ASHG 2023 Platform Abstracts

As of November 11, 2023

Instructions

1. View the Table of Contents on the next page.
    a. The Table of Contents is sorted chronologically by the Session Number.
2. Click on the Abstract Title of the abstract you would like to view, and you will be taken to the page the abstract is on.

Table of Contents

Session 008: Advancing diagnoses for rare genetic diseases

Location: Conv Ctr/Ballroom A/Level 3

Session Time: Thursday, November 2, 2023, 8:30 am - 10:00 am

Title: Interpretation of whole genome sequencing for clinical care in rare disease on a national scale: the UK experience

Authors:

D. Kasperaviciute[1], S. Walker[1], K. Savage[1], L. Abrahams[1], E. Serra[1], J. Ellingford[1], J. Lopez[1], D. Smedley[2], R. Mein[3], E. Thomas[1], Z. Deans[3], S. Hill[3], R. Scott[1], A. Rendon[1]; [1]Genomics England, London, United Kingdom, [2]Queen Mary Univ. of London, London, United Kingdom, [3]NHS England, London, United Kingdom

Abstract:

Building on the 100,000 Genomes Project, the NHS in England launched a centralised whole genome sequencing (WGS) diagnostic service. For patients with rare diseases, WGS is offered for 34 distinct clinical indications. Prior to June 2023, 46,347 genomes from 22,877 families were processed, with an overall diagnostic yield of 30% among interpreted cases. Currently, ~600 whole genomes are being processed per week, creating a challenge for variant interpretation, which heavily depends on manual review and classification.

Our variant triaging approach enables scalable and equitable WGS interpretation for clinical purposes, and we continue to refine it balancing sensitivity and specificity. Together with NHS interpreting laboratories, we defined the minimal interpretation workflow required for each case, which in a trio-based analysis, assuming complete penetrance, reduces the number of variants to be reviewed by ~84% compared with a purely genotype-led approach. This workflow involves reviewing small variants, copy number variants (CNVs) and short tandem repeat (STR) expansions in genes relevant to the referral phenotype (as defined in PanelApp), followed by reviewing the variants prioritised by Exomiser and de novo variants genome-wide.

In our results, 40.5% of diagnostic variants had pathogenic or likely pathogenic classifications in either internal or external databases at the time of interpretation. Therefore, sensitivity and interpretation efficiency can be increased by reviewing these variants first. We have developed the Clinical Variant Ark (CVA) database, which currently hosts information on >4,000,000 variants, >7,000 phenotypes and >11,000 pathogenic/likely pathogenic variants. It accelerates interpretation by highlighting previously classified variants. Specificity is substantially improved with the use of pipeline-matched internal variant allele frequency thresholds, on average reducing the number of prioritised variants per trio by 30% for small variants and 91% for CNVs.

Among the diagnoses made outside our core interpretation set, about half are due to more recent gene-disease associations. The rest are accounted for by a variety of more complex scenarios, including mosaicism, genetic heterogeneity, complex variants and others.

Over 77% of patients receiving a clinical WGS test consent to add their genome for research purposes to the National Genomic Research Library. This enables ongoing research-driven analysis for new diagnoses, including complex variants and new gene discovery, which are returned to NHS laboratories for reporting. This approach has already proved to be successful in the 100,000 Genomes Project.

Title: Boosting Rare Disease diagnostic yield with Artificial Intelligence: Validation of the eVai "Suggested Diagnosis" framework

Authors:

G. Nicora[1], F. De Paoli[1], S. Berardelli[1,2], I. Limongelli[1], E. Rizzo[1], R. Bellazzi[2], P. Magni[2], **S. Zucca**[1]; [1]enGenome srl, Pavia, Italy, [2]Univ. of Pavia, Pavia, Italy

Abstract:

Introduction
The ultimate goal of variant interpretation is to identify the DNA variant(s) causing a genetic disease in a given patient. Genetic diagnosis is pivotal for treatment of Rare Diseases, but Rare Disease diagnostic yield still varies from 35% to 55% depending on the disorder, leaving about 200,000 patients without a diagnosis. In this context, we implemented an Artificial Intelligence (AI) approach to prioritize patient's variants and support genetic diagnosis. This solution already resulted as a top performing predictor in the CAGI (Critical Assessment of Genome Interpretation) 6 RGP challenge, led by the Broad Institute of MIT and Harvard. Here, we show the new features of our approach and we extensively report its validation on real cases from different collaborations.
Materials and Methods
We have implemented an AI system that prioritizes variants based on 1) their pathogenicity, 2) the phenotypic similarity between genes and patient's phenotypes encoded in Human Phenotype Ontology terms, and 3) inheritance information. This framework, called "Suggested Diagnosis", is integrated in eVai (www.engenome.com), a SaaS platform for variant interpretation. Additionally, we have equipped the "Suggested Diagnosis" with an Explainability (XAI) layer, thus enabling users to understand the AI reasoning process behind prioritization, and we developed an approach to tag putative Incidental Findings (IF) based on patient's phenotypes and guidelines (PMID:34012068) in a list of prioritized variants. We further validated our framework on different cohorts of a) 85 samples from the "Deciphering Developmental Disorder" (DDD) study (PMID:25533962) b) 19 cases that underwent prenatal screening c) 9 cases with immune deficiency syndromes analyzed within a collaboration with the Clinical Immunogenomics Research Consortium Australasia (CIRCA) and d) on 14 synthetic samples created by adding 14 causative variants from the Rare Genome Project to samples from 1000 Genomes project (1KGP, PMID:31584097).
Results
The Suggested Diagnosis prioritizes the causative variants in the top 5 positions in 76% (65/85) DDD cases and in 79% (15/19) cases from the prenatal screening cohort. In the CIRCA cohort, all the 9 causative samples were prioritized within the top 15th positions, with 4 cases in the first position. In the synthetic samples, the causative variants were all ranked in the top 9th positions, with 6/14 in the first position.
Conclusion
We have validated our AI model to suggest diagnosis on heterogeneous cohorts, and we have made our predictions explainable, thus promoting human trust in AI predictions.

Title: Heterozygous Loss-of-Function Variants in *SMC3*: Lessons for the 'Medium-Hanging Fruit' Era of Mendelian Disease Gene Discovery

Authors:

P. Boone[1], M. Ansari[2], K. Faour[3], A. Shimamura[4], A. Blatnik[5], B. Helm[6], G. Grimes[5], H. Firth[7], A. Breman[8], D. Ben-Isvy[1], L. Wang[1], E. Kao[3], E. Denhoff[3], E. Bijlsma[9], A. Iwata-Otsubo[6], T. DeRavel[10], V. Fusaro[11], A. Fryer[12], K. Nykamp[11], M. Singer-Berk[3], E. Groopman[3], A. O'Donnell-Luria[3], S. Ditroia[13], S. Ballal[3], S. Srivastava[3], S. Banka[14], F. Tuettelmann[15], D. Conrad[16], M. Talkowski[1], D. Fitzpatrick[5]; [1]Massachusetts Gen. Hosp., Boston, MA, [2]Western Gen. Hosp., Edinburgh, United Kingdom, [3]Boston Children's Hosp., Boston, MA, [4]Boston Children's Hosp., Dana-Farber Cancer Inst., Harvard Med. Sch., Boston, MA, [5]Univ. of Edinburgh, Edinburgh, United Kingdom, [6]Indiana Univ. Sch. of Med., Indianapolis, IN, [7]Cambridge Univ. Hosp., Cambridge, United Kingdom, [8]IU Sch. of Med., Indianapolis, IN, [9]Leiden Univ. Med. Ctr., Leiden, Netherlands, [10]UZ Leuven/ Leuven Univ. Hosp., Leuven, Belgium, [11]Invitae, San Francisco, CA, [12]Alder Hey Children's Hosp. Liverpool, Liverpool, United Kingdom, [13]Broad Inst. of MIT and Harvard, Cambridge, MA, [14]Univ. of Manchester, Manchester, United Kingdom, [15]Univ. of Münster, Münster, Germany, [16]Oregon Natl. Primate Res. Ctr./ Oregon Hlth.& Sci. Univ., Beaverton, OR

Abstract:

Heterozygous missense variants and in-frame indels in *SMC3* are a cause of Cornelia de Lange syndrome (CdLS) spectrum phenotypes, marked by intellectual disability, growth deficiency, and dysmorphism, via an apparent dominant-negative mechanism. Yet, *SMC3* loss-of-function variants are of unknown consequence, leading to hypotheses of alternative phenotypes or developmental lethality. We analyzed population databases to characterize mutational intolerance in this gene and used matchmaking servers, patient registries, and other resources to identify individuals with heterozygous, predicted loss-of-function (pLoF) variants in *SMC3*. Here, we show that among population genomic data, *SMC3* behaves as an archetypal haploinsufficient gene - highly constrained against pLoF variants and strongly depleted for missense variants. Among 13 carriers of *SMC3* pLoF variants, phenotypes were variable but coalesced on low growth parameters, developmental delay/intellectual disability, and dysmorphism reminiscent of atypical CdLS. Statistical comparisons to individuals with *SMC3* missense/in-frame indel variants demonstrated a milder presentation in pLoF carriers. Furthermore, several individuals harboring pLoF variants in *SMC3* were non-penetrant for growth, developmental, and/or dysmorphic features, some instead having intriguing symptomatologies with rational biological links to SMC3 including bone marrow failure, acute myeloid leukemia, and Coats retinal vasculopathy. Analyses of transcriptomic and epigenetic data suggest that *SMC3* pLoF variants reduce *SMC3* expression but do not effect a blood DNA methylation signature clustering with that of CdLS, and that the global transcriptional signature of *SMC3* loss is model-dependent. Our finding of substantial population-scale LoF intolerance in concert with a variable/incompletely penetrant phenotype in subjects carrying *SMC3* pLoF variants expands the scope of cohesinopathies, informs on their allelic architecture, and suggests that additional clearly haploinsufficient genes whose disease links were recalcitrant to the first decade of exome sequencing - in which the 'low-hanging fruit' chapter of the Mendelian catalog was completed - will have those links confirmed only by multi-method genomic analyses paired with careful phenotyping.

Title: Driving Diagnostic Advancements in Rare Diseases: Genomic Answers for More Kids

Authors:

A. Cohen[1,2,3], E. Farrow[1,3,4], J. Means[1], M. Gibson[1], E. Grundberg[1,3,4], I. Thiffault[1,2,3], S. Younger[1,3,4], T. Pastinen[1,3,4]; [1]Genomic Med. Ctr., Children's Mercy Kansas City, Kansas City, MO, [2]Dept. of Pathology and Lab. Med., Children's Mercy Kansas City, Kansas City, MO, [3]UKMC Sch. of Med., Univ. of Missouri-Kansas City, Kansas City, MO, [4]Dept. of Pediatrics, Children's Mercy Kansas City, Kansas City, MO

Abstract:

The Genomic Answers for Kids (GA4K) program aims to shorten the path to diagnosis for all individuals with rare genetic diseases by expanding the diagnostic capabilities of modern genomics, cataloging rare disease genomes and phenotypes in a shareable format, and increasing equitable access to genomic testing. In early 2022, we published data on the initial 1,082 affected individuals which included short-read exomes or genomes (srNGS) for all and PacBio HiFi long-read genomes (HiFi-GS) for a subset of undiagnosed individuals, combined with a multitude of variant calling and machine-learning variant prioritization strategies that altogether increased diagnostic rate by ~10%. Importantly, we also recognized major barriers to genetic testing such as lack of insurance coverage and/or access to specialist clinics. Here we present an update on the GA4K program. Genomic analyses now span over 4,500 affected individuals. Recruitment has expanded, with efforts to engage pediatric primary care providers in underserved, socioeconomically disadvantaged rural areas. Updated machine-learning tools incorporate prioritization of copy number variants and allow more efficient ascertainment of recurrent genes of uncertain significance, which are actively shared through GeneMatcher and other collaborations to support discovery of novel disease genes. In addition, we have developed 276 patient-derived iPSC lines to study candidate variants in different cellular contexts through RNA perturbations via single cell data and Iso-Seq. Importantly, we expanded what was already the largest dataset of long-read rare disease genomes from 472 to 1,326 HiFi-GS. Variants identified by srNGS were recapitulated, indicating that HiFi-GS could be considered a reliable first-tier genomic test. New variant calling strategies have allowed the identification of novel expansions, determination of duplication orientation, and improved detection of inversions. In addition, we have begun exploring methylation signatures in the context of promoter silencing and epigenetic drivers via direct 5-methyl-C detection. All GA4K data is available to researchers through NHGRI AnVIL, totaling 26,500 diverse datasets from over 8,000 individuals. The diagnostic rate continues around 30%, having surpassed over 1500 total diagnoses to date. A move towards HiFi-GS offers the promise to simplify cascade genomic testing and increase laboratory efficiency. The main roadblock to analysis productivity is the lack of software to integrate such comprehensive analyses following technological advancements. GA4K remains committed to exploring novel methods needed to solve the unsolved.

Title: Clinical utility of deep-RNAseq in Mendelian disorder diagnostics

Authors:

S. Zhao[1], J. Sinson[1], S. Li[1], J. Rosenfeld[1], P. Mezthly[1], K. Worley[1], L. Burrage[1], M. W. Hubshman[1], S. Ketkar[1], W. Craigen[1], L. Emrick[1], Undiagnosed Diseases Network, T. Clark[2], Z. Shipony[2], D. Lipson[2], C. Eng[1], B. Lee[1], P. Liu[1]; [1]Baylor Coll. of Med., Houston, TX, [2]Ultima Genomics, Newark, CA

Abstract:

RNAseq on patient-derived tissues/cells has emerged in recent years as a powerful tool to aid the interpretation of genetic variants and the validation of molecular diagnosis. Sequencing the human transcriptome at a depth between 30-100 million reads has been shown to be effective in revealing diagnostic findings associated with abnormal expression or splicing. However, it has not been systematically explored whether sequencing at a higher depth will result in improvements in disease diagnostics. Here, we aim to evaluate the clinical utilities and limitations of sequencing the human transcriptome at ultra-high depths ("Deep-RNASeq").
Using mostly natural sequencing-by-synthesis, a cost-effective sequencing method enabled by Ultima Genomics, we sequenced samples derived from four clinically accessible tissues (blood, fibroblast, lymphocyte, and iPSC) at gradient depths, i.e., from $1X10^8$ up to $1X10^9$ reads per human transcriptome. Unique molecular identifiers (UMIs) were used to benchmark the duplication rate and enable better quantification. Given that expression outlier and splicing outlier analyses are the most common RNA diagnostic methods, we computed the minimum exonic reads and junction reads required to enable sufficient statistical power for the outlier analyses. Based on the gradient RNAseq data, we determined the depth-power curve for different clinically accessible tissues. We demonstrate the advantage of deep RNAseq in addressing a key limitation of using transcriptome analysis for clinical diagnostic interpretations -- tissue-specific gene expression, i.e., not all disease-associated genes and isoforms are adequately expressed in clinically accessible tissues. For example, the proportion of Mendelian disease genes (OMIM genes) covered with > 50 read counts increased by 20% when sequenced at $1X10^9$ reads versus $1X10^8$ reads. In addition, we detected abundant isoforms and alternative splicing variations that are useful for both clinical interpretation and functional studies. We then performed ultra-deep RNAseq on a series of clinical cases and identified potential diagnostic splicing variants in *ITPR1*, *VARS2*, and *POLR3A*.
In conclusion, our study provided an evidence-based analysis to benchmark the clinical diagnostic performance of RNAseq at ultra-high depth. Ultra-deep RNAseq successfully revealed previously unrecognized isoforms and alternative splicing patterns that would not be well-represented at the current industry standard sequencing depth.

Title: Harnessing genotype and phenotype data from 3 million patients referred for clinical genetic testing for scalable, Bayesian variant interpretation

Authors:

T. Manders, Y. Kobayashi, A. Wahl, B. Eaves, A. Colavin; Invitae, San Francisco, CA

Abstract:

The continued expansion of genetic sequencing into more areas of clinical medicine - catalyzed by reduced costs and broadened clinical guidelines - has resulted in an explosion in the number of rare variants observed through testing for hereditary disease. Despite a commensurate increase in the amount of available data with potential use for interpreting these variants, approximately half of identified variants in hereditary testing are classified as Variants of Uncertain Significance (VUS). Patient phenotypes and family histories represent categories of these data with tantalizingly large potential to reduce VUS, but their effective utilization remains a manual process, and systematic, computational application of this evidence has proved challenging.

We have designed a scalable model that utilizes natural language processing (NLP) and Bayesian inference to predict variant pathogenicity from clinical phenotype data reported by ordering providers in test requisition forms. First, our model leverages an NLP classifier to learn features of the indication and family history fields that are predictive of a molecular diagnosis of a given condition. Next, a classifier combines the features and demographic information for each patient to assign a PatientScore - the probability that a given patient is affected with the condition. Finally, a generative, hierarchical Bayesian inference model is fit to the PatientScores and available variant pathogenicity labels. Sampling the posterior predictive distribution then yields a VariantScore - the probability that a given variant is pathogenic for the associated condition. Thus our model accounts for both uncertainty in the affected status of each patient observation and uncertainty due to the number of observations for each variant.

We applied this modeling strategy to Invitae's genotype/phenotype database composed of over 3 million patients referred for clinical genetic testing for hereditary conditions and 2 million classified genetic variants. Our approach yielded over 500 genes associated with a range of disease areas with high performance (AUROC > 0.8) as evaluated on a holdout set of labeled variants in each gene. Over 20,000 VUS across these genes and conditions received high confidence probabilistic predictions (VariantScore > 0.95 or < 0.05) that could result in new evidence for interpretation.

Our approach demonstrates the utility of a Bayesian model to effectively integrate genotype and phenotype data, with significant potential to reduce VUS, accelerate genetic diagnoses, and improve treatment personalization for patients and providers.

Session 009: Genomics in the neonatal intensive care unit

Location: Conv Ctr/Room 145A/Level 1

Session Time: Thursday, November 2, 2023, 8:30 am - 10:00 am

Title: Nationwide implementation of rapid genomic diagnosis of sick newborn infants in Japan

Authors:

T. Takenouchi, C. Nishida, H. Suzuki, D. Nakato, M. Yamada, F. Miya, K. Kosaki; Keio Univ. Sch. of Med., Tokyo, Japan

Abstract:

Background: Approximately 5-10% of sick newborn infants admitted to neonatal intensive care units have genetic disorders. An increase in the number of treatable genetic conditions call for rapid genetic diagnosis in this high-risk population. We aimed to establish a nationwide genome diagnostic network for sick newborn infants by using online genetic counseling connecting between the central genome analysis facility and remote neonatal intensive care units. **Methods:** Step 1: between April 2019 and January 2022, a pilot study of whole genome sequencing targeting sick newborn infants was conducted in 18 level 3-4 neonatal intensive care units with full-time neonatologists and medical geneticists. Step 2: between February 2022 and February 2023, the research network for rapid genome diagnosis of sick newborn infants was made available to all levels of ~500 neonatal intensive care units in Japan. The hospitals that do not have capacity of medical genetics were offered online genetic counseling by board-certified medical geneticists. **Results:** Within 12 months, the diagnostic network was expanded from 18 centers in 8 prefectures (Step 1) to 99 centers in 37 prefectures out of all 47 prefectures (Step 2). In STEP2, 38% of cases were registered through online genetic counseling from the central genome analysis facility. The overall diagnostic rates remained unchanged between Step1, i.e., 67/140 cases = 48%, and Step 2, i.e., 44% = 72/164 cases. Currently, the diagnostic network covers approximately one-thirds of all beds (~3,000) in neonatal intensive care units in Japan. **Conclusion:** A rapid expansion of genome diagnostic network within a short period of time reflected unmet and increasing needs for expedited genetic diagnosis in neonatal intensive care units. By using online genetic counseling, we have successfully established a nationwide diagnostic network to cover substantial portion of neonatal intensive care units. Further research is needed to improve turn-around-time and diagnostic rates.

