes and clinical consequences of clonal hematopoiesis from 100,002 whole genomes.

Authors:
J. Weinstock [1]; A. Bick [2]; S. Nandakumar [3]; V. Sankaran [3]; A. Reiner [4]; S. Jaiswal [5]; G. Abecasis [1]; P. Natarajan [6]; S. Kathiresan [2]; on behalf of the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

View Session Add to Schedule

Affiliations:
1) Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI; 2) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 3) Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA; 4) Fred Hutchinson Cancer Research Center, Seattle, WA; 5) Department of Pathology, Stanford University, Stanford, CA; 6) Division of Cardiology, Massachusetts General Hospital, Boston, MA

INTRODUCTION: Clonal Hematopoiesis of Indeterminate Potential (CHIP) is a clonal expansion of blood cells arising from a leukemogenic somatic mutation in hematopoietic stem cells. CHIP increases with age and has been associated with hematologic malignancy and coronary artery disease.

OBJECTIVES: Simultaneous somatic and germline whole genome sequence analysis now provides the opportunity to identify root causes of CHIP. Here, we analyze genomes from 100,002 participants of diverse ancestries in the NHLBI TOPMed program to identify inherited variation associated with CHIP and its clinical consequences.

METHODS: We identified CHIP through somatic variant calling with MuTect2 followed by filtering on known leukemogenic CHIP driver mutations. We analyzed associations between CHIP and clinical phenotypes including blood traits, inflammatory markers and stroke. We performed a genome-wide association study to identify common inherited germline variation associated with the development of CHIP. We applied burden tests to rare loss-of-function (LOF) variation in coding genes, and rare regulatory variation in hematopoietic stem cell enhancers.

RESULTS: We identified 4,587 individuals with CHIP, 90% of whom had a single driver mutation. >75% of the driver mutations were in one of three CHIP driver genes (DNMT3A, TET2, and ASXL1). CHIP prevalence was strongly associated with age ($p<10^{-300}$).

CHIP associated with blood cell trait RDW ($p=1.3 \times 10^{-4}$), and inflammatory markers IL-6 ($p=4.6 \times 10^{?5}$) and Lp-PLA2 ($p=1.8 \times 10^{?7}$). We found an increased risk for the first nononcologic incident event conferred by CHIP – ischemic stroke (Adjusted HR: 1.1, p=0.047), with larger CHIP clones (VAF>20%) conferring increased risk (HR: 1.2, p=0.008). CHIP driver gene specific phenotypic profiles were also observed.

Four genome-wide significant risk loci were associated with CHIP, including one locus at *TET2* that was African ancestry specific. Computational analyses of the TET2 locus implicated rs79901204 as the causal variant which disrupts a hematopoietic stem cell specific enhancer element. Rare LOF variants in *CHEK2*, a DNA damage repair gene, and rare regulatory variation in HAPLN1 enhancers, a bone marrow stromal cell protein, were the top associations with CHIP in rare variant analyses ($p=2.1 \times 10^{-5}$ and $p=2.0 \times 10^{-5}$ respectively).

CONCLUSION: Heritable variation altering hematopoietic stem cell function and the fidelity of DNA-damage repair increased the likelihood of developing CHIP.

A- A+

PgmNr 345: Heterozygous deleterious variants in *WBP11* cause multiple congenital anomalies in humans and mice.

Authors:
G. Chapman [1,2]; E.M.M.A. Martin [1]; A. Enriquez [1,2]; D.B. Sparrow [3]; D.T. Humphreys [1,2]; K.R. Iyer [1]; J.A. Greasby [1]; P. Leo [4]; E.L. Duncan [4]; C. Dimartino [5,6]; J. Amiel [5,6,7]; N.L.M. Sobreira [8]; D. Lehalle [9]; H.M. Rasouly [10]; A.G. Gharavi [10]; R.D. Steiner [11]; C. Raggio [12]; R. Blank [13]; C.T. Gordon [5,6]; R. Jobling [14]; P. Giampietro [15]; S.L. Dunwoodie [1,2,16]

View Session Add to Schedule

Affiliations:
1) Development and Stem cell biology division, Victor Chang Cardiac Research Institute, Sydney, NSW, 2010, Australia; 2) Faculty of Medicine, UNSW Sydney, NSW, 2052, Australia; 3) Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom; 4) Queensland University of Technology, Brisbane, QLD, 4000, Australia; 5) Laboratory of Embryology and Genetics of Human Malformations, Institut National de la Santé et de la Recherche Médicale (INSERM) UMR 1163, Institut Imagine, Paris, France; 6) Paris Descartes-Sorbonne Paris Cité Université, Institut Imagine, Paris, France; 7) Département de Génétique, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris (AP-HP), Paris, France; 8) Pediatric Clinical Research Unit, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, 21287, USA; 9) Centre Hospitalier Intercommunal Créteil, Créteil, 94000, France; 10) Division of Nephrology, Department of Medicine Columbia University, 1150 Saint Nicholas Avenue, New York 10032; 11) Department of Pediatrics, University of Wisconsin School of Medicine and Public Health; 12) Hospital for Special Surgery, Pediatrics Orthopedic Surgery, New York, NY 10021, USA; 13) Medical College of Wisconsin, Medicine, Milwaukee, WI 53226, USA; 14) The Hospital for Sick Children, Toronto, Ontario, Canada; 15) Robert Wood Johnson School of Medicine/Rutgers University Department of Pediatrics, New Brunswick, NJ 08901; 16) Faculty of Science, UNSW Sydney, NSW, 2052, Australia

---

Vertebral malformations are relatively common, affecting 1 in 2,000 live births and can result in abnormal curvature of the spine or, in severe cases, in neonatal death due to restrictive respiratory insufficiency. Although vertebral malformations often occur in isolation, 20% of cases have additional defects, most commonly congenital heart disease. Relatively few genes have been linked to vertebral malformations, mostly associated with rare and severe forms, a notable exception being the discovery that *TBX6* deletion together with a common haplotype cause 10% of congenital scoliosis cases.

In order to identify additional genetic causes of vertebral malformations we performed exome sequencing on the DNA of 45 trios and families. Rare predicted pathogenic variants that segregated with disease were selected. Several likely causative variants were identified in genes previously linked to vertebral defects in humans. In addition, two unrelated patients with Klippel-Feil Syndrome had novel heterozygous stop gain variants in the *WBP11* gene. *WBP11* encodes an evolutionarily conserved activator of splicing, which has not been previously linked to human disease. Similar truncated *WBP11* constructs lack splicing activity (Llorian *et al.* 2005 *J. Biol. Chem.* 280: 38862-9), suggestive of a haploinsufficient mechanism of disease. Eight additional patients in five families were identified via GeneMatcher, four families with heterozygous truncating *WBP11* variants and one with a predicted pathogenic variant in *WBP11*. Although patient phenotypes varied, common malformations included cervical vertebral fusions, oesophageal atresia, kidney and heart defects. We generated a mouse model of *WBP11* haploinsufficiency by creating an 8bp deletion in exon 5 using CRISPR-Cas9 targeting. *Wbp11* homozygous null mouse embryos died prior to E8.5 indicating that *Wbp11* is an essential gene for development. Mice heterozygous for a *Wbp11* null allele were not found in the expected Mendelian ratio with many dying either late in gestation, postnatally or as adults from hydrocephaly. Importantly, defects of the axial skeleton, brain and kidneys were common in *Wbp11* heterozygous mice, similar to our patients, confirming *WBP11* mutation as a cause of multiple congenital defects in humans and mice.

A- A+

PgmNr 346: Understanding the physiological role of *DDRGK1* from Shohat-type SEMD: Novel regulatory mechanism of matrix proteins.

Authors:
Y. Bae; M. Weisz-Hubshman; A. Egunsola; M. Jiang; Z. Yu; Y. Chen; B. Lee

View Session Add to Schedule

Affiliation: Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

DDRGK1 is a substrate of ufmylation process, a post-translational modification mediated by UFM1, one of ubiquitin-like proteins. Various cellular and molecular functions of DDRGK1 were demonstrated by others' *in vitro* and genetic mouse studies. In our previous study, we elucidated the important physiological role of DDRGK1 by identifying homozygous mutation of *DDRGK1* as a causative mutation of Shohat-type SEMD (spondyloepimetaphyseal dysplasia), a rare form of chondrodysplasia. Zebrafish study showed that *DDRGK1* loss-of-function mutation led to defects in cartilage development. Furthermore, *Ddrgk1*$^{-/-}$ mice were embryonic lethal (E11.5) and, also resulted in delayed chondrogenic mesenchymal condensation in the limb buds. Mechanistically, we showed that DDRGK1 forms a complex with SOX9 and inhibits ubiquitination-mediated proteasomal degradation of SOX9. We also found a decreased Sox9 protein level and *Col2a1* transcript in the absence of *Ddrgk1* in ATDC5 cells. Hence, our discovery of DDRGK1 as a causative gene in Shohat-type SEMD provides that DDRGK1 is required for maintenance of SOX9 protein stability during chondrogenesis. Here, to elucidate cartilage-specific roles of DDRGK1, we generated *Prx1-Cre; Ddrgk1*$^{f/f}$ mice, where *Ddrgk1* is deleted in the osteochondroprogenitors of limb buds. Unlike *Ddrgk1*$^{-/-}$ mice, *Prx1-Cre; Ddrgk1*$^{f/f}$ mice survive and manifest severe chondrodysplasia including progressive shortening of the limbs with an abnormal growth plate and a delayed epiphyseal ossification. The proliferating zone in *Prx1-Cre; Ddrgk1*$^{f/f}$ mice was shortened, while the hypertrophic zone was elongated. Furthermore, *Agn1-CreERT2; Ddrgk1*$^{f/f}$ mice were generated to understand the role of DDRGK1 in adult cartilage homeostasis, where *DDRGK1* is deleted in chondrocytes from articular cartilage and growth plate by tamoxifen inducible manner. We found disorganized growth plate and reduced level of proteoglycan in growth plate and articular cartilage by Safranin O staining. Hence, DDRGK1 plays critical roles of: 1) regulating chondrogenic mesenchymal condensation by regulating SOX9 protein stability and proper development of growth plate, and 2) maintaining the matrix proteins in adult cartilage. The investigation of molecular functions of DDRGK1 is currently underway, and this will provide novel mechanism of UFM1 modification in cartilage development and common age-related diseases.