Title: The Baby Bambi study: the first national rapid trio genome sequencing project in neonatal intensive care units

Authors:

D. Marom[1,2], A. Mory[3], S. Reytan[1], A. Yam[1,2], A. Kurolap[3], J. Grinshpun-Cohen[4], A. Singer[4], H. Baris Feldman[5,2], The Israeli NICU-Genomics consortium; [1]Genetics Inst. and Genomic Ctr., Tel Aviv Sourasky Med. Ctr., Tel Aviv, Israel, [2]Sackler Faculty of Med., Tel Aviv Univ., Tel Aviv, Israel, [3]The Genetics Inst. and Genomics Ctr., Tel Aviv Sourasky Med. Ctr., Tel Aviv, Israel, [4]Community Genetics Dept., Publ. Hlth.Services, Ministry of Hlth., Ramat Gan, Israel, [5]The Genetics Inst. and Genomic Ctr., Tel Aviv Sourasky Med. Ctr., Tel Aviv, Israel

Abstract:

Background: Numerous cohort studies have shown a ~36% diagnostic yield for genomic testing in critically-ill neonates. Average turnaround times (TAT) vary between 0.8-60 days. We report the results of the first national project testing the feasibility, diagnostic efficacy and clinical utility of rapid trio genome sequencing (rtGS) in neonatal intensive care units (NICU) throughout Israel.**Methods**: A prospective collaborative pilot between the Community Genetics Department in the Israeli Ministry of Health, all Israeli Medical Genetics Institutes (n=18) and NICUs (n=25) included rapid-trio genome sequencing (rtGS) performed at the Tel-Aviv Sourasky Medical Center Genomics Center in neonates fulfilling inclusion criteria. TAT of a rapid report was expected within 10 days. Clinical utility was assessed via questionnaires circulated to treating neonatologists. **Results**: Between October 2021 to December 2022, 130 neonates (70 males) underwent rtGS. Mean age at enrolment was 12±13 days. Diagnostic efficacy was 50% (65/130) for disease-causing variants, 11% (14/130) variants of unknown significance suspected to be causative, and one novel gene candidate (1%). Disease-causing variants included 12 chromosomal, 54 monogenic disorders and one with uniparental disomy. Mean TAT for rapid report was 7±3 days. Secondary analysis increased diagnostic yield by 2%. Overall response rate for clinical utility questionnaires was 82% (107/130). Genomic testing led to change in medical management in 22% (24/107) of cases, as reported by the treating neonatologists. Results led to immediate precision medicine in 9% (6/66) diagnosed infants; additional two (3%) were transferred to nursing homes, and three (4.6%) received palliative care. **Conclusions**: National rapid genome sequencing in critically ill neonates is feasible and diagnostically beneficial in a public healthcare setting.

Title: The impact of a diagnostic genetic evaluation in the Neonatal Intensive Care Unit: A qualitative analysis

Authors:

S. Abouhala, M. H. Wojcik, M. Del Rosario, I. A. Holm; Boston Children's Hosp., Boston, MA

Abstract:

Background: Many rare genetic disorders manifest within the first 28 days of infant life and result in admission to the neonatal intensive care unit (NICU), where prior research suggests that parents may experience high levels of stress, trauma, and uncertainty. However, further in-depth research into outcomes for these infants and their families is lacking. We therefore investigated parent-reported outcomes for infants with suspected genetic disorders, with a particular focus on quality of life, psychosocial outcomes, and unmet needs within the current system of care.
Methods: We conducted a mixed-methods study utilizing an exploratory sequential approach, in which parents whose infants received genetic and/or metabolic evaluation in the NICU were enrolled for a quantitative analysis of parental values and multidimensional outcomes for the infants. In order to further explore the factors that drive parent-child resiliency post-NICU discharge, we subsequently performed a qualitative investigation in a subset of 18 families identified by purposive sampling. Iterative thematic analysis of interview transcripts was performed in an inductive approach informed by grounded theory.
Results: We identified four main themes: (1) Rare Disease as “Culture Shock,” (2) Parental Trauma & Stressors, (3) Family Resiliency, and (4) Hospital System Recommendations. Analyzing these findings using an adapted version of the socio-ecological model, we found that the experiences of NICU parents can be understood and contextualized through observing parental stress at the individual, interpersonal, community, and public policy levels. Although early, rapid, and broad genomic testing was appreciated and alleviated some degree of parental stress, we found that the majority of parents whose infant remained undiagnosed even despite such testing reported decreasing interest in finding a diagnosis for their child over time, and instead began to adapt their expectations, form more practical goals, and find pride in their child’s medical and social resilience.
Conclusion: These data further enhance our understanding of outcomes of infants with rare disease post-NICU discharge and inform potential future interventions that may prevent or remedy adverse parental mental and psychosocial health outcomes. By applying these findings to the NICU setting, parental, neonatal, psychological, and social service providers may better meet the needs of families experiencing complex medical circumstances within the realm of genomic medicine.

Title: The Utah NeoSeq Project: rapid diagnostic whole genome sequencing of critically ill newborns

Authors:

S. Boyden[1], B. Moore[1], T. Nicholas[1], A. Farrell[1], P. Bayrak-Toydemir[2], R. Mao[2], L. Pizzo[2], T. Wen[2], J. O'Shea[2], R. Lewis[2], M. Fulmer[2], J. Zhao[2], N. Al-Sweel[2], C. Solorzano[1], R. Palmquist[1], E. Hernandez[1], S. Rynearson[1], C. Holt[1], D. Pattison[2], M. Deshotel[2], H. Best[2], L. Brunelli[1], A. Quinlan[1], G. Marth[1], M. Yandell[1], M. Tristani-Firouzi[1], J. Bonkowsky[1], J. Carey[1], B. Shayota[1], S. Malone Jenkins[1]; [1]Univ. of Utah, Salt Lake City, UT, [2]ARUP Lab., Salt Lake City, UT

Abstract:

The NeoSeq project is a collaborative effort between the University of Utah Hospital, the Utah Center for Genetic Discovery (UCGD), and ARUP Laboratories to diagnose newborns in the neonatal intensive care unit (NICU) suspected of having a genetic condition. Patient families are enrolled either prenatally or postnatally and rapid whole genome sequencing (WGS) is performed in parallel with the clinical standard of care. Comprehensive analysis to identify and prioritize variants is conducted jointly by UCGD and ARUP. Results are discussed in a multidisciplinary board review meeting, where candidate variants are evaluated for diagnostic potential, classified according to ACMG criteria, and selected to be reported to the family. Average turnaround time from consent to generation of a results letter is 5-6 days. The program is performed on a research basis, which allows for flexible and evolving analysis methodology and the ability to detect and report all variant types, including single nucleotide variants (SNVs), indels, structural variants (SVs), repeat expansions, and chromosomal abnormalities. When possible, research diagnostic variants are validated by a CLIA-certified testing lab. To date, we have enrolled 66 patients, of whom 24 (36%) were diagnosed and 10 (15%) had a compelling candidate or likely diagnosis, for an overall success rate of 51%. Selected negative cases are further pursued with long read WGS, and our positive cases include three that were solved upon yearly reanalysis and one that was diagnosed with the assistance of RNA sequencing. Each of the 34 confirmed or probable diagnoses involves a different gene or chromosomal locus, highlighting the genetic heterogeneity among NICU patients. In two patients a diagnostic variant was inherited from a parent initially ascertained as unaffected; in one case the parent was found to be affected upon re-examination and in the other case the variant was likely mosaic in the parent. One patient had a mixed compound heterozygote with an SNV and a deletion inherited in trans. Additional diagnostic SVs, identified by several methods, have included three interstitial deletions, a terminal deletion, an unbalanced translocation, and a complex inversion flanked by small deletions that was not reportable by a clinical lab performing WGS-based testing. Our findings underscore the importance of using multiple complementary approaches for variant detection and an expansive analytical strategy to maximize diagnostic yield.

Title: Phenome-wide association studies in the neonatal intensive care population identify phenotypic features that correlate with genetic diagnosis

Authors:

B. Schuler[1], E. McArthur[2], J. M. Sucre[1], D. Ruderfer[1], L. Bastarache[1]; [1]Vanderbilt Univ. Med. Ctr., Nashville, TN, [2]Vanderbilt Univ., Nashville, TN

Abstract:

Introduction: Genetic disorders affect up to 50% of patients admitted to the neonatal intensive care unit (NICU) and contribute substantially to their morbidity and mortality. Making a genetic diagnosis requires strong genotype-phenotype associations to prompt clinical genetic testing and interpret sequence variants. While there are resources that curate genotype-phenotype associations, there are fewer curated perinatal phenotypes for many genetic diagnoses. Large observational cohorts raise the possibility of producing a new type of clinical description, one that leverages the tremendous volume of data generated during a NICU hospitalization. More complete curation of perinatal phenotypes that are sensitive to clinical context may be used to identify neonates in need of genetic testing. Methods: We performed a retrospective analysis of electronic health record (EHR) features from over 22,000 individuals that have received care in our NICU between 2005-2021. This "NICU Database" is coupled with the Clinical Genetics Database (CGdb) which consists of clinical genetic diagnoses for more than 6,000 individuals. Over 1,300 individuals overlap between the two databases allowing us to associate genetic diagnoses with EHR phenotypes collected in the NICU setting. We performed phenome-wide association studies (PheWAS) to identify features in the NICU Database that were associated with any genetic diagnosis in the CGdb. We tested 160 EHR features-including birth defects, surgeries, maternal factors, laboratory values, and clinical interventions-to determine if each feature was associated with any genetic diagnosis. Results: PheWAS revealed 60 NICU EHR features associated with genetic diagnosis including congenital heart disease (odds ratio [OR]=5.8, p=3.2e-176), presence of tracheostomy or tracheotomy (OR=5.6, p=5.6e-13), cleft palate (OR=4.5, p=6.5e-22) and requiring oxygen at 36 weeks gestation (OR=3.5, p=2.2e-22). PheWAS additionally identified NICU EHR features that made a genetic diagnosis less likely including maternal drug use (OR=0.60, p=1.3e-7), chorioamnionitis (OR=0.46, p=0.005), and multiple births in the pregnancy (OR=0.40, p=4.8e-8). Conclusions: Investigation of EHR features in the NICU Database revealed phenotypes that were positively correlated with genetic diagnoses as well as features that made an underlying genetic diagnosis less likely to be present. Comprehensive quantification of EHR-features that support or refute genetic diagnosis can direct genetic testing strategies in the NICU population. Future work will attempt to leverage these to data to aid in earlier genetic diagnoses in this population.

Title: A pilot prospective study of BeginNGS, an artificial intelligence-enabled genome sequencing system for newborn screening of 412 childhood genetic diseases

Authors:

K. Ellsworth[1,2], M. Wright[1,2], L. Olsen[1,2], M. W. Baker[1,3], S. Caylor[1,2], T. Defay[4], A. Feigenbaum[1,2,5], L. Guidugli[1,2], S. Khan[1,2], Y. H. Kwon[1,2], A. Lam[1,2], J. Le[1,2], J. Lenberg[1,2], L. Madhavrao[1,2], R. Mardach[1,2,5], D. Oh[1,2], L. Protopsaltis[1,2], R. Reimers[1,2], M. Saad[1], E. Sanford[1], G. Scharer[1], J. Schleit[1,2], B. Schultz[1,2], L. D. Smith[1], M. Tokita[1,2], L. Van Der Kraan[1,2], K. Wigby[1,2,5], M. J. Willis[1], M. Yandell[1,6], C. A. Hobbs[1,2], S. F. Kingsmore[1,2]; [1]Rady Children's Inst. for Genomic Med., San Diego, CA, [2]Rady Children's Hosp., San Diego, CA, [3]Wisconsin State Lab. of Hygiene, Ctr. for Human Genomics and Precision Med., Dept. of Pediatrics, Univ. of Wisconsin Sch. of Med. and Publ. Hlth., Madison, WI, [4]Alexion, Astra Zeneca Rare Disease, Boston, MA, [5]Dept. of Pediatrics, Univ. of California San Diego, San Diego, CA, [6]Fabric Genomics, Oakland, CA

Abstract:

Newborn screening (NBS) dramatically improves outcomes in selected, severe, childhood disorders by identification and treatment at or before symptom onset. Expansion of the NBS Recommended Uniform Screening Panel (RUSP) has lagged identification and approval of effective therapeutic interventions leading to delayed diagnosis and suboptimal outcomes in several hundred severe, early childhood onset, single locus (mendelian) genetic diseases. Begin Newborn Genomic Screening (BeginNGS) is one of several programs exploring the use of genome sequencing (GS) to supplement RUSP-NBS for these diseases. BeginNGS is unique with respect to the scope of NBS (currently 412 genetic diseases) and use of automated variant identification and virtual guidance regarding confirmatory testing, specialist referral, and urgent management. While clinical implementation is incomplete, here we report interim results of a first prospective clinical trial of BeginNGS. BeginNGS-1 is a scoping study to evaluate current performance and inform the design of a future, fully powered clinical study. Eligibility is limited to neonates aged 1 to 10 days who are admitted to the level IV NICU at Rady Children's Hospital, San Diego, and who are not suspected of having a genetic disease. Patients whose parents or guardians provide informed consent are enrolled. BeginNGS-1 enrollees receive RUSP-NBS and rapid, clinical diagnostic GS, which is performed immediately from dried blood spots with return of results to neonatologists or primary care pediatricians. The BeginNGS screen will be performed after completion of enrollment by automated re-analysis of GS with a pre-qualified set of ~50,000 variants and a modified GEM tool (Fabric Genomics). Of 97 newborns screened thus far for eligibility, 65 (67%) were eligible for enrollment. 13 (20%) of 65 were discharged prior to enrollment, 27 (42%) of 65 declined enrollment, and 25 (38%) of 65 were enrolled. To date, diagnostic GS identified reportable findings in 5 (28%) of 18 newborns that may have clinical significance: one patient with a likely pathogenic variant and four patients with variants of uncertain significance. Median age at enrollment was 8 days and time to result was 10 days. Comparison of results of diagnostic GS, BeginNGS, RUSP-NBS, and confirmatory testing in the first 50 enrollees will be presented, together with implications for ongoing BeginNGS development and design of future clinical trials.

Session 010: Multi-ancestry methods: This is the way

Location: Conv Ctr/Ballroom C/Level 3

Session Time: Thursday, November 2, 2023, 8:30 am - 10:00 am

Title: MultiSuSiE improves multi-ancestry fine-mapping in All of Us whole-genome sequencing data.

Authors:

J. Rossen[1], H. Shi[2], B. J. Strober[3], M. J. Zhang[4], M. Kanai[5], Z. R. McCaw[6], L. Liang[7], O. Weissbrod[8], A. L. Price[9]; [1]Harvard Sch. of Publ. Hlth., Boston, MA, [2]Genentech, San Bruno, CA, [3]Harvard Sch. of Publ. Hlth., Brookline, MA, [4]Harvard Univ., Roxbury Crossing, MA, [5]Broad Inst. of MIT and Harvard, Cambridge, MA, [6]Insitro, South San Francisco, CA, [7]Harvard Sch Publ. Hlth., Boston, MA, [8]Harvard T. H. Chan Sch. of Publ. Hlth., Boston, MA, [9]Harvard Sch Pub Hlth, Boston, MA

Abstract:

Leveraging data from multiple ancestries can greatly improve fine-mapping power due to differences in linkage disequilibrium and allele frequency. However, existing multi-ancestry fine-mapping methods have difficulty efficiently searching the space of potential configurations of causal variants, limiting power. In addition, lower imputation accuracy in African ancestry individuals can compromise both calibration and power.

We developed MultiSuSiE, an extension of the sum of single effects (SuSiE) model to multiple populations. SuSiE is a powerful, versatile, and fast approach to fine-mapping causal variants in a single population (Wang et al. 2020 JRSSB). SuSiE sums across multiple single effect models (each with a single causal variant), fitting and residualizing phenotypes for each single effect model in turn. In MultiSuSiE, each single effect model still assumes a single causal variant, but effect sizes are allowed to vary across populations via a multivariate normal prior informed by cross-population genetic correlations; MultiSuSiE fits and residualizes phenotypes for population-specific effect sizes of each single effect model in turn. Like SuSiE, MultiSuSiE accommodates single or multiple causal variant fine-mapping and can be applied to either individual-level data or GWAS summary statistics with in-sample LD. In simulations with real genotypes, MultiSuSiE maintains correct calibration and identifies more true causal variants than single-population SuSiE and two alternative multi-ancestry fine-mapping methods.

We applied MultiSuSiE to the All of Us cohort, analyzing whole genome-sequencing data for 47,000 African and 47,000 European ancestry individuals (19 million SNPs with MAF > 1% in at least one population) and 5 quantitative traits; we initially focused on single causal variant fine-mapping. MultiSuSiE identified 30% more high-confidence causal variants than SuSiE applied to 94,000 Europeans from All of Us (134 vs. 103 variants with PIP > 0.5), consistent with our simulations. Functional enrichment of fine-mapped causal variants across a broad set of functional annotations was similar for the two methods, validating these results. Interestingly, the improvement of MultiSuSiE over SuSiE was not recapitulated in analyses of 16 traits from the UK Biobank with imputed genotypes, likely due to differences in imputation quality between European and African samples. In conclusion, our findings demonstrate higher power for MultiSuSiE and we advocate for using whole-genome sequencing data over imputed genotypes for fine-mapping due to inaccurate imputation in some ancestries, particularly in African populations.

Title: PIPSORT: Multi-ancestry fine-mapping to identify shared vs. ancestry-specific causal variants

Authors:

T. Mirmira, M. Gymrek; Univ. of California San Diego, La Jolla, CA

Abstract:

Genome-wide association studies (GWAS) quantify associations between variants and traits. To identify possible causal variants for a trait, fine-mapping methods are applied to the GWAS results. Due to linkage disequilibrium (LD) patterns, variants in strong LD with the true causal variant often have strong associations with the trait. Statistical fine-mapping aims to overcome this challenge, but still struggles to distinguish between variants in high LD. However, LD patterns tend to vary across populations, enabling multi-study fine-mapping methods, such as MsCaviar and SuSiEx, to break down LD patterns and improve the resolution of causal variant identification. Existing tools assume that causal variants are shared across studies. We introduce PIPSORT, a multi-ancestry fine-mapping method that removes this assumption. PIPSORT leverages information from multiple studies while accounting for potential differences between them. We build on the Bayesian framework from MsCaviar and make two key modifications: (1) we allow the set of variants in each study to differ and (2) we explore a larger search space of possible configurations to allow the causal variants to differ across studies. With these modifications, we can compute study-specific posterior inclusion probabilities, which help identify ancestry-specific variants while allowing causal variants to be shared.

PIPSORT can be used to classify signals into two categories: (1) *shared* signals are present and causal (possibly with different effect sizes) across two studies and (2) *ancestry-specific* signals are present and causal in one study but monomorphic, rare or not present and therefore not causal in the other. Existing multi-study fine-mapping tools can only identify *shared* signals.

Using a cohort of European (EUR) samples and a cohort of African (AFR) samples from the UK Biobank, we simulate phenotypes for different categories of causal signals to evaluate PIPSORT. The simulations test a range of parameters (e.g. allele frequencies, LD, and effect size) and identify situations under which fine-mapping methods are likely to struggle. PIPSORT achieves similar results compared to MsCaviar for *shared* signals but only PIPSORT is able to detect the difference in the underlying causality for *ancestry-specific* simulations. Using the same cohorts, we evaluate PIPSORT with real phenotypes (LDL) from the UK Biobank on a set of 403 loci. We find that 12.2% of fine-mapped variants are *ancestry-specific*, of which 43.5% are AFR-specific and 57.5% are EUR-specific. Our work develops a general multi-study fine-mapping tool that uses cross-study information without requiring causal variants to be shared.

Title: Multi-ancestry gene expression prediction identifies novel genes regulating complex traits specifically in individuals of African descent.