A- A+

PgmNr 347: The *Maenli* long non-coding RNA locus controls *En1* expression and limb development.

Authors:
A. Superti-Furga [1]; L. Allou [2]; S. Balzano [3]; A. Magg [2,4,5]; M. Quinodoz [1,3]; B. Royer-Bertrand [1,3]; R. Schöpflin [2,4,5,6]; W.L. Chan [4,5]; E. Speck-Martins [7]; D. Rocha de Carvalho [7]; L. Farage [8]; C. Marques Lourenço [9]; S. Rajagopal [10]; S. Nampoothiri [11]; B. Campos-Xavier [1]; C. Chiesa [1]; F. Niel-Bütschi [1]; L. Wittler [12]; B. Timmermann [13]; S. Unger [1]; C. Rivolta [3,1]; P. Grote [14]; S. Mundlos [2,4,5]

View Session Add to Schedule

Affiliations:
1) Div. of Genetic Medicine, Lausanne Univ. Hospital (CHUV) and Uni Lausanne, Lausanne, Switzerland; 2) RG Development & Disease, Max Planck Inst. for Molecular Genetics, Berlin, Germany; 3) Medical Genetics Unit, Dept. of Computational Biology, Uni Lausanne, Lausanne, Switzerland;; 4) Inst. for Medical and Human Genetics, Charité-Universitätsmedizin Berlin, Berlin, Germany; 5) Berlin-Brandenburg Ctr for Regenerative Therapies, Charité-Universitätsmedizin Berlin, Berlin, Germany; 6) Dept. of Computational Molecular Biology, Max Planck Inst. for Molecular Genetics, Berlin, Germany; 7) Genetic Unit, SARAH Network of Rehabilitation Hospitals, Brasilia, Brazil; 8) Inst. de Cardiologia do Distrito Federal, Brasilia, Brazil; 9) Faculdade de Medicina, Centro Universitario Estácio de Ribeirão Preto, Ribeirânia, Ribeirão Preto – SP, Brazil; 10) Dept. of Medical Genetics, Tamil Nadu Dr. M.G.R. Medical University, Chennai, India; 11) Dept. of Pediatric Genetics, Amrita Institute, Cochin, India;; 12) Dept. of Developmental Genetics, Max Planck Inst. for Molecular Genetics, Berlin, Germany; 13) Max Planck Inst. for Molecular Genetics, Sequencing Core Facility, Berlin, Germany; 14) Inst. of Cardiovascular Regeneration, Ctr. for Molecular Medicine, Goethe University, 60439 Frankfurt am Main, Germany

---

The homeobox gene *Engrailed-1* (*EN1*) is conserved from drosophila to mouse and man. In drosophila, it controls segmentation and wing formation; in the mouse, it drives development of the hindbrain as well as outgrowth and dorso-ventral differentiation of limb buds. We report here that homozygous deletions approx. 200 kb downstream of *EN1* result in severe limb malformation with mesomelic shortening of the legs and dorsalisation of ventral structures including ventral nails. We identified a 4-exon non-coding lncRNA in the deleted region that specifically regulates *EN1* expression in the limb ("master of EN1 in the limbs"; *Maenli*). To investigate *Maenli* function, an allelic series was engineered in mouse ES cells and used to generate homozygous mice. Deletion of exon 1, insertion of a poly(A) stop-cassette, or inversion of the *Maenli* promoter resulted in loss of *EN1* expression in the limb and consecutive dorsal transformation of ventral paw structures similar to the previously described inactivation of *EN1*, but normal brain. Concomitant monoallelic deletion of *Maenli* and *EN1* in double heterozygous mice did not rescue the limb phenotype, indicating that activation of *EN1* by *Maenli* requires the two elements to be in *cis*. Thus, the lncRNA *Maenli* drives the expression of En1 in the limbs, and its biallelic abrogation in the human results in a phenotype of limb shortening with impaired dorsoventral differentiation (En1-related dorsoventral syndrome, ENDOVES) with intact brain structure. In contrast, with biallelic inactivation of *EN1* itself results in a combined limb-brain phenotype with absence of the cerebellum, as we have observed in a child homozygous for a premature termination variant in the *EN1* coding region.

Limb-specific *cis*-activation of *En1* by *Maenli* and the loss of function-associated phenotype show that genetic variants in lncRNA can result in congenital malformations. The two distinct phenotypes associated with *En1* and *Maenli* illustrate a novel conceptual frame of a full phenotype from a protein-coding gene and of a “nested” phenotype from a regulatory locus on that gene, both being inherited as true recessives. This suggests that involvement of ncRNAs in human Mendelian diseases could be more prevalent than currently appreciated.

A- A+

PgmNr 348: Variants in the HIF-1 pathway are associated with Ollier disease and Maffucci syndrome.

Authors:
S. Robbins [1,2]; R. Martin [1]; B. Cohen [3]; C. Haldeman-Englert [4]; M. Cernach [5]; G. Gottesman [6]; G. Semenza [7]; D. Valle [1]; N. Sobreira [1]

View Session Add to Schedule

Affiliations:
1) McKusick-Nathans Institute in the Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 2) Predoctoral Training Program in Human Genetics, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 3) Division of Pediatric Dermatology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 4) Fullerton Genetics Center Mission Health, Asheville, NC, USA; 5) Escola Paulista de Medicina, Sao Paolo, Brazil; 6) Center for Metabolic Bone Disease and Molecular Research, Shriners Hospitals for Children - St. Louis, St. Louis, Missouri, USA; 7) Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

Ollier disease (OD, OMIM 166000) and Maffucci Syndrome (MS, OMIM 614569) are two related rare disorders characterized by multiple enchondromas, which commonly develop in the extremities around joints, resulting in deformity, leg length discrepancy, and pathological fractures. MS is distinguished from OD by the development of vascular anomalies. Both disorders are cancer predisposition syndromes; chondrosarcomas develop in ~30% of individuals. Gain-of-function variants in *IDH1* and *IDH2* have been described in ~80% of the enchondromas, chondrosarcomas and vascular anomalies in patients with OD and MS. However, no predisposing germline variants have been reported to date. We hypothesize that OD and MS are tumor predisposition syndromes with germline or early post-zygotic causative variants and additional tumor variants involved in the formation of the benign and malignant tumors. Locus heterogeneity for the germline variants also seems likely. To search for causative germline variants, we performed exome sequencing of leukocyte DNA in 27 probands. We show that 9 (33%) had one or more rare variants in one of 5 genes in the HIF-1 pathway (*IDH2*, *KDM4C, HIF1A, VHL,* and *EGLN1* in 3, 1, 1, 4, and 2 probands respectively). To further investigate the role of HIF-1 in the pathogenesis of OD and MS, we performed RNA-seq of fibroblast RNA from 3 individuals with variants in *IDH2, KDM4C, VHL* with OD and MS at normoxia and hypoxia and compared to 3 control fibroblast lines. We show that patient fibroblasts have significantly less differentially expressed HIF-1-regulated genes in response to hypoxia suggesting that the HIF-1 pathway variants in our patients lead to HIF-1 pathway dysregulation. Also, to further investigate the pathways dysregulated in patients with OD and MS, we performed RNA-seq of 4 enchondromas and 3 chondrosarcomas in comparison to a human control chondrocyte line. The genes in the KEGG pathway “proteoglycans in cancer” (KEGG:05205) were enriched in both the normoxia fibroblasts and chondrosarcomas differentially expressed gene sets; this pathway encompasses many signaling cascades, e.g. HIF-1, mTOR, and MAPK. Also, chondrocyte differentiation and cartilage development genes (*HOXA11, GDF6,* and *PKDCC*) were significantly differentially expressed in patients’ fibroblasts, enchondromas and chondrosarcomas. Based on our results we suggest that dysregulation of HIF-1 and related pathways are responsible for tumorigenesis in patients with OD and MS.

A- A+

PgmNr 349: Mutations in *ANAPC1*, encoding a scaffold subunit of the anaphase promoting complex, cause Rothmund-Thomson syndrome type 1.

Authors:
P. Campeau [1]; N.F. Ajeawung [1]; T.T.M. Nguyen [1]; L. Lu [2]; T.J. Kucharski [3]; J. Rousseau [1]; S. Molidperee [1]; J. Atienza [1]; I. Gamache [1]; W. Jin [2]; S.E. Plon [2]; B.H. Lee [2]; J.G. Teodoro [3]; L.L. Wang [2]

View Session Add to Schedule

Affiliations:
1) University of Montreal, Montreal, Quebec, Canada; 2) Baylor College of Medicine; 3) McGill University

Rothmund-Thomson Syndrome (RTS) is an autosomal recessive disorder characterized by poikiloderma, sparse hair, short stature and skeletal anomalies. Type 2 RTS, which is defined by presence of bi-allelic mutations in *RECQL4*, is characterized by increased cancer susceptibility and skeletal anomalies, while the genetic basis of RTS Type 1, which is associated with juvenile cataracts, is unknown. We studied 10 individuals with RTS Type 1 from seven families and identified a deep intronic splicing mutation of the *ANAPC1* gene, a component of anaphase promoting complex/cyclosome (APC/C), in all affected individuals, either in the homozygous state or in *trans* with another mutation. Fibroblast studies showed that the intronic mutation causes the activation of a 95 bp pseudoexon leading to mRNAs with premature termination codons and nonsense-mediated decay, decreased ANAPC1 protein levels, and prolongation of interphase. Interestingly, mice heterozygous for a knockout mutation have an increased incidence of cataracts. Our results demonstrate deficiency in APC/C as a cause for RTS Type 1 and suggest a possible link between the APC/C and RECQL4 helicase, since both proteins are involved in DNA repair and replication.

A- A+

PgmNr 350: Mutations in ephrinB2-EphB4-RASA1 signaling underlie Vein of Galen Malformation.

Authors:
Zeng [1,2]; A. Hunt [3]; S.C. Jin [2]; S. Conine [3]; A. Allocco [3]; D. Duran [3,20]; C. Nelson-Williams [1]; M. Sorscher [4]; E. Loring [1]; J. Klein [5]; M. DiLuna [3]; C. Matouk [3]; S. Alper [6]; M. Komiyama [7]; A. Ducruet [8]; J. Zabramski [8]; A. Dardik [9]; B. Walcott [10]; C. Stapleton [11]; B. Aagaard-Kienitz [12]; G. Rodesch [13]; E. Jackson [14]; E. Smith [15]; D. Orbach [5,15]; A. Berenstein [4]; K. Bilguvar [1,16]; M. Vikkula [17]; M. Gunel [1,3]; R.P. Lifton [1,2]; K.T. Kahle [3,18,19]

View Session Add to Schedule

Affiliations:
1) Department of Genetics, Yale School of Medicine, New Haven, CT, USA; 2) Laboratory of Human Genetics and Genomics, The Rockefeller University, New York, NY, USA; 3) Department of Neurosurgery, Yale School of Medicine, New Haven, CT, USA; 4) Department of Neurosurgery, Icahn School of Medicine at Mount Sinai, New York NY, USA.; 5) Department of Neurosurgery, Boston Children’s Hospital, Boston MA, USA.; 6) ivision of Nephrology and Center for Vascular Biology Research, Beth Israel Deaconess Medical Center; and Department of Medicine, Harvard Medical School, Boston, MA USA; 7) Department of Neurointervention, Osaka City General Hospital, Osaka, Japan; 8) Department of Neurosurgery, Barrow Neurological Institute, Phoenix AZ, USA; 9) Department of Surgery, Yale School of Medicine, New Haven CT, USA; 10) Department of Neurological Surgery, University of Southern California, Los Angeles, CA, USA; 11) Department of Neurological Surgery, Massachusetts General Hospital and Harvard Medical School, Boston, MA USA; 12) Department of Neurological Surgery, University of Wisconsin, Madison, Wisconsin; 13) Service de Neuroradiologie Diagnostique et Thérapeutique, Hôpital Foch, Suresnes, France; 14) Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD USA; 15) Department of Neurointerventional Radiology, Boston Children’s Hospital, Boston MA, USA; 16) Yale Center for Genome Analysis, West Haven CT, USA; 17) Human Molecular Genetics, de Duve Institute, Université catholique de Louvain, Brussels, Belgium; 18) Department of Pediatrics, Yale School of Medicine, New Haven CT, USA.; 19) Department of Cellular and Molecular Physiology, Yale School of Medicine, New Haven, CT, USA; 20) Department of Neurosurgery, University of Mississippi Medical Center, Jackson MS, USA.

Vein of Galen Malformation (VOGM), characterized by the direct connection between choroidal circulation and the median prosencephalic vein in the brain without an intervening capillary bed, is the most common (1 : 25000) and severe type of neonatal brain arteriovenous malformations. Given the sporadic nature of the disease, VOGM was thought to be mainly attributable to *de novo* mutations. However, recent genetic studies using whole exome sequencing (WES) identified genes with significant burden of rare transmitted variants as well as somatic second-hit in risk genes. The large spectrum of disease manifestations observed in mutation carriers, ranging from mild cutaneous vascular lesions to VOGM, suggested variable expressivity and incomplete penetrance of VOGM genes.

In a previous study of 55 VOGM patients, including 52 parent-proband trios, we identified significant burden of rare damaging transmitted mutations in the ephrin receptor signaling pathway, including *EPHB4*, which reached genome-wide significance. Subsequent recruitment efforts have increased our sequenced cohort to almost double the size with 238 samples, the largest VOGM cohort in the world.

In the analysis of this expanded cohort, we have identified novel mutations in genes we previously discovered, including 1 *de novo* and 1 transmitted stopgain mutation in *RASA1*, and 1 transmitted deleterious missense (D-mis, defined by MetaSVM 'D') mutation in *EPHB4.* Notably, we identified 2 loss of function mutations (defined as stopgain, frameshift, or splice site variants) that are closely clustered in *ITGB1*, a gene in the same experimentally supported interactome as *RASA1* and *EPHB4*. Both *ITGB1* mutations disrupt a highly conserved tyrosine motif (NPxY) in the cytoplasmic tail, which is essential for the binding of ITGB1 to downstream proteins and integrin signaling regulated by tyrosine phosphorylation. Additionally, a D-mis mutation in *ACVRL1*, a known regulator of the ephrinB2-EphB4-RASA1 signaling, was identified in a VOGM proband from a three-generation family with history of Hereditary Hemorrhagic Telangiectasia (HHT) and multiple cases of VOGM. This *ACVRL1* mutation cosegregates with the disease phenotype in the family, establishing *ACVRL1* as a *bona fide* VOGM risk gene. The identified damaging mutations in the ephrinB2-EphB4-RASA1 signaling axis collectively account for ~18% of the VOGM probands, highlighting the key functionality of this pathway in the etiology of VOGM.

A- A+

PgmNr 351: gnomAD-SV: An open resource of structural variation for medical and population genetics.

Authors:
R.L. Collins [1,2,3]; H. Brand [1,2,4]; K.J. Karczewski [1,2]; X. Zhao [1,2,4]; J. Alföldi [1,2]; A.V. Khera [1,2]; L.C. Francioli [1,2,5]; L.D. Gauthier [1,2,6]; H. Wang [1,2]; N.A. Watts [1,2]; M. Solomonson [1,2]; A. O'Donnell-Luria [1,2]; A. Baumann [6]; R. Munshi [6]; C. Lowther [1,2,4]; M. Walker [1,2,6]; C. Whelan [6,7]; E. Valkanas [1,2,3]; J. Fu [1,2]; A. Philippakis [6]; E. Lander [1,8,9]; S. Gabriel [1]; B.M. Neale [1,2,3,7]; S. Kathiresan [1,2,5,10]; M.J. Daly [1,2,3,7,11]; E. Banks [6]; D.G. MacArthur [1,2,3,5]; M.E. Talkowski [1,2,3,4,7]; The Genome Aggregation (gnomAD) Consortium

View Session Add to Schedule

Affiliations:
1) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA.; 2) Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA; 3) Division of Medical Sciences, Harvard Medical School, Boston, MA; 4) Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 5) Department of Medicine, Harvard Medical School, Boston, MA; 6) Data Science Platform, Broad Institute of Harvard and M.I.T., Cambridge, MA; 7) Stanley Center for Psychiatric Research, Broad Institute of Harvard and M.I.T., Cambridge, MA; 8) Department of Systems Biology, Harvard Medical School, Boston, MA; 9) Division of Health Sciences and Technology, M.I.T., Cambridge, MA; 10) Division of Cardiology, Massachusetts General Hospital, Boston, MA; 11) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

---

Structural variants (SVs) rearrange the linear and three-dimensional organization of the genome, which can have profound consequences in evolution, diversity, and disease. As national biobanks, disease association studies, and clinical genetic testing are increasingly reliant on whole-genome sequencing, population variation references have become integral for the evaluation and interpretation of genomic variation. Here, we constructed a reference atlas of SVs from 32X short-read whole-genome sequencing (WGS) of 14,891 individuals across diverse global populations (54% non-European) as a component of gnomAD. We discovered a rich landscape of 498,257 unique SVs, including 5,729 multi-breakpoint complex SVs across 13 mutational subclasses, and examples of localized chromosome shattering, like chromothripsis. SVs were non-uniformly distributed across the chromosomes and SV classes; likewise, mutation rate estimates varied substantially by SV class. Signatures of selection were strongest against inversions and complex SVs, which appeared to be attributable to both coding and noncoding effects. We discovered strong correlations between constraint against predicted loss-of-function (pLoF) SNVs and rare SVs that both disrupt and duplicate protein-coding genes, suggesting that existing per-gene metrics of pLoF SNV constraint do not simply reflect haploinsufficiency, but appear to capture a gene's general sensitivity to dosage alterations. Our SV pipelines detected 8,202 SVs per genome, including eight rare, gene-altering SVs, and we predicted that SVs constitute at least 25% of all rare loss-of-function events per genome. We observed large (≥1Mb), rare SVs in 3.1% of genomes (~1:32 individuals), and a clinically reportable pathogenic incidental finding from SVs in 0.24% of genomes (~1:417 individuals). We also estimated the prevalence of previously reported pathogenic recurrent CNVs associated with genomic disorders, which highlighted differences in frequencies across populations and confirmed that WGS-based analyses can readily recapitulate these clinically important variants. In total, gnomAD-SV includes at

least one CNV covering 57% of the genome, while the remaining 43% is significantly enriched for CNVs found in tumors and individuals with developmental disorders. The gnomAD-SV map is browsable online (https://gnomad.broadinstitute.org), which will allow broad, hands-on access to these results as a resource for medical and population genetics.

PgmNr 352: Recommendations for determining the clinical validity of functional studies for use in variant interpretation.

Authors:
S. Brnich [1]; A. Abou Tayoun [2]; M.S. Greenblatt [3]; C.D. Heinen [4]; D. Kanavy [1]; X. Luo [5]; S.M. McNulty [1]; L.M. Starita [6,7]; S.V. Tavtigian [8]; S.M. Harrison [9]; L.G. Biesecker [10]; J.S. Berg [1]; ClinGen Sequence Variant Interpretation Working Group

View Session Add to Schedule

Affiliations:
1) Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Al Jalila Children's Specialty Hospital, Dubai, UAE; 3) Department of Medicine and University of Vermont Cancer Center, Larner College of Medicine, University of Vermont, Burlington, VT; 4) Center for Molecular Oncology, UConn Health, Farmington, CT; 5) Department of Pediatrics-Oncology, Baylor College of Medicine, Houston, TX; 6) Department of Genome Sciences, University of Washington, Seattle, WA; 7) Brotman Baty Institute for Precision Medicine, Seattle, WA; 8) Department of Oncological Sciences and Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT; 9) Broad Institute of MIT and Harvard, Cambridge, MA; 10) Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) clinical variant interpretation guidelines permit the use of "well-established" functional assays at a "strong" level of evidence. However, they do not provide detailed guidance on how this evidence should be evaluated or what is required for study validation. Differences in functional evidence application in variant interpretation is a major contributor to discordance between clinical laboratories, preventing definitive benign or pathogenic variant classifications. This recommendation provides a structured approach for assessing functional assays in variant interpretation.

The Clinical Genome Resource (ClinGen) Sequence Variant Interpretation (SVI) Working Group held multiple in-person and virtual meetings to define what constitutes a well-established functional assay and how this evidence should be structured for curation. In this process, we considered expert opinions, feedback from the ClinGen Steering Committee, and tangible examples of functional evidence applied in variant interpretations published by ClinGen Variant Curation Expert Panels.