Authors:

K. Akamatsu[1], T. Amariuta[2,3]; [1]Sch. of Biological Sci. Undergraduate Program, UC San Diego, San Diego, CA, [2]Dept. of Med., Div. of BioMed. Informatics, UC San Diego, San Diego, CA, [3]Halıcıoğlu Data Sci. Inst., UC San Diego, San Diego, CA

Abstract:

Genetic models of gene expression can implicate mechanisms of disease-associated variants from GWAS. While genetic effects on gene expression have been extensively characterized by eQTL studies, data from non-European populations is limited. This restricts our understanding of disease to genes whose regulatory variants are common in European populations. Previous work has leveraged data from multiple populations to improve GWAS power (Okada 2014 *Nature*) and polygenic risk score accuracy (Marquez-Luna 2017 *Genet Epi*). However, multi-ancestry data has not yet been leveraged to enable powerful genetic models of gene expression.

We developed MAGEPRO (Multi-Ancestry Gene Expression PRediction Optimization), a method that learns powerful gene expression prediction models for non-European populations using eQTL summary statistics across diverse ancestries. MAGEPRO takes a two-step approach. First, it leverages limited individual-level genotype and gene expression data in a target non-European population and learns a noisy estimate of SNP-gene effect sizes in a lasso-regularized linear regression. Second, MAGEPRO estimates posterior SNP-gene effect sizes by finding an optimal linear combination of eQTL summary statistics from independent cohorts (including the noisy estimate from the target population) in a ridge-regularized linear regression.

We applied MAGEPRO to gene expression and genotype data for 80 African American individuals in GTEx across 19,601 genes expressed in whole blood. We acquired 10 eQTL summary statistics across European, African, East Asian, American, and Hispanic populations. MAGEPRO improved the cross-validation accuracy of African American gene expression models by 23% relative to the single population model ($p < 1.2x10^{-84}$) and by 37% relative to sample-size weighted meta-analysis across all eQTL datasets ($p < 1.2x10^{-49}$); notably, for 2,162 genes, our method increased the accuracy by over 100%.

We then used the MAGEPRO gene expression models to perform a transcriptome-wide association study (TWAS) for each of 11 complex traits in African individuals from the pan UK Biobank (average N = 6,295). MAGEPRO identified 27 trait-associated genes at 5% FDR that were not found by the single population model, 19 of which were not significant in European TWAS (FDR > 5%). These include two genes (*DNAAF5* and *YWHAG*) negatively associated with BMI in African individuals and have been linked to obesity (Davenport 2007 *Curr Biol*) and insulin secretion (Mugabo 2022 *JCI Insight*) in mice. The discovery of ancestry-specific regulation of complex traits such as BMI provides insight into disease biology and differences in global prevalence.

Title: GWAS Meta-Analysis of Admixed Populations (GMAX) uses local ancestry inference to identify associated loci in GSCAN meta-analysis.

Authors:

N. Benjamin[1], GWAS & Sequencing Consortium of Alcohol and Nicotine use(GSCAN), X. Wang[2]; [1]Pennsylvania State Univ., Hershey, PA, [2]Penn State Coll. of Med., Hershey, PA

Abstract:

Admixed populations possess ancestry from multiple continental source groups, resulting in the unique mosaic genome structure from distinct continental ancestries. Hence, it is important to properly analyze admixed population genomes, including heterogeneity in effect sizes and linkage disequilibrium structure. Existed methods, for example TRACTOR, have already shown that incorporating local ancestry information in genome wide association studies (GWAS) can increase the power of discovering variant-trait associations, especially for admixed populations. Despite this fact, there are no current methods that have incorporated local ancestry information in meta-analysis.Here, we develop a method, GMAX Local Ancestry Inference (GMAX-LAI), to cooperate local ancestry across the genome of admixed individuals for GWAS meta-analysis. We first estimate ancestry proportions at a given variant for admixed study by decomposing allele frequencies as a weighted sum of allele frequencies of continental ancestries. By comparing with RFMix, a commonly used individual level LAI method, our approach provides comparable estimation. These ancestral estimates are later incorporated in our mixed effect meta-regression model to model genetic effects in our meta-analysis.We apply our method to GSCAN (GWAS & Sequencing Consortium of Alcohol and Nicotine use) smoking and drinking traits with a diverse ancestry background (55% European, 15% African American, 6% of Latino/Hispanic American, 24% of East Asian.) For African American studies, the proportions range from 68%-74% for African ancestry and 26% -32% for European ancestry. While for Latino/Hispanic studies, the estimated average compositions are 69%-72% European, 6%-10% African and 19%-24% Native American. We also observe significant ancestry proportion difference across studies, reflecting substantial study-specific local ancestry genetic structure. By meta-analyzing 121 studies, our method identifies 205 loci associated with the ‘Drinks per Week’(DrnkWk) trait and 20 loci associated with the ‘Age at Smoking Initiation’ (AgeSmk) trait. Compared to a fixed effect model, which assumes effect sizes are the same across all studies, 4 loci (mapped to *ADAMTS9-AS2*, *RNU6-679P*, *RPL3P7* and *AFG1L*), and 2 loci (mapped to *PCCA-DT* and *SPIRE2*) are novel to our model, for DrnkWk and AgeSmk, respectively.Overall, our model highlights the benefits of including local ancestry information for admixed individuals under a GWAS meta-analysis setting. The application of our method to GSCAN provides a significant step forward in understanding the genetic architecture of tobacco and alcohol use in admixed populations.

Title: Pan-cancer and cross-population genome-wide association studies dissect shared genetic backgrounds underlying carcinogenesis.

Authors:

G. Sato[1], Y. Shirai[1,2], S. Namba[1], R. Edahiro[1], K. Sonehara[1,3,4], T. Hata[1], M. Uemura[1], K. Matsuda[5], Y. Doki[1], H. Eguchi[1], Y. Okada[1,2,3,4]; [1]Osaka Univ. Graduate Sch. of Med., Suita, Japan, [2]Immunology Frontier Res. Ctr. (WPI-IFReC), Osaka Univ., Suita, Japan, [3]Graduate Sch. of Med., The Univ. of Tokyo, Tokyo, Japan, [4]RIKEN Ctr. for Integrative Med. Sci., Yokohama, Japan, [5]Graduate Sch. of Frontier Sci., The Univ. of Tokyo, Tokyo, Japan

Abstract:

As well as environmental and somatic factors, hereditary components play an important role in cancer development. Integrating genomic data of multiple cancers allows *de novo* cancer grouping and elucidation of the shared genetic basis across cancers. While previous cross-cancer studies have identified pleiotropic associations and genetically correlated cancer pairs, most of them have been undertaken in Europeans. Here, we conducted the pan-cancer and cross-population genome-wide association study (GWAS) meta-analysis and replication studies on 13 cancers including 61,465 cancer cases and 188,550 controls from East Asians (Biobank Japan) and 43,098 cancer cases and 334,343 controls from Europeans (UK Biobank). We identified five risk variants of the specific individual cancers (e.g., rs2076295 at *DSP* on 6p24 associated with lung cancer) newly satisfying the genome-wide significance threshold and five novel pleiotropic variants (e.g., rs2525548 at *TRIM4* on 7q22 nominally associated with six cancers), further investigated using colocalization with gene expression for the putative target genes. Quantifying shared heritability among the cancers detected positive genetic correlations between breast and prostate cancer across East Asian and European ancestries, as further validated in a Finnish biobank. Shared genetic components between breast and prostate cancer increased the statistical power, and the large-scale meta-analysis and replication studies of 277,896 cases and 901,858 controls including the largest-to-date GWAS of said cancers identified 91 newly genome-wide significant loci. Using the summary statistics of the large-scale meta-analysis, we performed the enrichment analysis of pathways and cell-types to acquire further insights into shared genetic backgrounds between the two cancers. The pathway enrichment analysis revealed six common pathways including apoptosis and sexual hormone responses. Using single-cell RNA-seq datasets, the cell type-specific analysis showed that the polygenic risk of breast and prostate cancer was enriched in epithelial cells. Our comprehensive genomic study detected novel cancer risk loci including pleiotropic associations and highlighted the advantages of the pan-cancer approach. Furthermore, our results demonstrate that investigating genetically correlated cancers can enhance our understanding of carcinogenesis.

Title: Population Representation and Sampling in the Human Pangenome Reference

Authors:

R. Shemirani[1], X. Feng[2], L. Song[2], A. Arguello[3], R. M. Cook-Deegan[4], R. Durbin[5], E. E. Eichler[6], A. L. Felsenfeld[3], N. A. Garrison[7], I. M. Hall[8], E. D. Jarvis[9], B. A. Koenig[10], H. A. Lawson[11], B. Paten[12], A. M. Phillippy[3], A. B. Popejoy[13], T. Wang[14], H. Li[2], K. H. Miga[12], E. E. Kenny[1]; [1]Inst. for Genomic Hlth., Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Dept. of Data Sci., Dana-Farber Cancer Inst., Boston, MA, [3]Natl. Human Genome Res. Inst., NIH, Bethesda, MD, [4]Consortium for Sci., Policy, and Outcomes, Arizona State Univ., Washington, DC, [5]Dept. of Genetics, Univ. of Cambridge, Cambridge, United Kingdom, [6]Univ. of Washington, Seattle, WA, [7]Univ. of California, Los Angeles, Los Angeles, CA, [8]Yale Univ., New Haven, CT, [9]Vertebrate Genome Lab., The Rockefeller Univ., New York, NY, [10]Univ. of California, San Francisco, San Francisco, CA, [11]Dept. of Genetics, Washington Univ. Sch. of Med., St. Louis, MO, [12]Univ. of California, Santa Cruz, Santa Cruz, CA, [13]Univ. of California, Davis, Sacramento, CA, [14]Washington Univ. in St. Louis, St. Louis, MO

Abstract:

The human genome reference has been the bedrock of biomedical research and personalized medicine for over 20 years. Yet the current linear reference derives from a single individual at each locus, which can result in a reference bias in variant calling which may limit downstream applications in genomic research and medicine. The goal of the National Human Genome Research Institute-funded Human Pangenome Reference Consortium (HPRC) is to produce a pangenome reference comprising the complete diploid genomes of hundreds of diverse individuals. A better representation of the landscape of human genomic variation in the reference will uncover greater breadth and complexity of variants and lead to increased accuracy of variant calling. However, the process for selecting individuals to best represent human diversity is challenging. There is an inherent need to balance approaches that maximize genetic variation with those that optimize for representation across the human coalescence, in addition to accounting for archaic introgressions and the evolutionary history of complex structural variants. There is also a need to embed strategies that account for the ethical, legal, and social implications raised when recruiting diverse participants to an open data effort. In this work, we describe the processes for selecting the first 350 participants for the human pangenome. We highlight two novel methods using extant genomic data to select individuals; (a) an iterative algorithm for maximizing the representation of known common variation, and (b) an algorithm that uses haplotypes identical-by-descent as an approximation of the ancestry graph of the datasets. For the latter approach, we demonstrate that maximal coverage of genetic divergence through limited sampling can be translated into a set cover problem, which we then solve using an adaptive particle swarm approximate optimization method. We evaluate the degree of genomic diversity represented in the first 261 genomes selected from the 1000 Genomes (1000G) Project, which achieves >99.9% coverage of common variants (MAF>1%) in the out-of-sample 1000G dataset, using data from the Genome Aggregation Database and All of Us Research Project. We also note the limitations of such approaches considering the bias in representation of human diversity in current genomic databases and highlight the need for continuing plans for prospective recruitment, community engagement, and international outreach in HPRC. This work contributes to the establishment of improved standards for how we represent human genomic diversity to build a global reference resource.

Session 011: New advances in genome interpretation and functional studies

Location: Conv Ctr/Room 207A/Level 2

Session Time: Thursday, November 2, 2023, 8:30 am - 10:00 am

Title: The quantitative minigene assay provides supporting evidence for variant classification in molecular diagnostics.

Authors:

C. Bender[1], B. Guan[1], M. Pantrangi[2], N. Moore[1], M. Reeves[1], A. Naik[1], H. Li[1], D. Blain[1], A. Agather[1], L. Huryn[1], C. Cukras[1], K. Goetz[1], R. B. Hufnagel[1]; [1]Natl. Eye Inst./NIH, Bethesda, MD, [2]Columbia Univ. Irving Med. Ctr., New York, NY

Abstract:

Minigene assays permit investigations of the effects of genetic variants on mRNA splicing *in vitro*. The assay plays an important role in variant classification as it has the potential to provide a functional support for variant pathogenicity according to the ACMG/AMP sequence variant interpretation guidelines. Typically, the assay is performed by transfecting DNA constructs containing reference and variant sequences into a cell line, after which the mRNA is analyzed via qualitative measures such as gel electrophoresis, fragment analysis, and Sanger sequencing. These methods are effective in identifying alternatively spliced mRNA products. We hypothesized that some variants may not lead to the use of alternative splice sites or exon skipping, but instead affect mature mRNA expression by altering splicing efficiency. To test this hypothesis, we designed a quantitative minigene assay where quantitative PCR (qPCR) and digital droplet PCR (ddPCR) were used to measure minigene transcript levels. We used a new efficient method to make minigene constructs by employing a gene assembly approach. We then applied our quantitative minigene assay on four noncanonical splice site variants in the *RS1*, *RPE65*, *PRPF8* and *OCA2* genes that were identified from patients with inherited eye diseases. State-of-art *in silico* splicing prediction tools suggest these variants are unlikely to cause splicing defects. Gel-electrophoresis and Sanger sequencing showed that the noncanonical splicing variants produced the same-sized mRNA products as the reference constructs, further suggesting that these four variants did not alter splicing sites. However, quantitative assays demonstrated that they caused markedly diminished levels of mature mRNA in the minigene assay. Upon applying the quantitative assay as a piece of strong supporting evidence for pathogenicity, we were able to reclassify the variants tested to be pathogenic or likely-pathogenic. By embracing these quantitative methodologies, we identified a new class of disease-associated splicing variants that typically evade detection by in silico prediction or conventional qualitative methods either *in vitro* or i*n vivo*. This quantitative assay provides critical functional evidence for classification of splice variants and could reveal many suspected benign splice variants are actually pathogenic.

Title: Rapid generation of transgenic mouse mimicking variant of uncertain significance, VUS, clarifies its pathogenicity

Authors:

S. Hayashi[1], Y. Suzuki[1], D. Fukushi[1], K. Yamada[1], H. Miyahara[2]; [1]Inst. for Dev.al Res., Aichi Dev.al Disability Ctr., Kasugai, Japan, [2]Inst. for Med. Sci. of Aging, Aichi Med. Univ., Nagakute, Japan

Abstract:

Background: In the last two decades, high-throughput genetic analysis has identified a vast number of not only pathogenic variants but also variants of uncertain significance (VUS) in genetic disorders. Even though *in vitro* or *in silico* analyses have provided many evidences to understand VUS, they are sometimes inadequate to explain a correlation between VUS and proband's phenotypes like behavioral problems or organ malformation. **Methods:** To resolve the problems, we have generated mice model mimicking precisely the targeting VUS by the iGONAD method. The method directly induces genome-editing reagents, guide RNA, ssDNA, and CAS9, into fertilized ovum to produce a transgenic mouse only in 19 days, thus it enables rapid assessment of the mouse phenotype. In our strategy, during the back-crossing N1~N2 mice are used for quick screening of VUS: *in vitro* analysis like transcript or protein analysis, and N5 or more are used for *in vivo* precise assessment like behavior analysis or histology of brain. It takes only a few months for the quick screening, and more a few months for the precise assessment. **Result:** Until now we have generated 16 lines of transgenic mice to explain the probands' phenotypes including familial epilepsy with intellectual disability (ID), familial cases of ID and gait disorder, complexed brain malformation, or large joint contracture in a known syndrome, and so on. We have successfully generated the transgenic mice include not only point mutation but also more complexed genomic rearrangement: *e.g.* insertion-deletion, over 10kb-sized deletion, or tag-insertion detectable by western blotting. As an example, we have interpreted VUS at a splice site of *TENM4* identified in a familial case of epilepsy and ID. The transgenic mouse mimicking the VUS certainly reproduced the exon-skipping observed in the probands. The mouse showed increased seizure susceptibility to pentylenetetrazole, primary cultured neurons showed less growth of the neurites, and brain showed microscopic structure abnormalities, *e.g.* decreasing of neural cells and myelinating fibers. **Conclusion**: These findings clarify a pathogenicity of the VUS in *TENM4*, and suggest that our current strategy is useful in understanding VUS in genetic disorders rapidly and precisely.

Title: Splice variants in silent genes resolved with CRISPR activation

Authors:

L. Jolly; Univ. of Adelaide, Adelaide, Australia

Abstract:

Resolution of variants of uncertain significance (VUS) benefits from functional studies to help confirm pathogenicity. For VUS predicted to impact splicing, analysis of gene splicing at the RNA level, ideally studied in the context of the patient genetic and cellular background, is often sufficient to reclassify the variant. Patient blood and/or skin derived cell lines are frequently used for RNA analyses, however, >1400 Mendelian disease genes are refractory to robust splicing assessment from RNAseq data generated from these clinically accessible tissues as they are not sufficiently expressed. To overcome this limitation, we adapted a CRISPR-based gene activation technology to induce the expression of otherwise non-expressed disease genes in patient cell lines. Thus far, we activated the expression of >20 non-expressed disease genes tested, including those causing epilepsy (e.g., *PCDH19*, *SCN1A*), intellectual disability (e.g. *MYT1L*, *PAK3*), blindness (*USH2A*), Osteomyelitis (*IL1RN*) and neuromuscular disorders (*DMD*) among others. Activation levels range from 10 to >20,000-fold. We provide examples where combining gene transactivation in patient cells with short- and long-read RNA sequencing resolved damaging impacts of VUS on mRNA splicing and/or surveillance by nonsense mediated mRNA decay. This study highlights the utility of CRISPR gene activation as a technology able to provide functional evidence of variant pathogenicity in non-expressed Mendelian genes in patient cells derived from clinically accessible tissues . The approach is adaptable to any gene and can be re-used to study different variants within a gene.

Title: Break-induced replication mediated by inverted repeats underlie formation of pathogenic inverted triplications

Authors:

C. Grochowski[1], J. Bengtsson[2], M. Y. Lun[2], M. Gandhi[2], H. Du[1], M. Mehaffey[2], K. Park[2], W. Höps[3], J. Korbel[3], M. Mahmoud[1], L. F. Paulin[1], S. Jhangiani[1], D. Muzny[1], J. Fatih[1], R. Gibbs[1], M. Pendleton[4], E. Harrington[4], S. Juul[4], A. Lindstrand[5], F. Sedlazeck[1], D. Pehlivan[1], J. Lupski[1], C. Carvalho[2]; [1]Baylor Coll. Med., Houston, TX, [2]Pacific Northwest Res. Inst., Seattle, WA, [3]European Molec Biology Lab, Heidelberg, Germany, [4]Oxford Nanopore Technologies Inc, New York, NY, [5]Karolinska Inst.t, Stockholm, Sweden

Abstract:

The human genome is dynamic and has been shaped by various mutational and evolutionary processes including the formation of complex genomic rearrangements. The duplication-triplication/inverted-duplication (DUP-TRP/INV-DUP) conformational structure, a DNA rearrangement end-product of structural variant mutagenesis, is a complex structural variant generated by replicative repair of one-ended double-stranded DNA formed during fork collapse. It is often mediated by a pair of inverted low copy repeats (IP-LCRs) followed by iterative template switches resulting in at least two breakpoint junctions *in cis*. This aberration can result in a disease state largely due to copy-number variants affecting dosage sensitive genes with a frequency up to 26% in some disease cohorts. Although it has been identified as an important signature of pathogenicity to genomic disorders, common diseases and one of the most prevalent aberrations in cancer genomes, its genomic architecture remains unresolved and is predicted to display at least four SV haplotypes. Here we studied the genomic architecture of DUP-TRP/INV-DUP by investigating 25 patients with X-linked neurodevelopmental disorders initially identified by array comparative genomic hybridization (aCGH). We resolved the previously unknown haplotype structure in 18 samples using a combination of short-read genome sequencing (GS), long-read GS (HiFi and ONT), optical genome mapping (OGM) and StrandSeq. Surprisingly, we found evidence for the existence of 4 out of 4 predicted structural variant haplotypes and observed that polymorphic inversions present in the ancestral chromosome generates further haplotype variations. Four different IP-LCRs were identified in this cohort with a size ranging from 917 bp to 140,621 bp, nucleotide similarity from 98.12% to 98.89% and distance apart encompassing 10,767 to 317,810 bp. The multimodal genomics approach refined the strand transfer or point of template switching between one pair of IP-LCRs revealing a DNA segment of ~2.2-2.7 kb of 100% nucleotide similarity that facilitates break-induced replication. A mutagenesis model was developed to infer the specific LCR used to mediate the non-allelic homology repair that aids in haplotype prediction. In aggregate, these data provide experimental evidence supporting our hypothesis that IP-LCRs act as a recombinant substrate in replication-based repair mechanisms. Moreover, we define the molecular features of IP-LCRs that generate genomic instability prone to the formation of copy-number associated inversions.