The SVI recommends a four-step process to determine the applicability and strength of evidence of functional assays for use in clinical variant interpretation. These steps are: 1. Define the disease mechanism in structured terms for gene, disease, molecular mechanism, and implicated biological pathway. 2. Evaluate general classes of assays and model systems used in the field for appropriateness to disease pathogenesis. 3. Evaluate validity of specific instances of assays as performed by different groups. Negative/positive controls and replicates are required for use in clinical variant interpretation. Benchmarking against known benign and pathogenic variants can increase the evidence strength, according to sensitivity/specificity or odds of pathogenicity. We recommend using "functionally normal" or "functionally abnormal" to describe a variant's functional impact as measured in a given assay. 4. Apply evidence to individual variant interpretation, using only the strongest evidence that best represents disease mechanism if multiple results are available.

This approach to functional evidence evaluation will help clarify ambiguity in the clinical variant interpretation process and help foster productive partnerships with basic scientists who have developed functional assays for various genes.

A- A+

PgmNr 353: Validation of scoring metrics to guide the classification of constitutional copy number variants.

Authors:
E. Riggs [1]; E. Andersen [2]; A. Cherry [3]; S. Kantarci [4]; H. Kearney [5]; A. Patel [6]; G. Raca [7]; D. Ritter [8]; S. South [9]; E. Thorland [5]; D. Pineda-Alvarez [10]; S. Aradhya [3,10]; C. Martin [1]

View Session Add to Schedule

Affiliations:
1) Autism & Developmental Medicine Institute, Geisinger , Lewisburg, Pennsylvania.; 2) ARUP Laboratories, University of Utah, Salt Lake City, Utah; 3) Department of Pathology, Stanford Health Care, Stanford, CA; 4) Cytogenetics and Genomics, Quest Diagnostics Nichols Institute, San Juan Capistrano, CA; 5) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; 6) Lineagen, Salt Lake City, UT; 7) Children's Hospital of Los Angeles, Los Angeles, CA; 8) Texas Children's Cancer Center, Baylor College of Medicine, Houston, TX; 9) AncestryDNA, Lehi, UT; 10) Invitae, San Francisco, CA

The American College of Genetics and Genomics (ACMG) and the NIH-funded Clinical Genome Resource (ClinGen) are in the process of updating the technical standards for classification and reporting of constitutional copy number variants (CNVs). This update will include points-based scoring metrics designed to guide users through a process for evaluating evidence and assigning classifications (e.g., pathogenic, uncertain, etc.) for both copy number losses and gains. These metrics were developed through an iterative process using expert opinion on the sources and relative strengths of various types of evidence. Through discussion and case examples, the committee assigned relative weights to each evidence type, including: the presence of known dosage-sensitive genes, overlap with CNVs reported in clinically affected individuals and individuals in the general population, case-control studies, segregation data, *de novo* occurrences, and the number of protein-coding genes included in the CNV.

The scoring metrics were refined through multiple rounds of internal and external testing. A total of 114 CNVs (58 deletions, 56 duplications), previously observed and reported by clinical laboratories, were evaluated by committee members and external reviewers using the scoring metrics. A subset of 47 of these CNVs (26 deletions, 21 duplications) were also evaluated using current classification methods as a baseline for comparison. The testing process aimed to answer three questions: 1) how often reviewers match the original clinical laboratory classification; 2) how often reviewers evaluating the same CNV reach the same classification (i.e., concordance); and 3) how appropriate were the classifications assigned using the scoring metrics, in the opinion of the reviewers. Overall, reviewers' ability to arrive at classifications concordant with the original clinical laboratory increased from 70.2% at baseline to 79.1% using the scoring metrics, and conflicting classifications that may impact medical management decreased from 39.2% to 23.4%. Classifications calculated using the metrics were considered appropriate by reviewers 89.3% of the time. With increased education, familiarity, and experience, we expect to see steady improvements in inter-laboratory concordance. We will continue to study trends in inter-laboratory concordance using these metrics, as well as usability and user experience, and plan to use this information to guide future improvements of the scoring metrics.

A- A+

PgmNr 354: StrVCTURE: A supervised learning method to predict the pathogenicity of structural variants.

Authors:
A.G. Sharo; S.E. Brenner

View Session Add to Schedule

Affiliation: University of California, Berkeley, CA 94720

We have developed a method to classify the impact of germline structural variants for Mendelian disease. Whole genome sequencing has shown success in clinical cases that thwart traditional diagnostic methods. However, preliminary studies show that at least half of these cases remain unresolved after whole genome sequencing. Structural variants (genomic variants larger than 50bp) may be the genetic cause of a portion of these unresolved cases. Historically, structural variants have been difficult to detect with confidence from short-read sequencing. As long-read and linked-read sequencing methods become increasingly available, researchers, and eventually clinicians, will have access to thousands of reliable structural variants of unknown disease-relevance. Filtering these structural variants by allele frequency is more difficult than filtering single nucleotide variants, as many structural variants will have endpoints not precisely matching those previously observed. This is problematic, as small changes in structural variant endpoints can vastly change their impact. Methods to predict the pathogenicity of these structural variants will be needed to realize the full diagnostic potential of long read sequencing.

To address this emerging need, we have developed StrVCTURE (Structural Variant Classification Through Understanding Regions with Exons), a classifier to distinguish pathogenic from benign structural variants that overlap exons. We make use of exon-specific, gene-specific, and conservation features in a random forest classifier that classifies pathogenic variants. Although existing databases of structural variants reflect size biases from sequencing techniques, we leverage multiple databases to construct a matched training set of pathogenic and rare, putatively benign structural variants. However, our method and its assessment are still constrained by acquisition bias in databases of pathogenic variants, and noise due to imprecise endpoints. Overall, we find some unexpected features to be important, such as the number of exons in an affected gene, while other features, such as whether an affected exon is constitutive, were surprisingly uninformative. We are currently using this tool to evaluate the pathogenicity of de novo and inherited structural variants revealed by linked-read sequencing of an undiagnosed cohort. Our classifier is available for download at compbio.berkeley.edu/proj/StrVCTURE.

A- A+

PgmNr 355: A digital diagnosis: The use of artificial intelligence in clinical exome analysis.

Authors:
A.B. Potter [1]; T.D. O'Brien [1]; N.E. Campbell [1, 2]; A. Frankenstein [1]; A. Kulkarni [1]; J.H. Letaw [1, 2]; C.S. Richards [1,3]

View Session Add to Schedule

Affiliations:
1) Knight Diagnostic Laboratories, Oregon Health & Science University, Portland, OR; 2) Department of Computational Biology, Oregon Health & Science University, Portland, OR; 3) Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR

Whole Exome Sequencing (WES) has been used as a clinical diagnostic tool for inherited rare disorders for over a decade. However, with massive numbers of variants detected in a typical human exome (~75k), identification of the causal variant can be challenging and time-consuming. Focused analysis utilizes human phenotype ontology (HPO) terms to generate specific genes related to patient phenotype. Identifying the causal gene and variant requires up-to-date information on newly described gene-phenotype and gene-disorder associations. Automated tools are ideal to search this vast array of new data sources. To evaluate whether an artificial intelligence (AI) tool can improve WES in our clinical laboratory, we evaluated our previously analyzed WES clinical cases. For all WES cases, medical records were searched to identify patient phenotypes, which were then converted to HPO terms for use with this tool. To determine whether an AI-supplemented system is equivalent to using a ranking-based non-AI gene prioritization tool, the positive cases detected by AI were analyzed using a gene prioritization tool. We analyzed 114 WES clinical cases using AI. The AI tool correctly identified the reported causal variant in 27/28 previously reported positive cases (96% concordance). AI solved complex cases including a mitochondrial pathogenic variant, a pathogenic large CNV, and a homozygous pathogenic missense variant due to uniparental disomy, demonstrating the broad scope of this tool. Further, analysis with the AI tool resulted in positive findings for 5 previously reported negative cases using new gene-phenotype or gene-disorder information not available at the initial time of analysis. In addition to increasing our positive rate of detection, this AI tool drastically reduced the time required for analysis. In contrast, the gene prioritization tool often did not prioritize the causal variant within the top 10 variants listed, and thus was less useful. In our experience the AI tool outperforms the ranking tool in finding diagnoses accurately and efficiently. Therefore, we have now adopted the AI tool to supplement our WES analysis. While AI has been criticized as a ‘black box’ approach, we would argue that the appropriate use of AI coupled with standard analysis tools is a valuable asset in the clinical genomics laboratory.

A- A+

PgmNr 356: Breaking the interpretation bottleneck: Examining the utility of an automated genomic interpretation algorithm in a clinical genetic lab.

Authors:
L. Meng [1,2]; R. Attali [3]; R. Dominguez-Vidana [2]; C. Taborda [2]; K. Bui [2]; Y. Regev [3]; N. Mizrahi [3]; A. Lev-Libfeld [3]; T. Talmy [3]; P. Smirin-Yosef [3]; R. Xiao [1,2]; I. Machol [2]; C. Eng [1,2]; F. Xia [1,2]; S. Tzur [3]

View Session Add to Schedule

Affiliations:
1) Molecular and Human Genetics, Baylor College Medicine, Houston, Texas.; 2) Baylor Genetics, Houston, Texas; 3) Emedgene Technologies, Tel Aviv, Israel

---

The utilization of WES and WGS in clinical practice has become widespread in recent years. However, the genotype-phenotype interpretation remains challenging and time consuming. By automating the variant prioritization and classification processes, machine learning technologies can unblock the genomic interpretation bottleneck, and improve the power and efficiency of the analysis.

In this study, 180 previously solved whole-exome sequencing cases (57 singles, 123 trios) were submitted to ‘Ada’, a machine learning based model which provides an automated interpretation of sequencing data. The model was previously trained on hundreds of solved cases, using combinations of variant attributes, with the aim to create a shortlist of an average of 10 variants that are most likely to resolve the case, and 104 candidate variants that might also be related with the patient phenotypes. It evaluates sequencing variant related with rare disorders only. As a key performance indicator, we measured the number of times the variants that were reported by the ABMGG board certified lab directors, appeared in the list of top 10 most likely variants suggested by ‘Ada’.

Results of auto-analysis of 180 cases show that ‘Ada’ accurately identified all reported variants as most likely or candidate variants. In addition, in trio cases the reported variants were designated as ‘most-likely’ in 98% (121/123) of the times, and in singleton cases in 91% (52/57) of the times. Overall, reported variants were designated in top 10 most likely list in 96% (173/180) of the cases, and 98% (176/180) in top 20.

Standard variant filtering based on variant call quality and minor allele frequency typically generates a list of 1000-2000 variants for whole-exome sequencing. By applying the ‘Ada’ algorithm, the number of variants requiring additional manual review was reduced to only 10 most likely variants, while analysis sensitivity was kept at 98% for trios, and 91% for singletons. This machine learning algorithm is currently being applied to automatically reanalyze a large number of historic unsolved cases, which would be impossible to analyze manually. This will allow clinical labs to achieve accurate and sustainable exome reanalysis with only minimal resources.

Overall, the study demonstrates the utility of machine learning models in automating genomic interpretation for large numbers of cases and providing effective decision support for clinical labs.

A- A+

PgmNr 357: Low-pass whole-genome sequencing versus chromosomal microarray analysis: Prospective implementation in prenatal diagnosis.

Authors:
Z. Dong [1,2]; H. Wang [3]; M. Chau [1,2]; T.Y. Leung [1,2,4]; S.W. Cheung [4,5]; Y.K. Kwok [1,2]; C.C. Morton [6,7,8,9,10]; Y. Zhu [3]; K.W. Choy [1,2,4]

View Session Add to Schedule

Affiliations:
1) Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, China; 2) Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China; 3) Maternal-Fetal Medicine Institute, Bao'an Maternity and Child Health Hospital Affiliated to Jinan University School of Medicine, Key Laboratory of Birth Defects Research, Birth Defects Prevention Research and Transformation Team, Shenzhen, China; 4) The Chinese University of Hong Kong-Baylor College of Medicine Joint Center For Medical Genetics, Hong Kong, China; 5) Department of Molecular and Human Genetics, Baylor College of Medicine Houston, Texas, USA; 6) Department of Obstetrics and Gynecology, Brigham and Women's Hospital, Boston, Massachusetts, USA; 7) Harvard Medical School, Boston, Massachusetts, USA; 8) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 9) Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA; 10) Manchester Center for Audiology and Deafness, University of Manchester, Manchester Academic Health Science Center, Manchester, UK

Purpose: Emerging studies suggest that low-pass whole-genome sequencing (WGS or GS) provides additional diagnostic yield of clinically significant copy-number variants (CNVs) compared with chromosomal microarray analysis (CMA). However, a prospective back-to-back comparison study evaluating the accuracy, efficacy, and incremental yield of low-pass GS as compared with CMA is warranted.

Methods: A total of 1,023 women undergoing prenatal diagnosis were enrolled. Each sample was subjected to low-pass GS and CMA for CNV analysis in parallel, and CNVs were classified according to the guidelines of the American College of Medical Genetics and Genomics.

Results: Low-pass GS not only identified all 124 numerical disorders or pathogenic or likely pathogenic (P/LP) CNVs detected by CMA in 121 cases (11.8%, 121/1,023), but also defined 17 additional and clinically relevant P/LP CNVs in 17 cases (1.7%, 17/1,023). In addition, low-pass GS significantly reduced the technical repeat rate from 4.6% (47/1,023) for CMA to 0.5% (5/1,023), and requires only 50-ng DNA as input.

Conclusions: In the context of prenatal diagnosis, low-pass GS identified additional and clinically significant information with enhanced resolution and increased sensitivity of detecting mosaicism as compared with CMA. Our study provides strong evidence for replacing routine CMA with low-pass GS as a first-tier prenatal diagnostic test.

Key words: molecular karyotyping; low-pass whole-genome sequencing; cryptic copy-number

variants; low-level mosaicism

A- A+

PgmNr 358: Low-pass WGS as an alternate cost-effective solution to chromosomal microarray analysis.

Authors:
J. Shen [1]; T. Chiang [1]; W. He [1]; M. Wang [2]; Y. Yang [1]; X. Wang [1]

View Session Add to Schedule

Affiliations:
1) AiLife Diagnostics, Pearland, TX 77584, USA; 2) Veritas-Genetics, China

---

Clinical diagnostic labs conventionally use chromosomal microarray analysis (CMA) as the standard approach to detect intra-chromosomal copy-number variants (CNVs) and chromosomal numerical anomalies, however recent advances clearly demonstrate the utility of low-pass whole-genome sequencing (WGS) as an alternative method with comparable diagnostic yields at significant cost-saving. Using our in-house CNV-seq detection and annotation pipeline, we analyzed a set of 85 patients (81 prenatal, 4 postnatal) referred for CNV analysis. Briefly, our pipeline strategy (1) consists of multiple CNV callers to enhance detection confidence, with the lower detection limit of 100 kb for 1x WGS, (2) adapts readily to different sequencing coverages, and (3) accounts for maternal cell contamination (MCC) when computing mosaicism and aneuploidies. Finally, it employs a comprehensive set of public resources for annotating, filtering and labelling CNVs. We report a total of 34 cases with aneuploidies (40%), 5 cases with heterozygous pathogenic CNVs (5.9%: 4 losses and 1 gains), and 3 cases with mosaic CNVs or aneuploidy (3.5%), resulting in an overall diagnostic yield of 49%. Among the 85 cases, 34 were tested for CMA and the results (13 cases with aneuploidy and 2 cases with intra-chromosomal CNVs) from both methods are 100% concordant. Interestingly, our CNV-seq pipeline detected a pathogenic intra-genic CNV of 40 kb (RPS6KA3) that was also identified by CNV calling from exome data. This work supports the utility of low-pass WGS in the clinical setting as a cost-effective alternative to CMA. It should be noted that low-pass WGS does not detect triploidy and MCC, for which alternative methods such as STR analysis should be used.

A- A+

PgmNr 359: Yield overpowers risk in amniocentesis performed in low-risk pregnancies.

Authors:
M. Shohat [1]; K. Hod [2]; B. Azaria [2]; I. Abadi-Korek [2]; Y. Segal [1]; R. Berger [1]; R. Moshonov [2]

View Session Add to Schedule

Affiliations:
1) Medical Genetics, Maccabi Health Services, Petah Tikva, Israel, Israel; 2) Assuta Medical Center , Habarzel, Atidim, Tel Aviv, Israel.

Background: New genetic technologies for detecting abnormalities in the amniotic fluid have improved the detection and prevention of severe diseases. This study aimed to investigate the yield versus the risk of such technologies, and if they should be recommended in pregnancies with low risk for genetic diseases.
Methods: A two-part study using clinical data from a large database of 30,830 singleton pregnancies at gestational age 16-23 weeks. First, the prevalence rates of unbalanced chromosomal anomalies detected by chromosomal microarray analysis (CMA) and karyotype were evaluated in low-risk pregnancies. Second, miscarriage rates at mid-trimester were compared between women who underwent amniocentesis (n=30,830) and women who did not undergo this procedure (n=98,590).
Results: Among low-risk pregnancies with normal karyotypes (n=4,174), the severe genetic diseases rate detected by CMA was 1:97, with no significant difference between maternal age groups. The likelihood of Down syndrome was negligible in this group. Even after the non-invasive prenatal test, the risk for genetic abnormalities with severe
consequences that can be detected only by CMA was 1:139. The overall miscarriage rate following amniocentesis was 1:1,401, and this rate was lower in low-risk pregnancies compared to non-low-risk pregnancies, and was not associated with the physicians' experience. Furthermore, the miscarriage rate following amniocentesis was lower than that of women with the same gestational age who did not undergo this procedure.
Conclusion: Given the low probability for miscarriage following amniocentesis, compared to the prevalence rate of severe genetic abnormalities detected by CMA, even women with low-risk pregnancies should be offered to undergo amniocentesis with CMA.

A- A+

PgmNr 360: The role of compound heterozygotes and *de novo* mutations in early human pregnancy loss.

Authors:
S.K. Garushyants [1]; E. Nabieva [1]; M.D. Logacheva [1,2]; T.V. Neretina [2]; A. Fedotova [1,2]; V. Moskalenko [2]; N. Libman [3]; R. Bikanov [3]; E. Pomerantseva [3]; D. Pyankov [4]; I. Kanivets [4]; T. Serebrennikova [5]; E. Glazyrina [5]; G.A. Bazykin [1]

View Session Add to Schedule

Affiliations:
1) Center for Life Sciences, Skolkovo Institute of Science and technology, Moscow, Russian Federation; 2) Lomonosov Moscow State University, Moscow, Russia; 3) Genetico LLC, Moscow, Russia; 4) Genomed, Moscow, Russia; 5) Progen, Moscow, Russia

Only a third of human conceptions result in live births. While a large fraction of pregnancy losses are associated with chromosomal abnormalities, the causes of spontaneous abortion in remaining cases remain largely unknown. A few recent studies employ whole-exome sequencing to uncover the genetic basis of prenatal lethality in euploid embryos, but they focus mainly on embryos of advanced gestational age or with pronounced developmental pathologies. Accumulation of more clinical data, especially with an emphasis on early-term cases, will add to the emerging picture of the diversity of causes of embryonic lethality.

Here, we report on the progress of a whole-exome sequencing study of spontaneous euploid abortuses below 22 weeks gestational age, that aims to determine the inherited or *de novo* genetic variation that could have led to pregnancy loss. Most abortuses in our sample have no evident developmental pathologies. We plan to sequence 100 mother-father-abortus trios; to date, we have collected 78 and sequenced and analyzed 38 trios. The mean gestational age was 12 weeks. As candidate causative inherited or *de novo* variants, we prioritized those in genes associated with embryonic lethality in mice, and in genes with high intolerance to protein truncating variants (PTVs). We only considered nonsynonymous variants with Polyphen2 score >0.996 and CADD score > 20. For inherited variants, we required allele frequency below 1% and absence in homozygous state in the general population. For *de novo* variants, we required absence in the general population. Furthermore, we only considered mutations in genes with demonstrated gene expression during embryogenesis.

We were able to detect candidate variants in 15/38 trios. In 7 of these cases, we found compound heterozygotes in essential genes *PNPT1*, *LAMA5*, *ZNF628*, *SSFA2*, *SH3RF3*, and *UTRN*. Mutations in *LAMA5* were observed in two cases in our sample. In mice, mutations in *LAMA5* have been shown to cause embryonic lethality. In 8 of the remaining cases, we propose that the cause of pregnancy loss is associated with *de novo* mutations (DNMs). In one case, we observed 6 nonsynonymous DNMs, with the most damaging one in *SOX4* (p.Gly420Asp). In another case, we observed 3 probably damaging nonsynonymous mutations: in *APOBEC3G* (p.Thr218Ile), *GSN* (p.Val522Phe), and *PKHD1* (p.Gly1321Glu). We thus demonstrate that *de novo* and compound heterozygous single-nucleotide variants may be important contributors to human pregnancy loss.