Title: Sexually dimorphic and cell type specific gene expression signatures of disease progression in Rett syndrome mouse model and human Rett cortices

Authors:

O. Sharifi[1], V. Haghani[2], K. Neier[2], K. Fraga[2], I. Korf[3], S. Hakam[2], N. Johansen[2], G. Quon[2], D. Yasui[3], J. Lasalle[3]; [1]Univ. of California, Davis, davis, CA, [2]Univ. of California, Davis, Davis, CA, [3]Univ. of California Davis Sch. of Med., Davis, CA

Abstract:

Rett syndrome (RTT) is an X-linked dominant neurodevelopment disorder caused by mutations in *MECP2*, encoding methyl CpG binding protein 2. RTT infants appear normal at birth, followed by a regression in cognitive and motor functions at 6-18 months of age. RTT affects females and *MECP2* mutations are rarely seen in males due to their paternal origin or early lethality. Random X-chromosome inactivation in heterozygous RTT females results in a mixture of *MECP2* wild-type and mutant expressing cells, creating a “mosaic” brain. The mosaic cellular heterogeneity and the developmental regression calls for longitudinal gene expression profiling in RTT brain at the single cell level. We engineered and previously published a mouse based on a *MECP2* exon 1 mutation from a RTT patient that ablates production of the MeCP2e1 isoform. This MeCP2e1 deficient mouse model recapitulates specific behavioral, motor, and metabolic defects observed in RTT patients. *Mecp2e1*$^{-/+}$ female mice exhibit mosaicism and progressive disease progression starting at 6-8 weeks, while the *Mecp2e1*$^{-/y}$ male mice exhibit rapid disease progression resulting in premature death by 16-18 weeks. Therefore, we utilized single nuclear RNA sequencing (snRNA-seq) before and across disease progression (E18, P30, P60 for both sexes, then P120 for males and P150 for females) to understand cell type specific defects in mosaic *Mecp2e1*$^{-/+}$ female and *Mecp2e1*$^{-/y}$ male cortex.We used snRNA-seq 5’ to profile cell type specific gene expression patterns based on the Allen Institute’s cell specific markers for mouse cortex. First, we show that the proportion of cell types stays consistent over time for both sexes. Differentially expressed gene (DEG) analysis showed a surprising decrease in DEGs in females over disease progression, in contrast to an expected increase in males. In *Mecp2e1*$^{-/+}$ females, gene pathways of synaptic function and retrograde endocannabinoid signaling are already dysregulated at a pre-symptomatic stage (P30) in inhibitory neurons. Upon symptom onset (P60), these dysregulated gene pathways are next observed in excitatory neurons, followed by astrocytes at the late disease stage (P150) in *Mecp2e1*$^{-/+}$ cortices. In contrast, DEGs identified in *Mecp2e1*$^{-/y}$ cortices do not show significant enrichment in synaptic pathways. Furthermore, a comparison of DEGs and KEGG terms between RTT human and mouse cortices shows a more significant overlap between human RTT and female *Mecp2e1*$^{-/+}$ mouse than male *Mecp2e1*$^{-/+}$ mouse. These results demonstrate profound sexual dimorphism in cortical cell type gene dysregulation consistent with loss of MeCP2e1 underlying RTT disease progression.

Title: Cell-free DNA as a Novel Biomarker for Disease Progression and Response to Treatment in Hutchinson-Gilford Progeria Syndrome.

Authors:

A. Thaivalappil, W. A. Cabral, U. L. Tavarez, M. R. Erdos, F. S. Collins; Natl. Human Genome Res. Inst., Bethesda, MD

Abstract:

Hutchinson-Gilford Progeria Syndrome (HGPS) is a premature aging disorder that affects tissues of mesenchymal origin. Most individuals with HGPS harbor a de novo c.1824C>T (p.G608G) mutation in the gene encoding lamin A (LMNA), which activates a cryptic splice donor site resulting in production of a toxic protein termed “progerin”. Clinical manifestations include growth deficiency, lipodystrophy, cardiovascular defects and bone dysplasia. Currently lonafarnib, a farnesyltransferase inhibitor, is the only FDA-approved treatment for progeria. For development of new therapeutics, a reliable biomarker is needed to demonstrate qualitative or quantitative efficacy of disease progression or treatment response in preclinical or clinical trials. We have developed a novel liquid biopsy approach to characterize phenotypic progression in two HGPS mouse models, as assessed by plasma concentration of cell-free DNA (cfDNA). cfDNA are short circulating DNA fragments released into the bloodstream through cellular breakdown and active DNA release. With a High Sensitivity DNA chip, we observed elevations of cfDNA in heterozygous and homozygous HGPS mice compared to age-matched counterparts, including a 107% increase in cfDNA levels of 12 week homozygotes compared to wild type. Digital droplet PCR (ddPCR) amplification of short and long interspersed retrotransposable elements (LINES/SINES) within the cfDNA sequences correlated with plasma cfDNA concentrations. Additionally, ddPCR provided a secondary validation of cfDNA trends, demonstrating a 119% elevation in homozygote cfDNA compared to wild type at 12 weeks. Quantification of plasma cfDNA via LINE/SINE copy number also greatly improved the sensitivity of the assay, enabling quantification with as little as 5uL of murine plasma. Validation of cfDNA levels as a clinical biomarker for therapeutic response was achieved by demonstrating quantitative reductions in plasma LINE and SINE copy number up to 56% at 8 weeks from mice treated *in vivo* with a locus-specific DNA base editor that corrects the mutation and partially rescues the HGPS phenotype. Thus, plasma cfDNA has the right properties to serve as a reliable biomarker for disease progression and treatment response in mouse models and can now be tested in human trial samples.

Session 012: Novel genetic variations associated with cancer risk and outcomes

Location: Conv Ctr/Room 146B/Level 1

Session Time: Thursday, November 2, 2023, 8:30 am - 10:00 am

Title: Large-scale genome-wide association study of 398,238 women identifies nine novel loci associated with high-grade serous epithelial ovarian cancer risk

Authors:

D. Barnes[1], J. P. Tyrer[1], J. Dennis[1], G. Chenevix-Trench[2], D. F. Easton[1], S. A. Gayther[3], M. R. Jones[3], A. C. Antoniou[1], P. D. P. Pharoah[3], Consortium of Investigators of Modifiers of BRCA1and BRCA2, Ovarian Cancer Association Consortium; [1]Univ. of Cambridge, Cambridge, United Kingdom, [2]QIMR Berghofer, Brisbane, Australia, [3]Cedars Sinai Med. Ctr., Los Angeles, CA

Abstract:

Twenty loci have been associated with high-grade serous (HGS) epithelial ovarian cancer (EOC), the most common and aggressive form of EOC. Almost all EOCs diagnosed in *BRCA1* and *BRCA2* pathogenic variant carriers are HGS. We used genotype and imputed data from the Ovarian Cancer Association Consortium (OCAC), the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA), the UK Biobank and FinnGen to identify novel HGS EOC susceptibility loci and develop enhanced polygenic scores (PGS) for the disease.

We imputed >22 million variants for 152,861 women of European ancestry from OCAC and CIMBA genotyped using the custom OncoArray or iCOGS array to the TOPMed reference panel. Associations were assessed separately by consortium: OCAC (104,887 controls and 15,361 cases), UK Biobank (244,529 controls and 848 cases) and CIMBA (28,939 unaffected and 3,674 affected) and were combined using meta-analysis. OCAC and CIMBA data were used to develop a PGS using the Select and Shrink with Summary Statistics method. The PGS was trained on 150,658 women from FinnGen (149,394 controls and 1,264 cases) and validated in UK Biobank.

Eleven novel independent variants (three with MAF>5%, two with MAF 1-5%, six with MAF<1%) at nine loci were associated with HGS EOC, resulting in 98 credible causal variants. rs78378222, a *TP53* 3' UTR SNP, was strongly associated with HGS EOC (per G allele relative risk (RR)=0.70, 95%CI:0.62-0.78, $P=1.76x10^{-9}$). rs78378222-G has been associated with an increased risk of developing skin, brain and prostate cancers. rs78378222-G has been associated with triple-negative breast cancer, with a near identical protective RR to the HGS EOC association that we estimated.

The PGS included 64,518 variants and was associated with an OR of 1.46 per SD (95%CI:1.37-1.54) in the validation dataset (area under ROC curve=0.607, 95%CI:0.590-0.623).

This study represents the largest GWAS for HGS EOC to date. The results highlight that improvements in imputation reference panels and increased sample sizes can identify HGS EOC associated variants that previously went undetected. Furthermore, they can lead to improved PGS that can be included in cancer risk prediction algorithms to improve personalized risk prediction for HGS EOC.

Title: Multi-ancestry GWAS identifies a novel osteosarcoma susceptibility locus at 12q24.13

Authors:

D. Gianferante[1], B. Gorman[2], S. Cleland[2], M. Yeager[1,3], M. Dean[1], L. G. Spector[4], K. A. Janeway[5], R. Gorlick[6], A. Patiño-Garcia[7], F. Lecanda[8], M. Serra[9,10], C. Hattinger[9,10], K. Scotlandi[9], F. Amary[11], I. L. Andrulis[12], J. S. Wunder[13], M. L. Ballinger[14,15], D. M. Thomas[14,15], P. J. Lupo[16], L. Morton[1], M. M. Hudson[17], G. T. Armstrong[17], S. Bhatia[18], L. E. Egolf[1,3], A. Vogt[1,3], J. Liu[1], B. D. Hicks[1,3], N. D. Freeman[1], W-Y. Huang[1], A. Lori[19], W. Diver[19], S. A. Savage[1], S. J. Chanock[1], L. Mirabello[1]; [1]Div. of Cancer Epidemiology and Genetics, NCI, NIH, Rockville, MD, [2]Booz Allen Hamilton, McLean, VA, [3]Cancer Genomics Res. Lab., Frederick Natl. Lab. for Cancer Res., Fredrick, MD, [4]Dept. of Pediatrics, Univ. of Minnesota, Minneapolis, MN, [5]Dana-Farber/Boston Children's Cancer and Blood Disorders Ctr., Harvard Med. Sch., Boston, MA, [6]Div. of Pediatrics, Univ. of Texas MD Anderson Cancer Ctr., Houston, TX, [7]Dept. of Pediatrics/Med. Genomics Unit and Program in Solid Tumors, Cima-Univ. de Navarra, Cancer Ctr. Clínica Univ. de Navarra (CCUN), Pamplona, Spain, [8]Ctr. for Applied Med. Res. (CIMA)-Univ. of Navarra, IdiSNA, and CIBERONC, Pamplona, Spain, [9]Lab. of Experimental Oncology, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy, [10]IRCCS Istituto Ortopedico Rizzoli, Osteoncology, Bone and Soft Tissue Sarcomas and Innovative Therapies, Pharmacogenomics and Pharmacogenetics Res. Unit, Bologna, Italy, [11]Royal Natl. Orthopaedic Hosp. NHS Trust, Stanmore, Middlesex, United Kingdom, [12]Lunenfeld-Tanenbaum Res. Inst., Sinai Hlth.System, Univ. of Toronto, Toronto, ON, Canada, [13]Litwin Ctr. for Cancer Genetics, Samuel Lunenfeld Res. Inst., Mount Sinai Hosp., Univ. of Toronto, Toronto, ON, Canada, [14]Garvan Inst. of Med. Res., Darlinghurst, Australia, [15]St Vincent's Clinical Sch., Faculty of Med., Univ. of New South Wales, Sydney, Australia, [16]Dept. of Pediatrics, Section of Hematology-Oncology, Baylor Coll. of Med., Houston, TX, [17]Dept. of Epidemiology and Cancer Control, St. Jude Children's Res. Hosp., Memphis, TN, [18]Inst. for Cancer Outcomes and Survivorship, Univ. of Alabama at Birmingham, Birmingham, AL, [19]Dept. of Population Sci., American Cancer Society, Kennesaw, GA

Abstract:

Osteosarcoma (OS) is the most common malignant bone tumor in children and adolescents, and there are few established risk factors for sporadic OS outside of rare germline variants which have been linked to ~25% of cases. The underlying germline genetic architecture of OS appears disproportionately weighted towards rare variants, but there has only been one previous OS GWAS of European-ancestry cases, which identified two common risk loci. Here, we conducted a large multi-ancestry GWAS of OS including 2,585 OS cases and 63,422 cancer-free controls from four ancestral groups to identify common germline genetic variants associated with susceptibility. **Methods:** We genotyped germline DNA from 1,843 OS cases and 59,616 controls of European ancestry, and 743 multi-ancestry cases and 3,110 matched multi-ancestry controls, including Latin-American-1, Latin-Amercian-2 and African-American sets (GRAF-pop classifications). After quality control and imputation, the adjusted (sex, principal components) mixed model logistic regression association results were combined across ancestral groups using an inverse-variance weighted fixed-effect meta-analysis. There was no evidence of genomic inflation in the three European case-control sets (λ=1.01) or the multi-ancestry meta-analysis (λ=0.99). Stratified analyses were conducted for young-onset (age<25 yrs) and adult-onset (age>25 yrs) cases. **Results:** One locus at 12q24.13 was associated with OS at genome-wide significance and was consistent across the three European ancestry case-control sets (odds ratio [OR] 1.25, 95%CI 1.2-1.4, P=2.3x10-8). The association was similar in the young-onset (OR 1.25; MAF 0.33) and adult-onset (OR 1.28; MAF 0.33) cases. This locus remained genome-wide significant in the multi-ancestry meta-analysis (OR 1.22, 95%CI 1.1-1.3, P=4.6x10-8); the top SNP was associated with OS in the Latin-American-2 and African-American case-control sets but not the Latin-American-1 set. Fine-mapping identified a single causal signal that reflects a broad association peak spanning 1 Mb. The top SNP is intergenic and located within a gene-dense region, and it is predicted to be a regulatory variant associated with gene expression and chromatin accessibility. There are two interesting nearby genes with significant eQTLs (P<2x10-8): *PTPN11*, a known oncogene with germline mutations causative of Noonan syndrome; and *MAPKAPK5*, a tumor suppressor gene involved in regulation of TP53 activity, a gene of known importance in OS etiology. This is the largest GWAS of osteosarcoma susceptibility to date that included multi-ancestral groups and identified a novel susceptibility locus on chromosome 12.

Title: Tumor transcriptomic profiling in racially and ethnically diverse colorectal cancer patients to identify new predictors of mortality

Authors:

H. Yin[1,2], J. Huyghe[2], D. Redwood[3], T. Harrison[1], A. Dawson[2], S. Thomas[2], B. Seaton[2], A. Koehne[2], H. Green[4], K. Nasir[5], J. Tiesinga[6], C. Whitlow[4], L. Hsu[2], J. Figueiredo[7,5], L. Li[4], T. Thomas[3], C. Li[1,2], U. Peters[2,1]; [1]Univ. of Washington, Seattle, WA, [2]Fred Hutchinson Cancer Ctr., Seattle, WA, [3]Alaska Native Tribal Hlth.Consortium, Anchorage, AK, [4]Ochsner Clinic Fndn., New Orleans, LA, [5]Cedars-Sinai Med. Ctr., Los Angeles, CA, [6]Alaska Native Med. Ctr., Anchorage, AK, [7]Univ. of Southern California, Los Angeles, CA

Abstract:

Tribal Health Organizations recognize the high rates of colorectal cancer (CRC) among Alaska Native peoples and are undertaking initiatives to address it. Studying tumor profiles in patients from diverse racial and ethnic groups will lead to novel molecular insights into CRC. We are generating RNA-seq data from a nested case-control study including 840 African American, Alaska Native, Hispanic, and non-Hispanic White CRC patients. Within each racial and ethnic group, we include 70 lethal CRC patients and 140 age-, sex-, stage-, and diagnostic year-matched non-lethal CRC patients. Using the tumor transcriptomic data, we aim to understand similarities or differences in gene expression features across race and ethnicity and evaluate their associations with CRC mortality. Our RNA-seq run included 273 patients who were equally distributed across four racial and ethnic groups. Quality control including an exonic rate >70% and >5.5 million uniquely mapped read pairs excluded 56 patients from downstream analysis. We estimated microsatellite instability (MSI) status, consensus molecular subtypes (CMS), and calculated a T cell-inflamed gene expression profile (GEP) score for 217 patients. We observed a lower proportion of Alaska Native patients in the “metabolic” subtype (8%) and lower proportion of African American patients in the “MSI and immune" subtype (6%), but the distribution of CMS was not different across race and ethnicity (P = 0.65). We observed differences in mean GEP scores between race and ethnicity (P = 0.001), with a higher mean score among Alaska Native patients (P = 0.0004) but a lower score among African American patients (P = 0.02), compared to non-Hispanic White patients. Differential expression and gene set enrichment analysis identified 32 genes (e.g. *ANGPT2*) and 23 pathways (e.g., Hedgehog signaling pathway) up-regulated, and 6 genes (e.g. *TTF2*) and 24 pathways (e.g., Antigen presentation pathway) down-regulated among lethal CRC patients compared to non-lethal CRC patients. Cell deconvolution using xCell inferred enrichment scores for 64 cell types. Immune cells were enriched among non-lethal CRC patients and fibroblasts were enriched among lethal CRC patients. We expect to have over 600 samples completed by summer 2023 and a more comprehensive prognostic index will be developed at that point. With the inclusion of patients from more diverse populations, our study will enhance the identification of novel clinically useful predictors of lethal CRC and potential novel therapeutic targets that could meaningfully reduce long-standing CRC disparities as well as increase the generalizability of findings across major racial and ethnic groups.