A- A+

PgmNr 361: Exome sequencing of 268 stillbirth cases emphasizes the importance of diagnostic sequencing and implicates highly constrained novel genes not currently associated with human disease.

Authors:
J.L. Giordano [1]; K. Stanley [3]; C. Buchovecky [2]; A. Thomas [2]; M. Ganapathi [2]; J. Liao [2]; A.V. Dharmadhikari [2]; A. Revah Politi [3]; M. Ernst [3]; N. Lippa [3]; H. Holmes [3]; R.M. Silver [4]; N. Stong [3]; V. Aggarwal [3]; R. Wapner [1]; D. Goldstein [3]; Stillbirth Collaborative Research Network

View Session Add to Schedule

Affiliations:
1) Dept of OB/GYN MFM, Columbia University, New York, New York; 2) Dept of Pathology and Cell Biology, Columbia University, New York, New York; 3) Institute for Genomic Medicine, Columbia University, New York, New York; 4) Dept of OB/GYN, University of Utah Health, Salt Lake City, Utah

---

Background The etiology of stillbirth remains largely unexplained despite detailed clinical and laboratory investigation. Genetic evaluations thus far indicate 10-20% of stillbirths are attributed to chromosomal abnormalities, however, the benefit of exome sequencing in this population is unknown. Therefore, we applied an exome-wide diagnostic framework to evaluate the diagnostic yield of exome sequencing in a cohort of 268 unexplained stillbirths.

Methods We generated whole-exome sequence data for 268 unexplained stillbirth cases with non-causative karyotype/microarray from the Stillbirth Collaborative Research Network and identified pathogenic/likely pathogenic variants in disease associated genes, which may lead to stillbirth. These included genes previously associated with stillbirth and those representing strong candidates for phenotype expansion to include stillbirth. We also evaluated whether stillbirth cases are more likely to carry variants in genes strongly depleted of loss-of-function (LoF) variation in humans when compared to 7,239 population-matched controls.

Results We identified diagnostic genotypes in 12 of 268 (4.5%) stillborn cases spanning 9 distinct Mendelian disease genes previously implicated in stillbirth *(PTPN11, ARID1B, KCNH2, MYBPC3, RYR2, HCN4, TRPM4, HNF1B, COL1A1)*. We further identified 11 cases (4.1%) with putative pathogenic mutations in 10 different disease genes that are not currently associated with stillbirth but are good candidates for phenotype expansion *(SMC3, FBN2, BCOR, EOGT, MIB1, DSC2, TPM1, GREB1L, MYT1L, DMD)*. Together, these findings indicate a potential cumulative diagnostic yield of 8.6% in known Mendelian disease genes. We also report a significant case-control enrichment of LoF variants in genes highly intolerant to such variants in the human population (OR = 2.29, 95% CI 1.61–3.19, logistic p-value = 1.91 x 10^{-6}). We found the risk signal in these highly constrained genes is driven primarily by genes not currently associated with human disease (OR = 2.31, 95% CI 1.54 – 3.36, logistic p-value = 2.35 x 10^{-5}). These findings suggest that a number of novel genes confer risk for stillbirth.

Interpretation Our findings establish exome sequencing as an important diagnostic modality in stillbirth. Notably, our data suggest the phenotype of stillbirth represents an expansion of currently described Mendelian diseases as well as an understudied phenotype caused by a number of novel genes essential for *in utero* survival.

A- A+

PgmNr 362: Exome sequencing for infants with a prenatally identified fetal structural anomaly: Diagnostic yield and changes in management.

Authors:
A.S. Freed [1]; S.V. Clowes Candadai [2,3]; M.C. Sikes [4]; J. Thies [4]; H.M. Byers [5]; J.N. Dines [1]; M.K. Ndugga-Kabuye [1]; K. Fogus [4]; H.C. Mefford [4]; C. Lam [4]; M.P. Adam [4]; A. Sun [4]; R. DiGeronimo [6]; K.M. Dipple [4]; G.H. Deutsch [7]; Z.C. Billimoria [6]; J.T. Bennett [1,4,8]

View Session Add to Schedule

Affiliations:
1) Department of Pediatrics, Division of Genetic Medicine, University of Washington, Seattle, WA; 2) Department of Laboratories, Seattle Children’s Hospital, Seattle, WA; 3) Patient-centered Laboratory Utilization Guidance Services (PLUGS), Seattle Children's Hospital, Seattle, WA; 4) Department of Genetics, Seattle Children’s Hospital, Seattle, WA; 5) Division of Medical Genetics, Department of Pediatrics, School of Medicine, Stanford University, Palo Alto, CA; 6) Department of Pediatrics, Division of Neonatology, University of Washington, Seattle, WA; 7) Seattle Children’s Hospital and University of Washington, Department of Pathology, Seattle, WA; 8) Center for Developmental Biology and Regenerative Medicine, Seattle Children’s Research Institute, Seattle, WA

Purpose:

Exome sequencing (ES) is rapidly becoming standard of care for diagnosis of critically ill neonates. Exomes are usually initiated after the infant is born, yet many present in utero with fetal structural anomalies (FSAs). Prenatal ES studies have focused on the diagnostic yield as a primary outcome; no studies have systematically evaluated whether prenatal molecular diagnosis could change management of the pregnancy or newborn. We hypothesized that prenatal diagnosis through ES would change in utero and/or neonatal management.

Methods:

In October 2016, Seattle Children’s Hospital instituted a process for rapid (~ 1 week TAT) trio ES for children in the ICU. Urgent (~4 week TAT) trio or duo ES was available for all hospitalized children. Data is reported through January 2019. Although we did not perform ES during pregnancy, we retrospectively chart-reviewed all of our cases for the presence of an FSA, and then determined if knowledge of genetic diagnosis during pregnancy would have changed management. Isolated growth restriction, macrosomia, and cleft lip were excluded.

Results:

42 neonates (defined as infants <1 year old and hospitalized since birth) underwent ES over a 15 month period. FSAs had been identified in 52% (22/42) of patients. The most common FSAs were hydrops fetalis (27%) and congenital heart defects (18%). In 45% (10/22) of the infants with an FSA, the ES was diagnostic. In 36% (8/22) of the infants, the ES process was initiated in the first 4 days of life. In those 8 patients, 5 had diagnostic results and 4 led to changes in medical management or a

change to palliative care. None of the results would have led to an in utero therapy. Two diagnoses may have led to an infant not being transferred to a tertiary care center and rather given palliative care at the delivery hospital. Two diagnoses could have led to earlier treatment with bisphosphonates and biotin for generalized arterial calcification of infancy and biotinidase deficiency.

Conclusions:

When analyzed retrospectively, ES had a high (52%) diagnostic yield for prenatally identified FSA and would have led to a change in neonatal medical management in about half of cases with a molecular diagnosis. We conclude that prenatal ES for FSA can provide diagnostic information that informs neonatal medical management. In the future, larger, prospective studies of ES for FSA will be needed to demonstrate the importance of prenatal ES to hospitals, laboratories, and payers.

A- A+

PgmNr 363: *MRP:* Exome rare-variant analysis via Bayesian model comparison prioritizes strong risk and protective effects across biomarkers and diseases.

Authors:
G.R. Venkataraman; M. Aguirre; Y. Tanigawa; M.A. Rivas

View Session Add to Schedule

Affiliation: Biomedical Data Science, Stanford University, Stanford, CA

Whole genome sequencing studies applied to large populations or biobanks with extensive phenotyping raise new analytical challenges. The need to consider many variants at a locus or group of genes simultaneously and the potential to study many correlated phenotypes with shared genetic architecture provide opportunities for discovery and inference that are not addressed by the traditional “one variant - one phenotype” association study. Here we introduce a model comparison approach we call **M**ultiple **R**are-Variants and **P**henotypes (MRP) for rare-variant association studies that considers correlation, scale, and location of genetic effects across a group of genetic variants, phenotypes, and studies. We use summary statistic data and apply univariate and multivariate gene-based meta-analysis models to identify those rare-variant associations that have protective or risk effects and can expedite drug discovery. We apply the method to 2,540 UK Biobank phenotypes across array genotype data for 337,151 white British individuals and exome data for a subset (34,395) and demonstrate that the model comparison approach can aggregate rare-variant association signals for greater power. We are able to find both previously-documented and novel associations between genes and several anthropometric phenotypes ($\log_{10}$ Bayes Factor > 5), including those between *GPR151* and arm and trunk fat masses; *LRP5*, *KCNK16*, *SSPO* and several heel bone mineral density phenotypes; *PAM* and hand grip strength; *PINX1* and nucleated red blood cell count; *PLCG2* and nucleated red blood cell percentage; *HELB* and age at menopause; *CD34* and pulse rate; *SLC45A2*, *TYR* and skin color; *APRT* and red hair color; and *DCLRE1A* and facial hair growth. Additional associations between genes and biomarkers include *SLC22A2* and Apolipoprotein-B levels; *CRP* and C-reactive protein; *SCL34A1* and covariate- and statin-adjusted Cystatin-C; and *SORT1*, *MSR1*, and *IGFBP3* and IGF-1 levels. Further associations between genes and sensory phenotypes, like *PMS1* and changes in the speed or amount of moving/speaking, also are uncovered. Most importantly, we find gene-disease associations between *MYOC* and intraocular pressure and *FLG* and atrial flutter/fibrillation. Overall, we show that the MRP model comparison approach is able to retain and improve upon useful features from widely-used meta-analysis approaches in order to prioritize actionable gene targets.

A- A+

PgmNr 364: Correlations between polygenic risk score predictors suggest a method to detect sub-phenotypes in GWAS cohorts.

Authors:
J. Yuan; H. Xing; A. Lamy; I. Pe'er

View Session Add to Schedule

Affiliation: Computer Science, Columbia University, New York, NY.

Genome-wide association studies for some diseases such as schizophrenia are thought to be confounded by the presence of heterogeneous clusters of affected individuals with unique sub-phenotypes. We present a method to detect this heterogeneity using only phenotype labels and predictors in the form of genotypic or transcriptomic data from a single cohort. While these predictors are often selected based on LD to be uncorrelated over the study cohort, we demonstrate that these variables exhibit predictable nonzero correlation patterns over subsets of individuals thresholded by polygenic risk scores (PRSs), such as all cases in a case/control study. These correlations are a mathematical consequence of all thresholded linear models including logistic and liability (probit) regression. We develop a test statistic to evaluate the overall nonzero correlation bias expected of a cohort absent any heterogeneity arising from sub-phenotypes or other sources. We further generalize this score to apply to a wide variety of GWAS scenarios, including those with quantitative predictors as in transcriptome-wide association, and those with quantitative phenotypes, by introducing a weighting function over PRSs in lieu of a discrete case/control split. We demonstrate through simulations that for PRSs with genomic variance explained as low as 5% and sample sizes in the range of 50k-100k, typical of modern GWAS, cohorts generated from distinct sub-phenotypes exhibit altered correlation patterns that can be distinguished by our method from the null scores expected of a single homogeneous PRS. Lastly, we apply this method to characterize the heterogeneity of genotype and expression data of schizophrenia patients in the Commonmind Consortium and the Psychiatric Genomics Consortium.

A- A+

PgmNr 365: Using model predictions as quantitative traits improves power in a genome-wide association study of acute ischemic stroke in the UK Biobank.

Authors:
P.M. Thangaraj [1,2]; N.P. Tatonetti [1,2,3]

View Session Add to Schedule

Affiliations:
1) Department of Biomedical Informatics, Columbia University, New York, NY; 2) Department of Systems Biology, Columbia University, New York, NY; 3) Department of Medicine, Columbia University, New York, NY

Large genetic repositories connected to electronic health records (EHR) promise the ability to perform thousands of genetic studies using routinely captured data. High-throughput identification of cases and controls can be difficult, however, due to time-consuming chart review and incompleteness of medical records. Machine learning methods can expand cohorts by assigning every patient a probability of disease. For example, genetic studies of stroke, a leading cause of death, require hundreds of thousands of patients. We hypothesized that the output of a supervised machine learning classifier can be used as a proxy variable for stroke and is an efficient strategy for expanding cohort size.

We tested our strategy in the UK Biobank, a prospective health study of over 500,000 participants with EHR and genetic data. To determine the feasibility of converting a binary trait to quantitative, we trained 5 classifiers on cases with acute ischemic stroke (AIS) and controls with no cerebrovascular disease using EHR-extracted features and tested on every patient in the UK Biobank for probability assignment of AIS. The top probabilities from the model-predicted cohort were significantly enriched for self-reported AIS patients without AIS diagnosis codes (65-250 fold over expected). For each of the models, 8,800-42,000 patients were found in the top 10% quantile, compared to 2,959 known AIS patients. We hypothesized that using the output model probabilities as a quantitative trait would boost power of a GWAS over the typical binary AIS assignment. For the binary trait, we fit a logistic regression, and for the model probabilities, we fit a linear regression against 26 SNPs known to be associated with stroke from previous GWAS. We found an increase in significant stroke SNPs by the model probabilities. Some p-values improved as much as five orders of magnitude (eg. pbin=0.00135, pqt=2.17e-8, rs635634) while a random sample of 1000 non-associated variants’ p-values remained largely unchanged (mean-fold change =1.08, 95% CI 1.04-1.11). Over 26 stroke markers, we found a mean-fold increase of 1.77 in significance (95% CI 1.05-3.02). Finally, we ran a GWAS with the elastic net and random forest models and found 20 and 402 variants that reached significance (p<5e-08), respectively, compared to mean 1.1 and 0.22 variants after permutation of the probabilities. We demonstrate that our quantitative proxy trait can significantly improve power over its respective binary trait.

A- A+

PgmNr 366: Detection of local genetic correlation by a scan statistic approach.

Authors:
H. Guo [1,2]; Q. Lu [3,4]; L. Hou [1,2,5]

View Session Add to Schedule

Affiliations:
1) Center for Statistical Science, Tsinghua University, Beijing, China; 2) Department of Industrial Engineering, Tsinghua University, Beijing, China; 3) University of Wisconsin, Madison, WI, United States of America; 4) Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, United States of America; 5) MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China

---

Genome Wide Association Studies (GWASs) have been carried out for many traits and diseases, yet our understanding of most traits' genetic basis remains incomplete. Investigation of the shared genetic architecture across multiple traits quickly gained popularity in the past few years and provided fundamental new insights into trait and disease etiology. In particular, methods have been developed to estimate the genetic correlation between traits, which is an efficient way to quantify the overall genetic sharing between complex traits. However, it is unclear which part of the genome contributed to the genetic correlation. In this work, we propose a computational method that scans the genome to detect small segments harboring genetic correlation signals. Compared to existing methods, our approach does not need pre-specified candidate regions, only uses GWAS summary statistics, is robust to sample overlap between studies, and accounts for linkage disequilibrium. Through extensive numerical simulations, we show that the proposed method successfully detects the biological signals in various settings. We have applied our method to study the local genetic correlation of two neurodevelopmental psychiatric disorders, attention deficit/hyperactivity disorder (ADHD; n=53,293) and autism spectrum disorder (ASD; n=46,350). Our analysis highlighted a specific region of 450 Kb on chromosome 20 (p11.22) showing a positive sharing between ASD and ADHD (p=0.022 under a stringent control for family-wise error rate). This region is a known risk locus for ASD but did not previously reach genome-wide significance in the ADHD GWAS (lowest p=1.3e-6 in the European population). These results demonstrated that the proposed approach is a powerful tool to detect genetic overlaps between complex traits.

A- A+

PgmNr 367: Mitigating batch effects in >23,000 WGS samples for a case-control GWAS of kidney disease.

Authors:
T. Soare [1]; W. Zhang [1]; A. Tebbe [1]; V. Mandal [1]; D. Borges-Rivera [1]; N. Chennagiri [1]; M. Kretzler [2]; G. Nadkarni [3]; R. Gbadegesin [4]; J. Wenke [5]; D. MacArthur [6,7]; J. Reilly [1]; P. Mundel [1]; L. Walsh [1]; T. Tibbitts [1]

View Session Add to Schedule

Affiliations:
1) Goldfinch Bio, Inc., Cambridge, MA, USA; 2) University of Michigan, Ann Arbor, MI, USA; 3) Mount Sinai Hospital, New York City, NY, USA; 4) Duke University Medical Center, Durham, NC, USA; 5) Nashville Biosciences, Nashville, TN, USA; 6) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 7) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

Background: Focal segmental glomerulosclerosis (FSGS) is scarring of the kidney that can lead to kidney failure. To discover loci associated with FSGS, we built the Kidney Genome Atlas (KGA), which currently contains WGS (>30X) on 23000 individuals, including 2000 cases of FSGS and other proteinuric disorders. Cases and controls were sourced from 5 clinical sites and 21 publicly-available cohort studies, with incomplete metadata on sequencing platform. Subtle differences in sequencing data acquisition (batch effects), when confounded with case-control status, lead to inflation of test statistics. We describe our data-driven approach to achieve a well-calibrated dataset for case-control GWAS.

Methods: We utilized standard WGS QC metrics, sex and ancestry inference, and relatedness to filter samples. To further determine and control for batch effects, we implemented a data-driven approach: (1) to identify clusters of similar sequencing technologies, conduct a PCA on depth of coverage (DP); (2) to quantify case-control dissimilarity at the dataset-level, implement a novel use of silhouette scores and permutations; and (3) to achieve a better-calibrated dataset for case-control GWAS, iteratively remove controls with high dissimilarity from cases. Using common variants, we conducted a GWAS and an expression quantitative trait loci (eQTL) analysis on a subset of 300 patients with transcriptomic data.

Results: DP was extracted at 118k high-confidence, independent single nucleotide variants. The top 5 PCs explained >96% of the variance in DP. We permuted case-control status among samples 1000 times and calculated the difference between the observed average silhouette score to this null distribution for a quantitative metric of case-control dissimilarity. We removed outlying controls to minimize this metric. A preliminary GWAS in AFR-ancestry individuals (n=904) at 5M common autosomal loci controlling for sex, 6 ancestry PCs, and 5 DP PCs showed that inflation was well controlled (lambda=1.03). Further, a preliminary eQTL analysis revealed 3562 genes had eQTLs in either microdissected glomeruli or tubules.

Conclusions: Integrating WGS samples from many sources presents challenges for downstream analyses, for example the presence of batch effects that may be confounded with case-control status. Minimizing a metric of case-control dissimilarity on DP in turn reduces inflation of tests statistics, providing a quantitative method to aid in achieving well-calibrated GWAS.

A- A+

PgmNr 368: Identifying robust trans-associations via a cross-condition mediation analysis and validating the trait-associations for trans-genes of GWAS loci.

Authors:
L.S. Chen [1]; F. Yang [2]; K.J. Gleason [1]; J. Wang [3]; J. Duan [4]; X. He [5]; B.L. Pierce [1,5]

View Session Add to Schedule

Affiliations:
1) Department of Public Health Sciences, Univ Chicago, Chicago, Illinois.; 2) Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado.; 3) Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA; 4) Center for Psychiatric Genetics, NorthShore University HealthSystem, Chicago, IL; 5) Department of Human Genetics, University of Chicago, Chicago, IL

Trans-acting eQTL effects explain a substantial proportion of gene expression variation. Yet given that they predominantly act in a tissue-specific manner, it is still underpowered to detect trans-eQTLs associated with expression levels of distal or inter-chromosomal genes in most tissue types and cellular contexts. It is reported that many trans-associations are (at least partially) mediated via cis-gene expression levels, i.e., SNPs affects cis-gene expression levels then further affects trans expression levels. Testing for trans-associations mediated by cis-genes complements the standard trans-eQTL tests and can detect trans-associations with interpretable mechanisms. Moreover, we show that trans-associations mediated by cis often have effects shared across tissue types. Therefore, we propose a Cross-Condition Mediation analysis (CCmed) method for detecting robust cis-mediated trans-associations based on only summary statistics of cis-association and conditional correlations of cis and trans expression levels from different tissue types/cellular conditions. We applied our method to the eQTL data from 13 brain tissue types of the GTEx project and identified thousands of trans-associations with effects shared across brain tissues. We further replicated our cross-tissue trans-association findings in two independent eQTL consortia, the eQTLGen and the CommonMind Consortium. In the second part of the work, by focusing a 108 known schizophrenia (scz)-associated GWAS SNPs, we identified the suspected trans-genes associated with the GWAS SNPs. We further proposed a novel mediation validation framework. Our validation framework considers multiple local eQTLs of the suspected trans-genes as multiple instruments to validate the trait-associations of trans-genes using only GWAS and eQTL summary statistics. In contrast to existing mediation methods, our validation method is robust to the presence of a small to moderate proportion of invalid instruments, i.e., allowing some SNPs with genetic effects on complex traits not completely mediated by trans-gene expression. By applying the proposed method to the putative trans-genes for the 108 scz risk loci, we validated the trait-associations of our trans-genes by recapitalizing on GWAS summary statistics from the Psychiatric Genomics Consortium (PGC).