Title: Rare variants associated with prostate cancer risk discovered from 269,920 male exomes influence risk of prostate cancer metastasis

Authors:

K-H. Wu[1], M. Kessler[2], J. Tang[1], J. Mower[1], N. Banerjee[1], Regeneron Genetics Center, DiscovEHR, Mayo Clinica Project Generation, Penn Medicine Biobank, M. Cantor[2], A. Ferrando[1], E. Jorgenson[2]; [1]Regeneron Pharmaceuticals, Tarrytown, NY, [2]Regeneron Genetics Ctr., Tarrytown, NY

Abstract:

Background Advancements in medicine and cancer screening techniques have led to a >99% five-year survival rate in patients with low grade and early-stage prostate cancer (PCa). However, the five-year survival rates are significantly lower in the case of for high grade tumors and particularly for advanced metastatic disease (32%). Hence, there is great interest in strategies to optimize screening approaches to identify PCa patients at risk of developing tumor metastasis. **Objectives** This study aimed to evaluate the effect of rare coding variants on PCa risk and then to study how identified variants affect PCa 1) age of onset and 2) time to metastasis. **Methods** We leveraged whole exome sequencing and electronic health record data from ~270K samples across 8 biobanks to conduct the largest Exome-Wide Association Study to date for PCa in 269,920 male participants (32,166 cases; 12%). We used these data to perform genetic analysis, which included testing for common variant associations and gene level associations. Summary statistics derived from common variants were further used to calculate Polygenic Risk Score (PRS). The gene level association was studied by aggregating predicted loss of function and/ or deleterious missense variants with minor allele frequency less than 1% in a gene burden framework. Discovery of rare variant signals was done across 7 biobanks, with UK Biobank held out and used to evaluate the difference of time-to-event between rare variant carriers and non-carriers. **Results** We identified six gene burdens, which are significantly ($p<2.5x10^{-6}$ =0.05/19,422) associated with increased risk of PCa, from the meta-analysis result of 7 discovery cohorts. In UK Biobank, significantly higher hazard ratios (HR) of 1) developing PCa (HR: 1.55; p: 1.03e-45) and 2) PCa metastasis (HR: 1.58; p: 0.002) were observed in PCa risk-associated rare variant carriers (n=10,382), compared to non-carriers. Moreover, we found that carriers of PCa risk increasing rare variants reached PCa lifetime risk of 13% (population incidence rate) four years earlier (at age 71) than individuals without such variants (at age 75). The increased risk among 10,382 PCa risk-associated rare variant carriers of developing PCa earlier in life was consistently observed across all ten common variant PRS decile groups. These results suggest that accounting for both common and rare germline variants can improve PCa screening, surveillance, and treatment decisions. **Conclusions** These results demonstrate that high effect rare variants influence PCa age of onset and metastatic risk, which suggest that they can be used to inform a more personalized approach to diagnosis and treatment.

Title: Germline polygenic risk scores are associated with immune cell infiltration in breast cancer

Authors:

Y. Liu[1], C. Peng[2], Y. J. Heng[3], G. M. Baker[3], D. G. Stover[4,5], R. M. Tamimi[6], P. Kraft[1,7]; [1]Dept. of Epidemiology, Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA, [2]Channing Div. of Network Med., Brigham and Women's Hosp., Boston, MA, [3]Dept. of Pathology, Beth Israel Deaconess Med. Ctr., Boston, MA, [4]Div. of Med. Oncology, Stefanie Spielman Comprehensive Breast Ctr., Columbus, OH, [5]Dept. of BioMed. Informatics, Ohio State Univ., Columbus, OH, [6]Dept. of Population Hlth.Sci., Weill Cornell Med., New York, NY, [7]Dept. of Biostatistics, Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA

Abstract:

Tumor immune microenvironment (TIME) plays key roles in tumor progression and response and resistance to immunotherapy. Studies have shown that germline genetic variants contribute to differences in TIME in cancer patients, with most of the prior studies focused on assessing the effect of individual common or rare variants. Here, we hypothesize that common variants underlying breast cancer risk or cancer-related traits, represented by polygenic risk scores (PRS), may jointly influence immune features in TIME. Our study was conducted in the Nurses' Health Study (NHS) and NHSII cohorts. We derived 145 immune traits, including immune cell fractions, cell type enrichment scores, and gene expression signatures related to immunomodulatory signaling from bulk gene expression profiles of 764 breast tumors and 598 adjacent normal tissue from 825 breast cancer patients using CIBERSORTx, TIMER 2.0, xCell, EPIC, quanTIseq, MCPcounter, and based on previous studies. Immunohistochemistry for four immune cell markers: CD4, CD8, CD20, and CD163, were conducted among 205 patients for a direct measurement of immune cell infiltration in breast tumor. Germline PRS were calculated for breast cancer, immune-related diseases, type 2 diabetes, ages at menarche and menopause, body mass index (BMI), BMI-adjusted waist-to-hip ratio (WHR), alcohol intake, and tobacco smoking. We assessed the associations between germline PRS and immune traits using linear and logistic regression adjusting for age at diagnosis and the top ten genetic principal components. Overall, we identified 243 significant associations between germline PRS and immune traits in breast cancer, including 144 associations in breast tumor and 99 associations in tumor-adjacent normal breast tissue ($P < 0.05$). Notably, we observed positive associations between BMI PRS and major histocompatibility complex class-I expression signatures (smallest $P = 8.26 \times 10^{-4}$) and WHR PRS and overall lymphocyte infiltration ($P = 2.45 \times 10^{-3}$) in breast tumor. Negative associations were observed between inflammatory bowel disease (IBD) PRS and interferon signatures (smallest $P = 5.37 \times 10^{-4}$) and breast cancer PRS and CD8+ T cell abundance (smallest $P = 2.72 \times 10^{-3}$) in tumor, and IBD PRS and B cell abundance in normal tissue (smallest $P = 3.37 \times 10^{-3}$). Our findings provide insights into the germline genetic determinants of the immune cell composition in TIME and the underlying genetic link between breast cancer and cancer-related traits. Further study of this germline-TIME link may inform us of novel therapeutic targets in breast cancer and improve personalized cancer immunotherapy.

Title: Polygenic risk score substantially adds to prostate specific antigen for prostate cancer screening in US and UK biobanks

Authors:

V. Plagnol[1], D. Thompson[1], R. M. Sivley[1], C. Nolan[1], C. Fisher[2], R. Linner[3], M. Oetjens[2]; [1]Genomics plc, Oxford, United Kingdom, [2]Geisinger, Lewisburg, PA, [3]Leiden Univ., Leiden, Netherlands

Abstract:

Prostate cancer (PrC) is the most common cancer among men in the United States, with the introduction of prostate specific antigen (PSA) in the 1980s having a profound impact on disease detection. Yet, the utility of PSA based asymptomatic screening for PrC remains a topic of debate, due to limited sensitivity and specificity of the test, combined with the typically slow progression of PrC.
Polygenic risk scores (PRS) for PrC provide a single metric summarizing an individual's inherited disease risk. Combining PRS with PSA testing can potentially improve the performance of screening strategies. Here, we examined the integration of clinical and genetic factors for risk stratification of prostate cancer among men in two large scale biobanks: the MyCode cohort from Geisinger and the UK Biobank.
For this study, we used a novel prostate cancer PRS and tested its predictive performance alone and in combination with PSA measurements. The predictive accuracy of PRS for diagnosed PrC was consistent between UK Biobank and Geisinger (odds ratio per SD 2.13 (95% CI 2.09 - 2.17) in UKB and 2.02 (95% CI 1.93 - 2.17) in Geisinger). Both datasets showed a higher PRS effect size among younger individuals; UKB odds ratio per SD of PRS 2.99 (95% CI 2.76 - 3.24) in 40-54yo men, compared to 1.86 (95% CI 1.80 - 1.93) in 70-79yo men. The overall incidence of diagnosed PrC was significantly elevated in Geisinger compared to UK Biobank.
Using the first recorded PSA measurement and diagnosis of PrC within 5 years as an outcome, the PRS contributed substantial discrimination for men with intermediate PSA levels. Among men with a first PSA measurement between 4-6 ng/mL close to the follow-up threshold, the 5 yr PrC diagnosis incidence varied between 9.2% (95% CI 6.2%-13.0%) in the bottom 20% of the PRS and 29.7% (95% CI 26.3%-33.3%) in the top 20% of the PRS (UKB). The joint predictive accuracy of the PRS and PSA was highly consistent between UK Biobank and Geisinger.
These results indicate that the PRS provided substantial and consistent discrimination across both biobanks for the diagnosis of PrC, especially in younger individuals, potentially identifying a subset of the population that would derive greater benefit from regular PSA screening. The integration of the PRS and PSA for risk stratification, particularly among those with a PSA measurement of intermediate risk, provides an opportunity to refine decision making around subsequent follow-up.

Session 013: Pharmacogenomics in the era of next-generation sequencing

Location: Conv Ctr/Room 147A/Level 1

Session Time: Thursday, November 2, 2023, 8:30 am - 10:00 am

Title: Assessing the underrepresentation of diversity in evaluating *CYP2D6* genotypes: a study of 490,558 genomes in the UK Biobank

Authors:

X. Jiang[1], F. Hu[1], S. Atanur[1], X. Zou[1], G. Alamgir[1], A. O'Neill[1], J. Harrow[1], A. Abbasi[1], S. Deevi[1], M. Fabre[1], Q. Wang[2], S. Petrovski[1], W. Rae[1], K. Smith[1]; [1]AstraZeneca, Cambridge, United Kingdom, [2]AstraZeneca, Waltham, MA

Abstract:

Introduction *CYP2D6*, is one of the most important pharmacogenes: depending on the corresponding genotype, standard doses may be ineffective or produce adverse effects. As this complex locus is not well captured using exome sequencing or genotype arrays, we used whole-genome sequencing (WGS) data from UK Biobank (UKB) to perform a phenome-wide association study of *CYP2D6* genetic-predicted metaboliser category with clinical outcomes. **Methods** We obtained *CYP2D6* star allele diplotypes using DRAGEN v3.7.8's specialised caller. Following standardised criteria from the Clinical Pharmacogenetics Implementation Consortium (CPIC), we used activity scores computed from genotypes to classify individuals into four categories. We subsequently investigated associations between *CYP2D6* metaboliser category contrasts and 15,909 binary outcomes within five ancestries and meta-analysed the results. **Results** We observed 95 distinct *CYP2D6* star alleles in 490,558 UKB genomes. Calls were Mendelian consistent in 98.8% of 1,014 trios. 48 (51%) had unknown effects on function according to CPIC, enriched for rare star alleles, resulting in the inability to assign a CPIC metaboliser category to 5-6% of African, Admixed American, and South Asian ancestry participants compared to ~2% of European and East Asian ancestry participants. We also detected 265 (0.05%) carriers of 76 rare (minor allele frequency ≤ 0.001) protein-truncating annotated *CYP2D6* variants, that were not captured by current star allele definitions. Of the 222,068 UKB participants with prescription data, 24.5% have been prescribed codeine. 6.8% and 1.6% of these participants are classified as poor (PMs) or ultra-rapid metabolisers (UMs) of *CYP2D6*. In multi-ancestry meta-analyses, *CYP2D6* normal metabolisers (NMs) and UMs had a higher risk of allergy to narcotic agents compared to PMs and intermediate metabolisers ($P=4.5x10^{-9}$; OR: 1.2). *CYP2D6* NMs and UMs were also associated with higher risk of kidney calculus ($P=8.9x10^{-8}$; OR: 1.2). **Conclusion** Lack of functional data for *CYP2D6* alleles disproportionately affects the ability to give pharmacogenetic advice to non-European ancestries. Only 0.05% of UKB participants carried a rare small truncating variant not considered in current classifications. Our phenome-wide study found two significant associations ($p<2.6x10^{-7}$) between *CYP2D6* categories and clinical outcomes. Future directions include studying associations with drug dose and treatment outcomes across a wider catalogue of *CYP2D6* metabolised medicines.

Title: CYP2C19 Loss-of-Function Variants and Clopidogrel Efficacy in Diverse Individuals with a Recent Myocardial infarction or Percutaneous Coronary Intervention in the All of Us Research Program

Authors:

A. Babbar[1], H. Mo[2], A. Williams[1], D. Schlueter[1], J. Keaton[1], T. Ferrara[1], S. Goleva[1], O. Stubblefield[2], T. Cassini[1], T. Tran[2], A. Awan[2], C. Zeng[1], J. Denny[1]; [1]NIH, Bethesda, MD, [2]Natl. Human Genome Res. Inst., Bethesda, MD

Abstract:

Antiplatelet therapy with clopidogrel and aspirin has been found to be effective in reducing subsequent events in patients with a recent myocardial infarction (MI) or percutaneous coronary procedure. Clopidogrel is a prodrug requiring conversion into its active metabolite, primarily by CYP2C19; as such, it has been found to be less effective in carriers of CYP2C19 loss-of-function (LoF) variants, representing ~25% of the population of European ancestry. The clinical impact of these LoF variants in other populations has been under-studied. We evaluated the association of CYP2C19 LoFs with clopidogrel response in multiple ancestral groups in the All of Us research program, which currently has 245,388 participants with whole genome sequencing and aims to recruit 1 million participants that reflect the diversity of the US population with a focus on historically understudied populations in biomedical research. Using the accessible GUI tools pre-installed at the All of Us workbench, we identified a total of 1937 participants that were on clopidogrel therapy after a recent MI and/or intracoronary stent placement, among whom ~32% were of non-European ancestry, including 365 participants of African ancestry. We identified 545 carriers of the CYP2C19*2 LoF variant using the linked whole genome sequencing data. We found that the CYP2C19*2 variant was associated with a higher risk of recurrent cardiovascular events in trans-ancestry populations (hazard ratio (HR): 1.53, 95% CI: 1.43-1.63, P = 6.56e-38), populations of European ancestry (HR: 1.50, 95% CI: 1.38-1.62, P = 5.9e-23) and populations of African ancestry (HR: 1.20, 95% CI: 1.02-1.40, P = 0.02). Additionally, we identified 762 carriers of CYP2C19*17 (gain of function variants). As expected, we found that this variant was associated with a lower risk of recurrent events in trans-ancestry populations (HR: 0.76, 95% CI: 0.71-0.81, P = 1.3e-17) and populations of European ancestry (HR: 0.73, 95% CI: 0.68-0.79, P = 5.4e-16), However, no association was found in populations of African ancestry (HR: 0.97, 95% CI: 0.83-1.13, P = 0.66). Our findings supported the functional impacts of CYP2C19*2 variants on clopidogrel response across multiple ancestral populations. More investigation is needed on CYP2C19*17 variants in African ancestry populations. Notably, this time-sequence phenotype algorithm was able to be accomplished with the graphical All of Us cohort builder tool. Further research should explore efficacy in other diverse populations and with other clopidogrel pharmacovariants.

Title: Rare variants from genes implicated in GWAS of hearing loss, but not from Mendelian hearing loss genes, are associated with cisplatin-induced hearing loss

Authors:

S. Guagliardo[1], P. C. Dinh[2], X. Zhang[3], M. R. Trendowski[3], S. Nakshatri[3], R. D. Frisina[4], Regeneron Genetics Center, L. B. Travis[2], E. Dolan[3], N. J. Cox[5], M. M. Shuey[5]; [1]Vanderbilt Univ., Nashville, TN, [2]Div. of Med. Oncology, Indiana Univ., Indianapolis, IN, [3]Dept. of Med., Univ. of Chicago, Chicago, IL, [4]Dept.s of Med. Engineering and Communication Sci. and Disorders, Global Ctr. for Hearing and Speech Res., Univ. of South Florida, Tampa, FL, [5]Vanderbilt Genetics Inst., Vanderbilt Univ. Med. Ctr., Nashville, TN

Abstract:

Cisplatin is a commonly used cytotoxic drug, and highly effective in combination chemotherapy for testicular cancer. Cisplatin use, however, is associated with various toxicities including tinnitus (TINN), hearing loss (HL), and peripheral neuropathy (PN). Due to the early age of onset for testicular cancer, these toxicities can be particularly detrimental to the long-term quality of life of survivors. Understanding the genetic predisposition to these toxicities is of particular interest to physicians as it provides the potential to personalize preventive care and assessment.

In a cohort of 1,680 testicular cancer survivors of genetically determined European Ancestry, we tested whether rare variation in genes associated with Mendelian or polygenic forms of TINN, HL, and PN were substantially enriched in patients experiencing these adverse outcomes. High-throughput whole exome sequencing was performed by Regeneron Pharmaceuticals, and 1,663 samples remained after undergoing various quality control filters. Rare variant gene burden testing was also done by Regeneron, aggregating various types of rare variants (predicted loss of function, missense variants, deleterious missense, and validated deleterious missense) at different minor allele frequencies (< 0.1, 0.5, 0.01, singleton) for a total of 16 burden masks. We performed gene set enrichment analyses on these burden results based on pre-specified gene lists for Mendelian genes: non-syndromic HL (n=154), syndromic HL (n=199) and PN (n=29). Results derived from the GWAS Catalog were used to curate gene sets for polygenic risk of TINN (n=97), HL (n=250), and PN (n=31).

There was significant enrichment of polygenic gene sets for TINN and HL ($p = 0.001$, $p = 0.02$, respectively), but not Mendelian gene sets. We identified no significant enrichment in PN polygenic gene sets possibly owing to phenotypic variability. Our results suggest that rare variation in genes associated with polygenic forms of TINN and HL are enriched in patients experiencing cisplatin-induced TINN and HL.

Title: Characterisation of *CYP2D6*, *CYP2B6*, and *CYP2A6* pharmacogenetic variation in African populations and development of the StellarPGx diplotype calling algorithm.

Authors:

D. Twesigomwe[1,2], Z. Lombard[2], S. Hazelhurst[1,3], the Wits-H3Africa/GSK ADME Collaboration; [1]Sydney Brenner Inst. for Molecular BioSci., Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa, [2]Div. of Human Genetics, Natl. Hlth.Lab. Service and Sch. of Pathology, Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa, [3]Sch. of Electrical and Information Engineering, Univ. of the Witwatersrand, Johannesburg, South Africa

Abstract:

Background and aims: Genetic variation is in part responsible for the variability in drug response across populations. However, the full catalog of pharmacogenetic variants, and their distribution, are yet to be established, in particular for African populations. This study therefore aimed to characterize variation in three core drug metabolism pharmacogenes, namely *CYP2D6*, *CYP2B6,* and *CYP2A6*, across diverse and/or previously underrepresented African populations. **Methods**: Given the challenges of diplotyping these hypervariable genes, we developed a novel bioinformatics pipeline (StellarPGx) to facilitate star allele (haplotype) detection from short-read whole genome sequence (WGS) data. This was followed by benchmarking based on simulated data and 109 real-world WGS datasets from participants with gold-standard diplotypes reported by the CDC's Genetic Testing Reference Materials Coordination Program. We then used StellarPGx and other existing tools to call *CYP2D6*, *CYP2B6*, and *CYP2A6* star alleles from high-coverage WGS data from 961 African participants and over 2000 genomes (for comparison) from global populations. **Results**: In our benchmarking analysis, StellarPGx (99%) and Cyrius (98%) had the highest concordance to ground truth *CYP2D6* diplotypes. From the genomic mining, we present frequencies across sub-Saharan Africa (SSA) for star alleles in *CYP2D6* (e.g. **17*, **29*, and **5*—frequency of 20%, 10%, and 8%, respectively), *CYP2B6* (e.g. **6* and **18*—frequency of 33% and 10%, respectively) and *CYP2A6* (e.g. **4*, **9*, and **46*—frequency of 3%, 8%, and 6% respectively), compared to other global populations. StellarPGx identified novel African-ancestry star alleles in *CYP2D6* (n=27), *CYP2B6* (n=18), and *CYP2A6* (n=31); seven of these alleles were validated via targeted Single-Molecule Real-Time (SMRT) resequencing. In addition, collaboration with experts from the Pharmacogene Variation Consortium (PharmVar) has enabled the validation of a further 12 novel *CYP2D6* star alleles. Our phenotype predictions for *CYP2D6* and *CYP2B6* indicate that the landscape of drug metabolizer phenotypes is non-uniform across SSA and differs to a large extent from phenotype distributions in global populations. **Conclusion**: Our findings underscore the need for investigating pharmacogene variation in the African context to reliably inform clinical pharmacogenomics implementation in Africa and the African diaspora. One African population cannot be used as a proxy for the whole continent. We recommend using StellarPGx for high-coverage WGS-based identification of known and novel pharmacogene haplotypes across all global populations.

Title: Pharmacokinetics of Tamoxifen and its major metabolites and the effect of the African ancestry specific CYP2D6*17 variant on the formation of the active metabolite, endoxifen.