A- A+

PgmNr 369: Phase 2/3 trial to assess the safety and efficacy of Lenti-D autologous hematopoietic stem cell gene therapy for cerebral adrenoleukodystrophy.

Authors:
F. Eichler [1]; C. Duncan [2]; P. Orchard [3]; S. De Oliveira [4]; A. Thrasher [5]; T. Lund [3]; C. Sevin [6]; P. Gissen [5]; H. Amartino [7]; N. Smith [8]; E. Shamir [9]; W. Chin [9]; E. McNeil [9]; P. Aubourg [10]; D. Williams [11]

View Session Add to Schedule

Affiliations:
1) Massachusetts General Hospital and Harvard Medical School, Boston, MA; 2) Boston Children's Hospital and Dana-Farber Cancer Institute, Boston, MA; 3) University of Minnesota Children's Hospital, Minneapolis, MN; 4) University of California, Los Angeles, CA; 5) University College London Great Ormond Street Hospital Institute of Child Health and Great Ormond Street Hospital NHS Trust, London, UK; 6) Hôpital Universitaire Hôpital Bicêtre-Hôpitaux Universitaires Paris Sud, Paris, France; 7) Fundacion Investigar, Buenos Aires, Argentina; 8) Women's and Children's Hospital, Adelaide, Australia; 9) bluebird bio, Inc., Cambridge, MA; 10) INSERM & Hôpital Bicêtre, Paris, France; 11) Dana-Farber and Boston Children's Cancer and Blood Disorders Center, Boston Children's Hospital, Harvard Medical School and Harvard Stem Cell Institute, Boston, MA

Cerebral adrenoleukodystrophy (CALD) is a rare, X-linked, metabolic disease in which dysfunctional ALD protein leads to the toxic accumulation of very-long chain fatty acids, primarily in the adrenal cortex and white matter of the brain. CALD is characterized by rapidly progressive inflammatory cerebral demyelination leading to irreversible loss of neurologic function and death. There is currently no treatment approved for CALD, although allogeneic hematopoietic stem cell (HSC) transplantation has been shown to have a beneficial effect on clinical indices of disease and long-term survival, if performed early.

Lenti-D Drug Product (DP), an autologous HSC gene therapy, is in clinical development for the treatment of CALD. In an open-label phase 2/3 study (ALD-102) of the safety and efficacy of Lenti-D, boys with early CALD (≤17 years and evidence of contrast enhancement on MRI, Neurologic Function Score ≤1, and Loes score >0.5 and ≤9) were fully myeloablated with busulfan and cyclophosphamide prior to infusion of autologous CD34+ cells transduced with elivaldogene tavalentivec (Lenti-D) lentiviral vector. The primary efficacy endpoint is the proportion of patients who are alive and free of major functional disabilities (MFD) at Month 24. The primary safety endpoint is the proportion of patients who experience acute (≥Grade 2) or chronic graft-versus-host disease (GVHD) by Month 24. As of April 2019, the trial was fully enrolled with 32 patients having received Lenti-D DP (median follow-up 21.2 months, min-max, 0.0-60.2). Fifteen patients have completed ALD-102 and are being followed in a long-term follow-up study; 14 patients remain in ALD-102. Two patients were withdrawn and referred for allo-HSCT before their Month 24 visit; another experienced rapid disease progression resulting in MFDs and death. Of patients who have reached (or would have reached) 24 months of treatment, 15 of 17 (88.2%) were alive and MFD-free. All Lenti-D DP-treated patients generally showed evidence of neurologic function stabilization at their last follow-up. To date, there have been no reports of graft failure, GVHD, or transplant-related mortality and recorded adverse events have proved consistent with myeloablative conditioning. There is no evidence of replication competent lentivirus or insertional oncogenesis.

These data suggest that Lenti-D DP stabilizes neurologic disease progression and appears to be a promising gene therapy for CALD.

A- A+

PgmNr 370: The first human single cell atlas of the Substantia nigra reveals novel cell-specific pathways associated with the genetic risk of Parkinson’s disease and neuropsychiatric disorders.

Authors:
C. Sandor [1]; D. Agarwal [2,3]; V. Volpato [1]; T. Caffrey [3]; J. Alegre-Abarrategui [3,4,5]; R. Wade-Martins [3]; C. Webber [1,3]

View Session Add to Schedule

Affiliations:
1) UK Dementia Research Institute, Cardiff University, Cardiff, UK; 2) Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, UK; 3) Department of Physiology, Anatomy , Genetics, University of Oxford, Oxford, UK; 4) Department of Neuropathology , University of Oxford, Oxford, UK; 5) Department of Medicine, Imperial College London, London, UK

We describe the first single-nuclei transcriptomic atlas of the human **Substantia Nigra (SN)**, a brain region playing an important role in reward and movement, and use this atlas to interpret the genetic architecture of many disorders impacting the SN.

To create this atlas, we sequenced ~ 17,000 nuclei from both the cortex and SN of five human brains. The SN contained a much higher proportion of oligodendrocytes as compared to the mainly neuronal cortex. We show that large difference in the glial/neuron cell proportion explain why many studies performed with bulk brain transcriptomic profiles fail to identify brain-specific associations with neurological disorders as the relevant genetic risk variants map to under-represented cell-populations.

By mapping genetic variants associated with different human traits to specific SN cell types, we show for the first time that common genetic variation for **Parkinson’s disease (PD)**, for which the symptoms are caused by the loss of the dopaminergic neurons (DaNs) within the SN, is indeed associated with DaN-specific gene expression and pathways previously associated with PD, such as mitochondrial organization and functioning, protein folding and ubiquitination. Unlike Alzheimer’s disease, we find no association between microglia and PD genetic risk suggesting a less causal role for neuroinflammation. We also identified a distinct and specific cell association of PD risk with oligodendrocyte-specific expression causally implicating endocytosis and autophagy networks within this cell type complementing reports that G2019S-*LRRK2* transgenic mice exhibit oligodendrocyte vacuolization leading to neuronal degeneration.

Beyond PD, we find SN cell types associated with different neuropsychiatric disorders, particularly **schizophrenia (SCZ)** and **bipolar disorder (BP)**. Using our matched cortex/nigral samples, we find distinct associations with SCZ genetic risk for both pyramidal (synaptic functioning) and dopaminergic neurons (lipid metabolic processes). As with PD, we also identified distinct glial cell signatures for SCZ and BP within oligodendrocyte and astrocyte cell populations, respectively. Specifically, BP risk associates with apoptotic processes in reactivate astrocytes. This atlas provides the first robust associations between genetic risk of multiple disorders and the midbrain cell types these risks likely manifests through, thereby directing our aetiological understanding.

A- A+

PgmNr 371: Genomic analyses in 3 million individuals identify genetic determinants of questionnaire response bias and study participation with potential implications for GWAS interpretation.

Authors:
A. Ganna [1,2,3]; G. Mignogna [2,3,4]; B. Hollis [5]; C. Carey [2,3]; R. Walters [2,3]; M.D. Van der Zee [6]; . 23andMe Research Team [7]; P. Joshi [8]; N. Pirastu [8]; B. Neale [2,3]; M. Nivard [6]; J.R.B. Perry [5]

View Session Add to Schedule

Affiliations:
1) FIMM, Helsinki, Finland; 2) Massachusetts General Hospital, Boston, USA; 3) Broad Institute, Cambridge, USA; 4) University of Milano Bicocca, Milan, Italy; 5) University of Cambridge, Cambridge, UK; 6) Vrije University, Amsterdam, Netherlands; 7) 23andMe, Inc. Mountain View, USA; 8) University of Edinburgh, Edinburgh, UK

Genetic association results are often interpreted with the assumption that voluntary participation in the study in itself does not affect allele frequencies distribution and the size of biases in recall of participants does not associate with genetic variants. To explore this in detail, we assessed multiple forms of bias in up to 3 million genotyped research participants from the UK Biobank and 23andMe, Inc.

Firstly, we identified 96 genome-wide significant signals marking autosomal allele frequency differences between men and women that could not be attributed to technical artefacts. For example, the *FTO* body mass index (BMI) raising allele was observed at higher frequency in men compared to women ($P=1.9\text{x}10^{-39}$). This association was replicated in UK Biobank ($P=4\text{x}10^{-5}$) and observed in younger age-restricted analyses ($P=1.8\text{x}10^{-5}$), suggesting the effect is not driven by survival bias. Mendelian Randomization analyses demonstrated this effect is a common feature of BMI-associated variants. We also observed sex discordant allele frequency differences across variants associated with educational attainment. Genetically higher educated men were more likely to be participants of 23andMe than women (r_g 0.34, $P=4\text{x}10^{-70}$). Intriguingly the direction of effect was opposite in UK Biobank, where higher educated women were more likely to participate (r_g 0.25, $P=3.3\text{x}10^{-9}$). We conclude that such genetic variants influence propensity to participate, that the effect size can be sex-specific and recruitment strategy specific.

Secondly, we identified 16 genome-wide significant signals associated with the likelihood of either being a non-responder to questions or not knowing the answer. Overall we saw strong genetic correlations with multiple aspects of health and wellbeing. We conclude that genetic variants may cause associations with self-reported health traits through bias in recall.

Finally, we developed a likelihood framework to identify individuals whose questionnaire answers consistently deviate from their genetic prediction and show how this indicator can capture response bias.

These findings demonstrate that GWAS sample sizes are now of sufficient size to detect participation and response bias at high significance, which may manifest differently by study design. Our ongoing work seeks to understand how these sources of bias may influence the identification and interpretation of genetic associations and causal inferences arising from such large-scale studies.