Authors:

C. Kanji[1], G. Nyabadza[1], C. Nhachi[2], C. Masimirembwa[1,3]; [1]African Inst. of BioMed. Sci. and Technology (AiBST), Harare, Zimbabwe, [2]Univ. of Zimbabwe, Harare, Zimbabwe, [3]Sydney Brenner Inst. for Molecular BioSci. (SBIMB), Univ. of the Witwatersrand, Johannesburg, South Africa

Abstract:

Tamoxifen is widely used in the treatment of hormone receptor-positive breast cancer. CYP2D6 is an enzyme that converts tamoxifen to its highly potent secondary metabolite, endoxifen. Studies have been conducted to determine the association between CYP2D6 genotype and endoxifen concentrations, but there is limited research on the association with reduced activity variants such as CYP2D6*17 which occur at high frequencies in people of African ancestry. We conducted a single-dose pharmacokinetic study in humans to investigate the effects of CYP2D6*17 variant, on the pharmacokinetics (PK) of tamoxifen and its active metabolites in 42 healthy black Zimbabweans. Subjects were grouped based on CYP2D6 genotypes as CYP2D6*1 or *2, CYP2D6*1/*17 or 2*/*17, and CYP2D6*17/*17. PK parameters for tamoxifen and its metabolites, N-desmethyl tamoxifen, 4-hydroxy-tamoxifen, and endoxifen, were determined. Metabolic ratios of endoxifen/ N-desmethyl tamoxifen were used to interpolate the predicted activity score. The non-parametric superposition tool in Phoenix WinNonlin was used to estimate tamoxifen dose increases to predict endoxifen steady-state plasma concentrations. The PK of endoxifen showed statistically significant differences among the three genotype groups. CYP2D6*17 gene carriers had significantly lower endoxifen exposure levels than CYP2D6*1 or *2 carriers, with the mean endoxifen $AUC_{0-\infty}$ in CYP2D6*17/*17 subjects being 452.01 (196.94) h*ng/mL, which was 5-fold lower than in CYP2D6*1 or *2 subjects. Individuals who were heterozygous or homozygous for CYP2D6*17 alleles showed a 2- and 5-fold decrease in Cmax, respectively, compared to the CYP2D6*1 or *2 genotype. Pharmacokinetic parameters of tamoxifen and the two primary metabolites, N-desmethyl tamoxifen and 4-hydroxy tamoxifen, did not show any significant difference in the three genotype groups. With observed differences in endoxifen exposure levels, we predicted an activity score value of 0.3 which is lower than the established activity score value of 0.5. Using single dose study and simulating to steady state we observed that CYP2D6*17/*17 patients failed to reach the 5.9 ng/ml endoxifen putative threshold, but further simulations showed that dose escalation to 40 mg per day resulted in all individuals homozygous for CYP2D6*17 having therapeutically effective endoxifen concentrations. The African-specific CYP2D6*17 variant showed reduced activity with a predicted activity score of 0.3, resulting in lower endoxifen exposure levels. CYP2D6-informed tamoxifen dosing could benefit IM patients who are homozygous for this variant.

Title: Faster CYP2A6 activity increases chronic obstructive pulmonary disease and lung cancer risk by influencing smoking: A mediation Mendelian Randomization study in the UK biobank.

Authors:

H. Giratallah[1,2], M. J. Chenoweth[1,2,3], J. Pouget[2,3], A. El-Boraie[1,2], C. Lerman[4], J. Knight[5], R. F. Tyndale[1,2,3]; [1]Dept. of Pharmacology & Toxicology, Univ. of Toronto, Toronto, ON, Canada, [2]Campbell Family Mental Hlth.Res. Inst., Toronto, ON, Canada, [3]Dept. of Psychiatry, Univ. of Toronto, Toronto, ON, Canada, [4]Univ. of Southern California, Los Angeles, CA, [5]Lancaster Univ., Bailrigg, United Kingdom

Abstract:

Background: Previous investigations, including candidate gene studies and our phenome-wide association study (PheWAS) in the UK Biobank, have linked genetic variation in CYP2A6, a critical enzyme involved in nicotine and nitrosamine metabolism, with the risk of Chronic Obstructive Pulmonary Disease (COPD) and lung cancer. Given CYP2A6's influential role in modulating various smoking behaviors, including cigarette consumption and smoking topography, we explored whether smoking consumption measures mediated the association between *CYP2A6* variation and disease risk. **Methods**: We implemented a two-step bidirectional mediation Mendelian Randomization (MR) approach. We used our weighted *CYP2A6* genetic risk score (wGRS) as an instrument for CYP2A6 activity and assessed obstructive chronic bronchitis and cancer of bronchus (lung) as outcomes. Our analysis incorporated three mediators: pack-years (chronic smoking exposure), the number of cigarettes smoked per day (CPD; self-reported smoking quantity), and the summation of nicotine's main metabolites, cotinine and 3'-hydroxycotinine (COT+3HC; a biological measure of nicotine intake). We sourced genetic instruments for these mediators from the summary statistics of genome-wide association studies. To rule out CYP2A6 activity as a mediator of the effect of smoking on disease risk, we conducted reverse mediation MR by swapping the exposure with the mediators. We also executed mediation analyses for mediators directly obtained from the UK Biobank, namely pack-years and CPD. All analyses were restricted to European-ancestry individuals and adjusted for age, sex, and genetic principal components. **Results**: Faster CYP2A6 activity (i.e., higher wGRS) was associated with higher pack-years, CPD, and COT+3HC (p <0.001). All three measures significantly mediated the estimated effects of CYP2A6 activity on obstructive chronic bronchitis and cancer of bronchus (lung) risk in mediation MR ($p_{indirect}$ <0.05). Pack-years and CPD indirect effects were also significant in mediation analyses (p <0.01). The percentage of effect mediated (PM) by all smoking mediators was higher for obstructive chronic bronchitis (PM >50%) than for cancer of bronchus (lung) (PM <30%). Smoking mediators did not alter CYP2A6 activity in reverse mediation MR. **Conclusions**: The mediation MR and mediation analyses support a mechanistic role of *CYP2A6* variation in smoking consumption, subsequently escalating the risk of COPD and lung cancer. The mediation effect of smoking consumption appears more substantial in the context of COPD.

Session 014: Structural variation and the human genome

Location: Conv Ctr/Ballroom B/Level 3

Session Time: Thursday, November 2, 2023, 8:30 am - 10:00 am

Title: Ancestral bias alters structural variant breakpoints

Authors:

P. Audano[1], C. Beck[2]; [1]The Jackson Lab., Farmington, CT, [2]The Jackson Lab. and UConn Hlth., Farmington, CT

Abstract:

High-quality genome assemblies and sophisticated algorithms have increased sensitivity for a wide range of variant types, and breakpoint accuracy for structural variants (SVs, ≥ 50 bp) has improved to near basepair precision. Despite these advances, many SVs in unique regions of the genome are subject to systematic bias that affects breakpoint location. This ambiguity leads to less accurate variant comparisons across samples, and it obscures true breakpoint features needed for mechanistic inference. To understand why SVs are not consistently placed, we re-analyzed 64 phased haplotypes constructed from long-read assemblies released by the Human Genome Structural Variant Consortium (HGSVC) for inconsistent breakpoints identifying 882 SV insertions and 180 SV deletions not anchored in tandem repeats (TRs) or segmental duplications (SDs). Sequence and assembly error has minimal impact on breakpoint accuracy, however, there is a strong effect by ancestry. We confirm that mismatches and small indels are enriched at shifted breakpoints, but these polymorphisms are lost as a result of breakpoints shifting. Long tracts of homology, such as SVs mediated by TEs, increase the number of imprecise breakpoints and the distance they are shifted over. Tandem Duplication (TD) breakpoints are most heavily affected with 14% of TDs placed at different locations across haplotypes resulting in a chimeric representation, and the true breakpoint is lost. While graph genome methods normalize SV breakpoints across many samples, the resulting SVs are sometimes inconsistent with biology highlighting a need to tune graph methods for breakpoint accuracy. The breakpoint inconsistencies we characterize collectively affect ~5% of the SVs called in a human genome and underscore a need for algorithm development to improve SV databases, mitigate the impact of ancestry on breakpoint placement, and increase the value of callsets for investigating mutational processes.

Title: Assessing tissue-specific effects of rare and structural variants towards gene regulation with the EN-TEx personal genome resource

Authors:

M. Jensen, J. Li, T. Michaels, A. Su, T. Galeev, S. Kumar, K. Xiong, B. Borsari, J. Rozowsky, M. Gerstein, EN-TEx Consortium; Yale Univ., New Haven, CT

Abstract:

Comprehensive tissue-specific analyses of how variants alter molecular phenotypes have greatly improved our understanding of genomic mechanisms for complex traits. For example, the EN-TEx resource, consisting of long read-based personal genomes and a full battery of 1,635 functional assays across 30 tissues in four donors, allowed us to systematically assess allele-specific activity of common single-nucleotide variants (SNVs). The long-read phased personal genomes in the EN-TEx resource are also ideal for studying the functional effects of a full spectrum of genomic variants, in particular rare SNVs and all types of structural variants (SVs), which are often under-represented in functional genomics studies. In this study, an extension of work from the EN-TEx Consortium, we aligned 381 tissue-specific and single-cell ATAC-Seq, DNAse-Seq, and ChIP-Seq datasets within EN-TEx onto personal genomes, and prioritized functional signals within variant regions adjacent to genes with altered expression. Alternative bioinformatic approaches such as graph-based and alt-aware mapping of sequencing data to personal genome haplotypes allowed us to better identify functional peaks in heterozygous regions, especially in novel insertions that are distinct from the reference genome. In these datasets, we found 293 SV-eQTLs per individual and linked key functional signals within these variants to genes with altered expression; for instance, deletion of an upstream H3K27Ac peak led to reduced expression of *ZFAND2A*. Globally, we found that open chromatin was reduced in the presence of nearby transposable elements, and 1.5% of identified allele-specific events could be attributed to SVs. We further identified 152 novel functional elements within genomic insertions per individual, and found that 16% of these variants could contribute to perturbed expression of nearby genes. For example, the tumor suppressor *ASMTL-AS1* showed 4.2-fold increased expression when coupled with an upstream duplication spanning a novel ATAC-Seq peak. Beyond SVs, we identified 620 rare SNVs per individual that disrupted candidate *cis*-regulatory elements (cCREs) near protein-coding genes, and prioritized variants in cCREs located near genes with tissue-specific altered expression. Overall, our study emphasizes the broad effects of rare variants and SVs in particular towards tissue-specific gene regulatory patterns, and highlights the utility of using personal genomes for accurate and comprehensive functional genomics studies.

Title: Integrative analysis of rare structural variants from long and short read sequencing and transcriptomic signals reveals novel determinants of undiagnosed mendelian disease.

Authors:

B. Ni[1], T. Jensen[2], P. Goddard[2], R. Ungar[2], B. Strober[3], N. Ersaro[4], T. Li[5], E. A. Ashley[2], M. Wheeler[2], S. B. Montgomery[2], M. C. Schatz[1], A. Battle[1]; [1]Johns Hopkins Univ., Baltimore, MD, [2]Stanford Univ., Stanford, CA, [3]Harvard Sch. of Publ. Hlth., Brookline, MA, [4]Illumina, Oakland, CA, [5]Johns Hopkins Sch. of Med., Baltimore, MD

Abstract:

Rare structural variants (SVs) are known to contribute risk to disease, yet few pipelines exist to reliably call SVs and estimate their effects. We developed a pipeline to call high-fidelity SVs and then used expression outliers (eOutliers) and splicing outliers (sOutliers) to prioritize functional rare SVs with SV-specific annotations. The pipeline integrates the genomic and transcriptomic signals in a probabilistic model of variant function called Watershed-SV. Trained using the GTEx v8 SV calls and outlier signals, Watershed-SV produced substantial improvements and prioritized more rare SVs that corresponded to eOutliers in held out individuals, confirming the large impact of rare SVs on gene expression. Watershed-SV also learned large weights for regulatory element annotations, enabling the prioritization of noncoding SVs.

We called SVs in two cohorts with RNA-seq and WGS using short and long read sequencing: the NHGRI Inherited Muscular Disease (IMD, n=26) and the Undiagnosed Disease network (UDN, n=69). We sequenced UDN patients with Oxford Nanopore Technology long reads and used Jasmine to merge SV calls from multiple tools (Sniffles, SVIM, and cuteSV). We called 267,769 rare SVs, which was three times the number found with short reads. For IMD, we used svtools and Parliament2 to call SVs. sOutliers and eOutliers were called against a background of healthy tissue from the GTEx project to determine outliers in disease-relevant genes. While short-read SV calling excels in finding duplications and deletions, we found more rare insertions and repeat expansions in UDN and observed enrichment of rare insertions for eOutliers.

Watershed-SV prioritizes on average 20 high-confidence functional SVs per patient in UDN, narrowing down to < 1% of rare SVs ascertained. Among them, genetic counselors identified a coding deletion depleted expression of *TANGO2* in a patient with undefined cardiomyopathy. In another case, we found a novel extreme VNTR expansion using long reads in the promoter of *FAM193B*, also identified by genetic counselors as a novel disease gene in a musculoskeletal disorder patient. In IMD, two coding inversions linked to both sOutlier and eOutlier in *DMD* gene were shown to cause Duchenne muscular dystrophy in previous studies. Additionally, a coding DEL (AF=0.0003) depleted expression of *PIP5K1C*, a gene linked to another severe muscle disorder, in a patient with limb girdle muscular dystrophy.

In summary, we developed a pipeline to integrate short-read sequencing, long-read sequencing, and transcriptomic outliers to identify rare functional SVs. This pipeline enabled the discovery of novel disease associated SVs in two rare disease cohorts.

Title: *MECP2* copy number variants studied by multiple approaches reveal impact of genomic structure to disease variability.

Authors:

J. Bengtsson[1], C. M. Grochowski[2], S. S. Bajikar[2], M. Gandhi[1], M. Y. Lun[1], J. Angad[3], R. S. Roy[1], M. Mahmoud[4], L. Paulin[4], M. Mehaffey[1], K. Park[1], H. Du[3], J. Eisfeldt[5], M. Pettersson[5], B. Suter[6], S. Jhangiani[4], D. Muzny[4], J. Fatih[2], R. Gibbs[7], A. Lindstrand[5], F. Sedlazeck[4], J. Lupski[2], H. Zoghbi[8], D. Pehlivan[2], C. C. M. Carvalho[1]; [1]Pacific Northwest Res. Inst., Seattle, WA, [2]Dept. of Molecular and Human Genetics, Baylor Coll. of Med., Houston, TX, [3]Baylor Coll. of Med., Houston, TX, [4]Human Genome Sequencing Ctr., Baylor Coll. of Med., Houston, TX, [5]Karolinska Inst.t, Stockholm, Sweden, [6]Section of Neurology and Dev.al NeuroSci., Dept. of Pediatrics, BCM, Houston, TX, [7]Baylor Coll. Med., Houston, TX, [8]Baylor Coll. of Med./HHMI/NRI, Houston, TX

Abstract:

Copy number variants (CNV) spanning the Xq28 region can result in at least three neurodevelopmental disorders. Xq28 overlaps the dosage sensitive gene *MECP2,* which is included in the smallest region of overlap (SRO) in copy-number gain causative of the *MECP2* duplication syndrome (MDS, MIM#300260). MDS has considerable interindividual clinical variability and disease severity which has puzzled researchers and clinicians for years. Our group has previously shown that ~ 30% of MDS is characterized by complex genomic rearrangements (CGRs) of intermediate complexity ($2 \leq$ CNV breakpoints < 5). To test the hypotheses that CGRs contribute to expression variability of dosage sensitive genes at Xq28, we have analyzed a large cohort consisting of 137 probands using combined genomics, transcriptomics, and a deep phenotyping approach. Genomic analyses including array CGH, optical genome mapping, and genome sequencing revealed the CNV size to be diverse, ranging from 64 kB to 16 Mb. Within these 137 individuals the proportion of structures is 47% tandem duplication (tandem DUP), 23% duplication extending to the telomere (terminal DUP), 23% Duplication-Triplication-Duplication (DUP-TRP/INV-DUP) and 7% other complex genomic structures. Genomic structural information has been incorporated with transcriptomic and phenotype analysis including Human Phenotype Ontology (HPO) semantic similarity scores. RNAseq data from lymphoblastoid cell lines indicated that the *MECP2* transcript in *MECP2* triplications is statistically different from tandem DUP, but not between other classes of genomic structures. Five probands carrying two copies of *MECP2* showed *MECP2* expression values greater than 1.5 quartiles from the median (fold change 2.6 or greater). Four are complex, i.e. two DUP-TRP/INV-DUP (two *MECP2* copies), two with terminal DUPs and one has a tandem DUP. In addition, the phenotype of MDS patients is generally more severe in cases with CGRs. For instance, MDS individuals display differences in overall survival, developmental delay, and earlier onset of seizures in order of increasing severity from tandem DUP, complex duplications, terminal DUP/translocations, and finally to triplications. Detailed genomic analyses have allowed similar genomic rearrangements to be separated into structural categories, which appear to correspond to increasingly severe phenotype as complexity increases pointing to the interplay between genomic structure and phenotype severity that cannot be explained purely by variation in *MECP2* copy-number. Utilizing this type of analytical approach will lead to advances in understanding the role of genomic rearrangements in disease.

Title: Unraveling the Contribution of Structural Variants on Cardiometabolic Traits in the UK Biobank.

Authors:

A. Basile[1], M. Byrska-Bishop[1], W. Clarke[1,2], A. Corvelo[1], U. Evani[1], X. Zhao[3,4,5], M. Talkowski[3,4,5,6], G. Narzisi[1], M. Zody[1]; [1]New York Genome Ctr., New York, NY, [2]Outlier Informatics Inc., Saskatoon, SK, Canada, [3]Program in Med. and Population Genetics, Broad Inst. of MIT and Harvard, Cambridge, MA, [4]Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [5]Dept. of Neurology, Massachusetts Gen. Hosp. and Harvard Med. Sch., Boston, MA, [6]Stanley Ctr. for Psychiatric Res., Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

Structural variants (SVs) are chromosomal rearrangements consisting of deletions (DEL), insertions (INS), duplications (DUP), inversions (INV), and translocations that are frequently defined as being longer than 50 base pairs (bp). This class of genomic variation accounts for greater genetic diversity than single nucleotide variation, and has been shown to contribute to phenotypic heterogeneity. Rare, *de novo*, and somatic SVs have been well implicated in sporadic human diseases and cancers, but the general contribution of SVs to common disorders has yet to be fully resolved.

Here, we perform an SV association study with 487,908 imputed UK Biobank samples to investigate the cardiovascular disease burden of common SVs, which have not been directly measured by traditional SNV associations. We extend our prior work by using our publicly released high-coverage 1000 Genomes Project (1kGP) SV integrated reference panel containing 73,452,337 SNVs/INDELs and 102,459 SVs from 3,202 samples. Moreover, we investigate the role of SVs in cardiovascular disease risk across multiple diverse ancestral populations. We performed imputation on array data from 487,908 UK Biobank subjects, employing Eagle2 and Minimac4. We confidently imputed 39,671 SVs (19,604 DEL; 14,923 INS; 210 INV; 4,934 DUP) with a Minimac R2 >0.3 and HWE P >1e-10. Association analysis was performed on each ancestry group, previously assigned by the Pan-UKB team (2020), and consisting of: African (n=6,802), Admixed American (n=994), Central/South Asian (n=9,107), East Asian (n=2,782), Middle Eastern (n=1,622), and European (n=409,520) samples. As a pilot for large scale cardiometabolic trait associations, GWAS was executed for triglycerides, total cholesterol, LDL, and HDL on imputed variants with MAF >0.5% and R2 >0.3 using SAIGE, and adjusting for age, sex, and principal components. We identified multiple significant SVs (P <1.7 x10-9; Bonferroni correction) associated with lipids. Among the most significant are a 97bp INS (AF=1.7%) in *PYY* associated with elevated triglyceride levels (Beta=0.25) and lowered HDL (Beta=-0.31), imposing an overall cardiovascular disease risk effect; a 51bp *NUP93* INS (AF=0.61%) associated with elevated HDL (Beta=0.35); and a 51bp *CEACAM22P* DEL with a cholesterol lowering effect (Beta=-0.22). The latter two genes have multiple lipid associations in the GWAS Catalog. We also replicated known SNV lipid associations in each assessed trait.

This work showcases the use of imputation based frameworks for studying the contribution of SVs to complex traits, and highlights the role of SVs in shaping the genetic landscape of cardiac traits.

Title: Copy-number variants as modulators of common disease susceptibility.

Authors:

C. Auwerx[1], M. Jõeloo[2], M. Sadler[1], N. Tesio[1], S. Ojavee[1], C. Clark[1], R. Mägi[2], Estonian Biobank Research Team, A. Reymond[1], Z. Kutalik[1]; [1]Univ. of Lausanne, Lausanne, Switzerland, [2]Univ. of Tartu, Tartu, Estonia

Abstract:

Copy-number variations (CNVs) have been associated with genomic syndromes but their impact on health later in life in the general population remains poorly described. Assessing four modes of CNV action (i.e., mirror, U-shape, deletion-only, and duplication-only models), we performed genome-wide association scans between the copy-number of CNV-proxy probes and 60 manually curated ICD-10 based clinical diagnoses in 331,522 unrelated white UK Biobank participants with replication in the Estonian Biobank. We identified 73 signals involving 40 diseases and stratified these in confidence tiers based on three statistical approaches: Fisher test, residual analysis, and Cox proportional-hazard modeling. CNVs always increased disease risk and caused earlier onset. Even after correcting for these signals, a higher CNV burden increased risk for 18 disorders, mainly through the number of deleted genes, suggesting a polygenic CNV architecture. Pathogenicity of individual CNV regions was driven both by the number and nature of encompassed genes, with CNVs affecting a larger number of genes being more pleiotropic (+0.16 associations per gene; $p = 1.5 \times 10^{-5}$) and genes encompassed by disease-associated CNVs being under stronger evolutionary constraint ($p_{pLI} = 1.3 \times 10^{-4}$; $p_{LOEUF} = 1.9 \times 10^{-7}$). Associations were supported by colocalization with both common and rare single nucleotide variant association signals, providing insights into the epidemiology of known gene-disease pairs typically studied in clinical cohorts (e.g., *BRCA1* and *LDLR* deletions increase ovarian cancer and ischemic heart disease risk, respectively), clarifying dosage mechanisms of action (e.g., both increased and decreased 17q12 dosage affects renal health), and identifying putative causal genes (e.g., *ABCC6* for kidney stones). Characterization of the pleiotropic pathological consequences of recurrent CNVs in adulthood indicated variable expressivity of these regions and the involvement of multiple genes. For instance, we find evidence for involvement of genes beyond *TBX1* in common cardiac phenotypes in 22q11.2 CNV carriers and identify novel non-neurological disease associations with 15q13 CNVs that specifically involve dosage of genes within an interval that does not span *CHRNA7*, the suggested candidate gene for neuropsychiatric afflictions. Together, our results shed light on the prominent role of CNVs in determining common disease susceptibility within the general population and provide actionable insights allowing to anticipate later-onset comorbidities in carriers of recurrent CNVs.

Session 015: The health equity puzzle: Piecing together disparities for a fairer future

Location: Conv Ctr/Room 202A/Level 2

Session Time: Thursday, November 2, 2023, 8:30 am - 10:00 am

Title: Social and psychological modifiers of genetic literacy in general and research samples

Authors:

G. Ramirez-Renta[1], I. Little[2], L. Koehly[3], C. Gunter[1]; [1]NIH, Bethesda, MD, [2]NIH, Charlotte, NC, [3]NHGRI, Bethesda, MD

Abstract:

Genetic literacy goes beyond knowledge of genetic terms, as it means having substantial genetic information and understanding to make decisions that support personal wellbeing and permit effective participation in social discussions about genetic issues. Several previous studies have explored the interaction of social modifiers, such as race or religious beliefs, and genetic understanding or attitudes towards genetics. In 2021, we created and disseminated a survey to two separate samples: 2050 members of the US general public and 2023 participants in a large genetic research study, SPARK. We assessed genetic literacy in three components (familiarity, knowledge, and skills), and found that genetic literacy had increased since data collected from a similar 2013 survey of 1016 individuals (Little et al, AJHG, 2022). To our knowledge, this is the largest study of genetic literacy; ours is also the only commonly used scale to actually measure genetic comprehension (Daly & Kaphingst, PEC Innovation, 2023). In the 2021 surveys, we also collected data on several social modifiers, including beliefs about racial essentialism (BARS) (Tawa, Cultur Divers Ethnic Minor Psychol, 2017) and religious beliefs (Botoseneanu et al, Health Educ Behav, 2017). Overall, we found no more than weak correlation between racial essentialism (the belief that racial groups form discrete genetic categories) and genetic literacy in our 2021 sample (n=4083). In contrast, our measures on religiosity uncovered two relationships: self-described religious belief (and denomination) does show a significant difference in genetic literacy scores. People who report that their religious teachings disagree sometimes with science also had significantly higher genetic literacy scores. Further, we saw that other modifiers, such as confidence in one's own genetic knowledge appeared to have improved over time, going from a moderately negative correlation in 2013 to a moderately positive one in both 2021 sample groups. These findings demonstrate the complexity inherent in endeavors to raise true genetic literacy in the population, as proposed by NHGRI in its strategic vision (Green et al, Nature, 2020), and demonstrate that context-specific genetics education is crucially important.

Title: Do Beliefs About Biologization of Race and Genetic Determinism Influence Views about the Ethical Acceptability of Using COVID-19 Host Genomics Information? A Survey of U.S. Health Professionals

Authors:

S. Jose, G. Geller, J. Bollinger, D. Mathews, J. Kahn, B. T. Garibaldi; Johns Hopkins Univ., Baltimore, MD

Abstract:

Background: Host genomics can play an important role in enabling targeted clinical and public health measures to control an infectious disease outbreak, such as COVID-19. The application of genomics to the infectious disease context raises several ethical, legal, and social implications (ELSI). Biologization of race and genetic determinism are two major ELSI issues that arise at this intersection of genomics and infectious disease. These issues are fraught with historical and present-day examples that reinforce the need to explore their potential impact on the use of host genomics in the infectious disease context. As predictive host genetic testing becomes available in the near future, health professionals will be an important stakeholder group involved in the implementation of these new technologies and they can provide key insights into these ELSI issues.
Methods: In 2021, a cross-sectional online survey was fielded to US health professionals. The survey explored their views about the ethical acceptability of using COVID-19 host genomic information in different settings, as well as their personal beliefs about biologization of race and genetic determinism.
Results: A total of 603 participants completed the survey. Overall, a majority of survey participants thought it was ethically acceptable to use COVID-19 host genomics to inform clinical and public health decisions. However, nearly 60% of the survey participants believed race to be a biological or genetic ancestral group. Only 5% believed that genomic risk factors should be prioritized over conventional risk factors when making decisions about COVID-19 in the hospital or public health settings. While a majority supported the overall use of host genomics in managing COVID-19, those who held highly deterministic views and/or believed that race was biological were more likely to support its use than those who did not.
Conclusion: These findings provide empirical evidence on how health professionals currently view the role of race in host genetic variation and the relative importance of genomic information over other factors. While identifying host genomic risk factors and understanding the links between disease risk and genetic ancestry may improve health outcomes, researchers should ensure that the host genomic findings are not misunderstood or miscommunicated. Acknowledging and accounting for the presence of these racialized and deterministic ideas among stakeholders and the histories embedded in genomics and infectious disease fields is critical for responsible implementation and to enable equitable access of infectious disease related genetic testing resources across all populations.

Title: The cost of inappropriate prediction algorithms on the health of minority individuals in large care-based Electronic Health Record

Authors:

J. Hirbo[1], N. Arora[2], P. Straub[1], N. Cox[3]; [1]Vanderbilt Univ. Med. Ctr., Nashville, TN, [2]Tulane Univ. Sch. of Med., New Orleans, LA, [3]Vanderbilt Univ Med Ctr., Nashville, TN

Abstract:

Individuals of African ancestry in the US bear significant and disproportionate health burdens, as demonstrated by most measures of health. Delivering high-quality, efficient, and equitable healthcare to all Americans require clinical systems that use appropriate models capturing the full complexity of diverse US populations. However, many of the clinical decisions are based on models and algorithms developed years — often decades — ago, based on information from small samples of European ancestry that fails to incorporate the diversity in our populations. Clinical laboratory tests are at the heart of the some of the most used algorithms for diagnosing chronic diseases, and assessing disease progression and treatment response, and clinicians take or recommend specific clinical follow-up if values fall outside predefined reference intervals (RI). The burden of RIs that fail to capture the diversity of populations in our healthcare systems fall disproportionately on US minorities, and costs incurred for the consequent repeat testing, under- and misdiagnosis, and under-/over-treatment of disease have never been systematically estimated. We performed a comprehensive analysis of population differences in clinical values and quantified consequent disparities in clinical follow-up between individuals of European and African ancestries in two large biobanks: Vanderbilt University Medical Centers biobank-BioVU (n=3.4 million) and All-of-US (n=400K). Out of over 1000 clinical values we analyzed 709 clinical laboratory values with at least 200 of individuals both ancestries and identified 306 (over 40%) that show systematic mean differences in levels. We further performed evaluation of subset of the 306 clinical lab values (50/306) with at least 1000 individuals to evaluate health inequities between the two ancestries. We use Vitamin D as a canonical example of these clinical laboratory values to show the role of genetics and other non-genetic factors in explaining the systematic differences observed. We then used prediction models that incorporates common and rare genetic variations, ancestry and social determinant of health parameters to refine individual levels on same scale across the two ancestries, to facilitate its clinical implementation. Our results have potential influence on health equity and will inform appropriate stakeholders on possible options for reducing inequities created by RIs poorly matched to the diverse populations.

Title: Frequency of known pathogenic variants across ancestries in the All of Us Research Program cohort

Authors:

E. Venner[1], D. Kalra[2], K. Patterson[3], K. Walker[4], S. Kalla[5], J. Smith[6], S. McGee[7], A. Radhakrishnan[8], A. Haddad[9], A. Musick[10], J. Karnes[11], P. Empey[9], G. Jarvik[12], R. Gibbs[2], The All of Us Research Program; [1]HGSC - Baylor Coll. of Med., Houston, TX, [2]Baylor Coll. Med., Houston, TX, [3]Univ. of Washington, Seattle, WA, [4]Baylor Coll. of Med. HGSC, Houston, TX, [5]Baylor Coll. of Med., Houston, TX, [6]Univ of Washington, Seattle, WA, [7]Univ. of Washington Genome Sci., Seattle, WA, [8]Dept. of Genome Sci., Univ. of Washington, Seattle, WA, [9]Univ. of Pittsburgh, Pittsburgh, PA, [10]All of Us Res. Program, NIH, Bethesda, MD, [11]Univ. of Arizona, Tucson, AZ, [12]Univ Washington Med Ctr., Seattle, WA

Abstract:

The All of Us Research Program has placed a strategic emphasis on the recruitment of individuals from a wide-range of demographic backgrounds. We employed our database of curated variants to analyze differences in the prevalence of recognized pathogenic variants across various predicted genetic ancestry groups. Recognizing disparities in pathogenic variant frequencies related to ancestry could highlight gaps in current genetic understanding, thereby underscoring the significance of the diverse recruitment approach employed by the All of Us Research Program.

We previously analyzed the v6 and release of the dataset, observing that even excluding HFE, which has a very large known effect, these rates differ significantly between European, African and Admixed American/Latino groups (Chi-square: $p < 0.001$) and are largely driven by divergent pathogenic variant (PV) rates in APOB ($p = 0.0001$) and PALB2 ($p = 0.004$). The v7 release of the All of Us Research Program's controlled tier data includes whole genome sequences (WGS) from 245,394 participants, of predominantly European (50%), African (22%), American Admixed/Latino (17%), East Asian (2%), South Asian (1.0%), Middle Eastern (0.2%) or 'Other' (8%) ancestry. The PALB2 divergence was replicated in the v7 analysis.

We additionally assessed known pathogenic variants and rare, predicted loss of function variants (pLoF) in a set of 525 genes known to cause recessive disease, including 288 where pLoF is a known disease mechanism. We observed a high rate of carrying PVs across the cohort, ranging from a rate of 1.02 variants per individual in the European ancestry population to 0.66 in the South Asian population. We observe multiple genes enriched or depleted for carrying PVs relative to the European ancestry population (Bonferroni corrected $p < 0.000016$), including F5, HBB, SERPINA1, DHCR7 and MEFV.

We have examined VUS prevalence in this dataset, observing an inverse correlation with the rate of pathogenic variants in some ancestry groups. Comparing the variance for pLoF variants to all other variant types across ancestry groups shows reduced pLoF variance, pointing to ascertainment bias of variant knowledge in public databases.

Ancestry-linked differences reveal areas of current health disparities and directs future work aimed at reducing those disparities, which will be a lasting legacy of the AoURP.

Title: Misconceptions and limitations of video education: Barriers to genetic testing in Black or Spanish-language patients with cancer.

Authors:

F. Barquet Ramos[1], K. Kilbridge[1], S. McGraw[2], J. E. Stopfer[1], A. T. Slack[1], B. Jefferies[3], M. Brule[4], E. Lizardo[3], S. Grumet[3], H. W. Pepprock[1], S. Hussein[4], A. Husband[1], J. E. Garber[1], P. Sanz-Altamira[4], D. Toppmeyer[3], H. Q. Rana[1]; [1]Dana-Farber Cancer Inst., Boston, MA, [2]MNW Consulting Group, Portland, OR, [3]Rutgers/Cancer Inst. of New Jersey, New Brunswick, NJ, [4]Dana-Farber Cancer Inst., Methuen, MA

Abstract:

Background: Pre-test genetic education for patients with cancer is increasingly delivered through alternative models, including video education (VE) to improve access to germline genetic testing (GT). Prior research on the application of VE models has been primarily conducted with white patients. Feedback from diverse patients is needed to improve the accessibility of VE. **Methods:** We conducted structured interviews with patients at Dana-Farber Cancer Institute and Rutgers Cancer Institute of New Jersey who had a cancer diagnosis and no prior GT. Patients either identified as black or Hispanic/Latinx who preferred to use Spanish. Using open-ended questions, participants described their thoughts about the veracity of 8 items from the Hereditary Cancer Multidimensional Measure of Informed Choices (MMIC) to gauge their knowledge of cancer genetics. Following this, they viewed and commented on a video explaining major components of pre-GT education. Health literacy was assessed using the Rapid Estimate of Adult Literacy in Medicine-Short Form for English-speaking and the Short Assessment of Health Literacy - Spanish for Spanish-speaking participants. **Results:** Among the 20 participants, 11 identified as black, and 9 were Spanish speakers. Participants were divided between adequate and low levels of literacy. Low literacy levels were identified among 36% (n=4) of English-speakers and 22% (N=2) of Spanish speakers. Analyses uncovered several themes that suggest participants had misconceptions about the purpose and benefits of GT. Their discussion of the MMIC items, prior to the VE, suggested that 8 (40%) participants were confused about the difference between GT and diagnosis, with higher rates among participants with low literacy. After the VE, nearly half of the participants did not understand the meaning of the results from GT, confounding positive or negative results with a cancer diagnosis (N=9, 45%). This misconception was more prevalent among those with low literacy: 4 of the 6 (67%) low literacy participants compared to 5 of the 14 (36%) adequate literacy participants. Following the VE, participants recognized some benefits of GT, including providing risk information for relatives (N=11, 55%) and guiding medical care (N=7, 35%). Overall, 14 of 20 participants (70%) liked the VE and would recommend it to others. **Conclusion:** This study suggests the importance of addressing confusion about the purpose of cancer GT among underrepresented populations. Our findings suggest VE can be one approach to enhancing understanding of GT. Future efforts are focused on modifying the video to increase clarity and better address misconceptions in a randomized trial.

Title: Landscape of diversity in the genomics workforce: a review and analysis of relevant literature

Authors:

T. Mosley[1], T. Gaye[2,3], L. Hindorff[4], V. L. Bonham[2,4]; [1]Scientific Workforce Diversity Office, NIH, Bethesda, MD, [2]Social and Behavioral Res. Branch, Natl. Human Genome Res. Inst., NIH, Bethesda, MD, [3]Eastern Virginia Med. Sch., Norfolk, VA, [4]Training, Diversity, and Hlth.Equity Office, Natl. Human Genome Res. Inst., NIH, Bethesda, MD

Abstract:

The genetics and genomics workforce falls short of reflecting the nation's diverse population. This lack of diversity limits performance, creativity, and outcomes of success in the scientific enterprise. As efforts to advance diversity and inclusion heighten across the nation, new and creative strategies and programs are required to address current gaps in efforts to diversify the genomics workforce and create an inclusive environment. We sought to understand and describe the current landscape of diversity within the genomics workforce. We performed a systematic literature search to curate relevant literature related to four key themes informed by the NHGRI Diversity in Genomics Workforce Initiative: 1) the importance of diversity; 2) overall knowledge or state of diversity within the biomedical workforce generally and 3) genomics workforce specifically; and 4) successful programs that have sought to increase the diversity of these workforces. Searches were constructed for five literature databases were searched using a combination of over 20 search terms such as "diversity", "underrepresentation", "genomic workforce", "biomedical workforce", "minority-serving institutions" and more to identify peer-reviewed articles published between 2000 - 2023. Using a directed content analysis approach, we will present key findings from the literature search related to the four key themes. The work presented will review diversity and diversity-focused programs within the genomics and biomedical workforce and provide recommendations for addressing current gaps in data, policies, and strategies needed to diversify the field of genomics. This study provides foundational data for advancing diversity and inclusion in the genomics research workforce.

Session 016: Advancing understanding of mechanisms in Alzheimer's Disease

Location: Conv Ctr/Room 207A/Level 2

Session Time: Thursday, November 2, 2023, 10:45 am - 12:15 pm

Title: Cell-type specific non-coding rare variant association tests reveal novel Alzheimer's disease relevant functional annotations and loci.

Authors:

A. Das[1,2], C. Terwagne[3], C. M. Lakhani[2], J-S. Lin[2], T. Raj[4], D. A. Knowles[1,2]; [1]Columbia Univ., New York, NY, [2]New York Genome Ctr., New York, NY, [3]The Francis Crick Inst., London, United Kingdom, [4]Icahn Sch. of Med. at Mount Sinai, New York, NY

Abstract:

Alzheimer's disease (AD) is a complex neurodegenerative disorder, and understanding the role of rare variants in disease development will further unravel its complex etiology. Multiple gene-based rare variant association tests exist to analyze the relationship between genotype and phenotype for complex diseases, but with the availability of improved coding and noncoding variant-level functional annotations, integrative analyses that prioritize relevant variants can enhance the discovery of disease-associated genes, pathways and cell-states. In particular, growing evidence suggests that neuroinflammation and dysfunction of microglia contribute to the development and progression of AD, but existing methods have not had sufficient power to examine rare noncoding effects that incorporate such cell-type specific information. To address this gap, we focus our analysis on rare variants within microglia-specific predicted enhancer and promoter regions calculated by the Activity-by-contact model (*Fulco et al., 2019*). Using the Functional Score Test (FST) method (*He et al., 2017*), we analyze whole genome sequence data from the Alzheimer's Disease Sequencing Project, consisting of 7,865 cases and 13,086 controls, along with 51 functional annotations and 7.9 million non-coding variants in microglia-specific enhancers/promoters. Our analysis yields per-gene burden, dispersion, and combined test statistics for functional annotation and gene significance. We further extend FST to estimate genome-wide functional annotation effects, providing insight into AD-relevant annotations. Incorporating functional annotations, we find 222 genes at genome-wide significance (Bonferroni adjusted $p < 5.2E-13$); this number decreases to 154 genes without them, indicating the value of functional information. We find that significant genes are associated with higher levels of gene expression across 14 brain tissues, and many findings are previously reported AD genes, including *ABCA7* and *SORL1*. From our extension of FST, we identify conservation scores, mappability scores, and population-specific allele frequencies as functionally significant for AD genome-wide. Our study establishes a computational pipeline that incorporates functional annotations of microglia-related rare variants to perform gene-based AD association tests, uncovering both disease-relevant genes and functional annotations. Future directions include analyzing effects within enhancer and promoter regions for other AD-relevant cell types and investigating population-specific rare variant effects.

Title: Haplotype characterization using short and long-read sequencing data of a protective region of segmental duplication for Alzheimer disease in African carriers of *APOE*ε4

Authors:

L. Bertholim Nasciben, K. Nuytemans, M. Lipkin Vasquez, J. Young, D. Dykxhoorn, F. Rajabli, L. Wang, W. Scott, D. Davis, R. Vontell, A. Griswold, M. Pericak-Vance, J. Vance; Univ. of Miami, Miami, FL

Abstract:

APOE ε4 is the strongest genetic risk for Alzheimer Disease (AD), but *APOE* ε4 carriers with African (AF) local ancestry surrounding the *APOE* locus have a lower AD risk compared to other ancestral backgrounds. Our group has shown that the A allele of rs10423769 (located 2MB upstream of *APOE*) reduces the OR for AD risk for *APOE* ε4 carriers from 7.2 to 2.1. The minor allele frequency (MAF) of this protective variant is 0.12 in individuals of AF ancestry vs. 0.003 in EU. This SNP is in a cluster of pregnancy specific beta-1 glycoprotein genes (*PSG*s) on a highly segmentally duplicated region of chr19. *PSG*s are polymorphic in copy number and SNVs and have been reported to be involved in changes in chromatin structure as cells age. While a novel protective factor, the mechanism of mitigating the effect of *APOE* ε4 is not clear. To gain insights into this mechanism, we characterized the haplotype and structural variations (SVs) in the region. We first utilized PLINK and Haploview to construct haplotypes of rs10423769_A allele carriers using short read whole genome sequencing data from the AD Sequencing Project (ADSP) consisting of nearly 36,000 individuals. We then generated long-read whole genome sequencing data (LRS) on the Oxford Nanopore PromethION on 16 individuals (nine heterozygous and 7 homozygous for rs10423769_A). Resulting FASTQs were aligned to GRCh38 and T2T-CHM13 using minimap2. We performed haplotype analysis on 1,962 individuals carrying rs10423769_A allele from ADSP. Using variants with a MAF ≥ 0.05, the resulting haplotype block was approximately 21Kb (Chr19:43099521-43120749). We identified other markers along this haplotype whose frequency is > 0.10 in AF populations and are found at a lower frequency in Latino/Hispanic population, but extremely rare in EU population. LRS yielded an average genome coverage of 18.3 ± 3.1X and read lengths of 10.4 ± 1.8kb using routine DNA extraction methods. High mapping quality of these reads confirmed that the variants making up this haplotype are specific to this region and not in another duplicated region of the genome. LRS data were used to investigate SVs in the region, but no consistent SV was evident at this coverage. Studying this highly segmentally duplicated region near *PSG*s not only will provide insights into the protective factor for *APOE* ε4 but will increase our knowledge of the potential properties of segmental duplications affecting the surrounding genome. We have identified an AF-specific haplotype protective against AD risk conferred by *APOE* ε4. Ongoing analyses including 3D chromatin structure and LRS at higher coverage will further help us to understand the mechanism involved in AD protection.

Title: PWAS with summary-level pQTL reference data of brain, CSF, and plasma tissues identifies 23 risk genes for Alzheimer's disease dementia

Authors:

T. Hu[1], Q. Dai[1,2], M. P. Epstein[1], J. Yang[1]; [1]Ctr. for Computational and Quantitative Genetics, Dept. of Human Genetics, Emory Univ. Sch. of Med., Atlanta, GA, [2]Dept. of Biostatistics and Bioinformatics, Emory Univ. Sch. of Publ. Hlth., Atlanta, GA

Abstract:

Characterizing the genetic mechanisms underlying Alzheimer's disease (AD) dementia is vital for new therapeutics. Proteome-wide association study (PWAS) can identify risk genes of AD dementia with genetic effects mediated through protein abundance. Existing PWAS analyses often train protein abundance imputation models on individual-level reference proteomic and genetic data of a single tissue type and only consider a single training method. We believe a more comprehensive PWAS that considers multiple training methods and is further applicable to summary-level pQTL data available on various tissues would likely improve PWAS power. To this end, we can use the recently published OTTERS tool to train multiple protein abundance imputation models (P-value thresholding, LASSO, Bayesian with continuous shrinkage prior, nonparametric Bayesian Dirichlet process regression) using summary-level pQTL data from different trait-related tissues. For a fitted imputation model per protein, we can then perform PWAS on independent GWAS data of the corresponding protein coding gene, and combine PWAS p-values across multiple models using an aggregated Cauchy association test to yield one omnibus PWAS p-value. We employed OTTERS to conduct PWAS with summary-level pQTL reference data of brain (1079 proteins, n=380), cerebrospinal fluid (CSF, 731 proteins, n = 835), and plasma (931 proteins, n=529) tissues, as well as the most recent GWAS summary data of AD dementia (n=762,917). We further validated significant PWAS risk genes detected by OTTERS using the PMR-Egger tool that can assess causal genetic effects mediated through protein abundance for the phenotype of interest. We identified 23 significant risk genes of AD dementia with FDR < 0.05 (8 with brain, 9 with CSF, and 10 with plasma reference pQTL data), including 4 genes significant with at least two of the tissues and 14 genes identified by previous GWAS, TWAS, or PWAS. We validated that 5 out of 8 risk genes with brain, 8 out of 9 risk genes with CSF, 7 out of 10 risk genes with plasma pQTL data had significant mediation effects (with Bonferroni correction) by PMR-Egger tool. By using the STRING tool, 15 out of 23 PWAS significant genes were found interconnected through a protein-protein interaction network, including the well-known AD risk gene APOE. These significant PWAS risk genes were enriched in important biological pathways involving lipoprotein metabolism and immune response. Our study showed the feasibility of using OTTERS to leverage summary-level pQTL data of multiple tissues for PWAS. Our findings also provide insights into the underlying mechanisms of risk genes for AD dementia.

Title: Single-cell epigenomic changes in Alzheimer's diseases across 1 million cells and six brain regions

Authors:

Z. Liu; Massachusetts Inst. of Technology, Cambridge, MA

Abstract:

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive function and memory loss. However, the underlying epigenetic mechanisms of the disease are still poorly understood. To address this, we generated a single-cell chromatin accessibility landscape from six brain regions of 100 individuals with varying degrees of Alzheimer’s disease pathology. By analyzing the single-cell epigenomic profiles of 1 million cells, we identified six major cell types and more than 30 distinct sub-cell types. We also revealed the cell-type-specific utilization of 359,420 candidate cis-regulatory DNA elements (cCREs) and identified upstream regulators for these cell-type-specific cCREs. Understanding large-scale chromatin structure changes in AD is crucial for uncovering underlying disease mechanisms. Previous studies explored 3D changes at the bulk level, but cell-type-specific 3D genome changes during AD progression remain unclear. To understand this, we computed the active and repressive compartments and observed their dynamics across brain regions and cell types during AD progression. Our findings indicate a global relaxation and smoothened epigenomic state, as well as a loss of chromatin identity and information during AD. Notably, we discovered brain region-specific and cell-type-specific epigenomic erosion, with the most severe erosions observed in the entorhinal cortex and hippocampus regions, as well as activated glial cells. Repetitive elements (REs) comprise a significant portion of the genome and have traditionally been considered as "junk" sequences. However, emerging evidence suggests that these REs can have regulatory roles and contribute to the regulation of gene expression and genome stability. We found that cCREs in brain cell types are also located in some RE families with a cell-type-specific activation pattern. However, the regulatory roles of these active repetitive elements are largely unknown. Thus, we also investigated the chromatin accessibility and regulatory roles of REs in different brain regions and cell types and their dynamics in AD. Moreover, to gain deeper insights into the genetic basis of AD, we performed an integrative analysis that combined GWAS data with information on cCREs in each brain region and cell type. We identified specific genomic regions and brain cell types that are enriched with AD risk loci. Overall, this comprehensive, large-scale, single-cell epigenomic atlas deepens our understanding of the epigenomic regulatory circuits and mechanisms underlying Alzheimer's disease.

Title: Variant-to-function mapping of late-onset Alzheimer's disease GWAS loci in human microglial models implicates *RTFDC1* as an effector gene at the *CASS4* locus

Authors:

E. Burton[1,2], M. Argenziano[3], K. Cook[2], C. Su[2], E. Manduchi[1], K. Hodge[4], L-S. Wang[1], G. Schellenberg[1], J. Pippin[2], A. Wells[2,1], S. Anderson[2], C. Brown[1], S. Grant[2,1], A. Chesi[1]; [1]The Univ. of Pennsylvania, Philadelphia, PA, [2]The Children's Hosp. of Philadelphia, Philadelphia, PA, [3]Univ. of South Florida, Tampa, FL, [4]Emory Univ., Atlanta, GA

Abstract:

Late-onset Alzheimer's disease (LOAD) impacts nearly 6 million Americans over the age of 65. There remains no effective therapy to slow or halt the disease progression. Classically, neurons have been the primary cell type of focus given their production of aggregating amyloid beta ($A\beta_{1\text{-}42}$), a known factor contributing to neurodegeneration. In contrast, recent genetic studies suggest that loci confer their effects via microglia, the resident macrophages of the central nervous system, to prolong inflammation in LOAD brains in the presence of $A\beta_{1\text{-}42}$ therefore exacerbating neurodegeneration. Indeed, recent LOAD GWAS have identified multiple loci in the proximity of genes related to microglial function, such as *TREM2* and *CR1*. However, GWAS itself cannot directly identify causal variants or effector genes, as a GWAS only reports the sentinel variant at a given locus, which serves as a tag for all variants in high linkage disequilibrium. In order to physically map interactions between GWAS-implicated regulatory elements harboring putative causal non-coding variants and their corresponding candidate effector genes in LOAD, we leveraged a combination of ATAC-seq and high-resolution promoter-focused Capture-C in two human microglial cell models, the immortalized HMC3 microglia-like cell line and iPSC-derived iMicroglia. This variant-to-gene mapping strategy in both models identified a novel proxy SNP, rs6024870 ($r^2 = 0.93$ to sentinel SNP rs6014724), at the '*CASS4*' locus, located within the 2nd intron of the *CASS4* gene. This variant resides within a microglia-specific open chromatin region and directly contacts the promoter of *RTFDC1*, a gene not previously implicated in LOAD. Deletion of the 380bp putative enhancer region harboring rs6024870 by lentiviral CRISPR-Cas9 in HMC3 cells reduced the expression of *RTFDC1* at the mRNA level, as measured by RNA-seq, by 20.5% ($P<0.0026$), and immunoblotting showed a downward trend in protein expression of RTFDC1. Gene set enrichment analysis of differentially expressed genes from RNA-seq showed that the top associated up-regulated pathways in enhancer knockout (KO) cells, as compared to wild-type cells, were pro-inflammatory interferon ($FDR_q=0.0$) and interleukin signaling ($FDR_q=0.011$) pathways. Additionally, enhancer KO cells secreted significantly more of the pro-inflammatory cytokines IL-6 ($P<0.05$) and IL-8 ($P<0.01$) than their wild-type counterparts. Together, we have identified a novel LOAD-associated distal microglial enhancer that drives *RTFDC1* gene expression. Our data suggests this mechanism primes the cells to a pro-inflammatory state, consistent with the neuroinflammation seen in LOAD patients.

Title: A *PSEN1* E280A sheep model of Alzheimer's Disease

Authors:

N. Mckean[1], R. Snell[1], H. Zetterberg[2], R. Handley[1], S. Rudiger[3], S. Bawden[3], P. Verma[3], J. Kelly[3], R. Faull[1], S. Reid[4], J. Pearson[5], J. Hardy[6], J. Gusella[7], M. Owen[8]; [1]Univ. of Auckland, Auckland, New Zealand, [2]Univ. of Gothenberg, Gothenburg, Sweden, [3]South Australian Res. and Dev. Inst., Adelaide, Australia, [4]The Univ. of Auckland, Auckland, Auckland, New Zealand, [5]Univ. Otago Christchurch, Christchurch, New Zealand, [6]Inst Neurology, UCL, London, United Kingdom, [7]Massachusetts Gen. Hosp., Boston, MA, [8]Univ Wales Col Med., Cardiff, United Kingdom

Abstract:

Alzheimer's disease (AD) is a devastating neurodegenerative disease. The prevalence is rapidly increasing due to an ageing population worldwide, creating a looming population health crisis. Many rodent models of AD have been created, but none capture the full symptomatology without massively overexpressing multiple human mutations in transgenes. The aim of this project was to create a large animal model of AD with a single mutation in a native gene. Sheep have larger bodies and a gyrencephalic brain similar to human. Importantly, they naturally develop plaques and tangles as they age. Major AD-related genes, such as *APP*, *PSEN1* and *PSEN2*, and the *APP* cleavage sites which produce the Aβ peptide, are highly conserved between human and sheep. Sheep are naturally fixed for the APOE4 allele associated with late onset AD in humans. They can also be efficiently bred in large numbers and simply maintained in a normal farming situation. JIVET technology also allows oocytes to be harvested from ewe lambs at six weeks of age, which can then be implanted into mature recipient ewes, drastically shortening generation times. The gene edited sheep were produced by injection of CRISPR-Cas9 RNP complexes, including a single strand donor DNA carrying the *PSEN1* E280A mutation, into single cell zygotes, which were then implanted. Five founder animals were born carrying the E280A mutation. One ewe is homozygous, two ewes are heterozygous, and two rams are hemizygous with frameshift mutations on their non-E280A allele. The founder lambs were genotyped by PCR and whole genome sequenced to confirm zygosity. No off-target edited was detected. All animals are outwardly healthy and growing normally. Plasma biomarker analysis of the *PSEN1* E280A animals revealed that $A\beta_{1-42}$:$A\beta_{1-40}$ peptide ratios are increased in the mutation carriers as expected. Regression analysis on genotype demonstrated genotype explained over 98% of the variation in the ratios (ANOVA P=1.2e-6). These results have since been replicated. One heterozygous ewe and one ram have been used to create F_1 lines carrying the mutation. The mutation carrying offspring from the ewe lamb have inherited the Aβ phenotype. The offspring from the ram have just been born and will be part of a life history study. If these animals can recapitulate the entire disease from a single gene cause, they will likely shed light on the underlying mechanisms of AD that have proven elusive with small animal models. These animals are also likely to be useful for preclinical pharmaceutical testing.

Session 017: Causal noncoding variants and the genes they impact

Location: Conv Ctr/Ballroom B/Level 3

Session Time: Thursday, November 2, 2023, 10:45 am - 12:15 pm

Title: Functional investigation of pleiotropic effects of an eQTL in the *CELSR2*/*PSRC1*/*SORT1* gene cluster that associates with low density lipoprotein cholesterol levels and resting metabolic rate in an Indigenous population in Arizona

Authors:

K. Freeland, P. Piaggi, M. Traurig, L. J. Baier, **K. Bandesh**; Natl. Inst. of Diabetes and Digestive and Kidney Diseases, Phoenix, AZ

Abstract:

Variation in the *CELSR2-PSRC1-SORT1* locus is a robustly replicated association signal for low density lipoprotein cholesterol (LDL-C) across different ethnic groups with variable diets and eating patterns. Elevated LDL-C is a key risk factor for obesity, a disease that is highly prevalent in an Indigenous population in Arizona. In a recent genome-wide association study of 5,202 indigenous people with fasting plasma LDL-C measures, the strongest association was a *CELSR2* variant rs12740374 ($P= 1 \times 10^{-22}$, 13 mg/dL mean decrease in LDL-C with minor allele genotype) which has a known role in modulating lipid levels and is an expression-QTL (eQTL) for *CELSR2* in liver, a key organ for LDL synthesis. Rs12740374 also showed an unexpected and novel association with reduced resting metabolic rate (RMR) in a subset of individuals who were clinically characterized for obesity (N= 509; P= 0.003; effect= -44.7 kcal/day after adjustment for body composition). Consistent with RMR association, rs12740374 is also a highly significant eQTL for *CELSR2* in skeletal muscle ($P= 2.3 \times 10^{-76}$, GTEx portal), and skeletal muscle mass is a major determinant of resting energy metabolism. We speculated that this pleiotropic effect of rs12740374 on multiple metabolic traits maybe due to a tagged functional variant. Therefore, we searched for additional variants in the locus and identified rs6670347 (T/C), which is in high linkage disequilibrium with rs12740374 in our indigenous population ($r^2= 0.98$) but an independent variant in other populations. Rs6670347 is similarly associated with LDL and RMR and is located in the core binding motif for the glucocorticoid (GC) receptor NR3C1 situated in a credible *CELSR2* enhancer. Whole transcriptome eQTL analysis of rs6670347 in skeletal muscle biopsies obtained from 202 indigenous people identified *CELSR2* as the most differentially expressed gene where the C allele (a minor allele, frequency= 0.2) correlated with higher *CELSR2* expression ($P= 1.9 \times 10^{-7}$). *CELSR2* is a G-protein coupled receptor associated with GC resistance. GCs primarily signal through NR3C1; their prolonged exposure induces lipid accumulation in skeletal muscle and increased levels of lipids may impair muscle oxidative capacity, thereby influencing the energy expenditure. We are currently investigating the impact of rs6670347 on *CELSR2* expression in skeletal muscle cells in response to GC treatments. In conclusion, we propose that variants in the *CELSR2/PSRC1/SORT1* locus have tissue-specific effects on multiple metabolic traits. In addition to the well-characterized role in the liver to influence LDL-C, variants may have an independent role on muscle metabolism via GC signaling.

Title: Comprehensive identification of coronary artery disease-associated variants regulating expression in human vascular smooth muscle cells.

Authors:

N. Barbera, L. Wallace, N. Perry, M. Civelek; Univ. of Virginia, Charlottesville, VA

Abstract:

BACKGROUND: Coronary artery disease (CAD) is a complex disorder with genetic and environmental influences. Genome-wide association studies (GWAS) have identified >300 loci associated with disease risk. Vascular smooth muscle cells (SMCs) play a critical role in the initiation and progression of atherosclerosis, the precursor to coronary artery disease. As part of this progression, SMCs undergo phenotypic changes mediated by changes in gene expression. We previously isolated SMCs from the ascending aortas of a cohort of 151 human heart transplant donors and found a subset of CAD-associated loci that regulate phenotypic changes relevant to the progression of atherosclerosis by altering gene expression in SMCs. However while these loci were identified, the specific causal SNPs within these loci remain unknown.METHODS: We performed a series of orthogonal experimental approaches on our primary SMCs to identify CAD-associated SNPs regulating SMC gene expression in an allele-specific manner: (1) lentivirus-based massively parallel reporter assays (lentiMPRAs) and (2) Cleavage Under Targets & Release Using Nuclease (CUT&RUN) assays in donor-matched SMCs along with (3) assay for transposase-accessible chromatin with sequencing (ATAC-seq). We integrated the results of these three experiments and compared them with our previously obtained expression quantitative trait loci (eQTL) data to identify CAD-relevant variant-gene pairs in SMCs. Finally we used CRISPRi experiments to confirm their allele-specificity of the identified SNPs. RESULTS: We identified dozens of CAD-associated SNPs within functionally relevant regions of SMCs showing allele-specific expression in our lentiMPRAs. By comparing these identified variants with our previously obtained expression quantitative trait loci (eQTL) data, we identified 10 putative disease-relevant variant-gene pairs, which were confirmed with CRISPRi experiments.SUMMARY: By combining multiple orthogonal experimental approaches, we found specific CAD-associated SNPs which regulate SMC genes in an allele-specific manner. Our results further characterize the causal variants that confer increased risk of CAD through their effects on vascular smooth muscle cells by identifying and validating previously unknown specific SNP-gene pairs.

Title: *MAPT* expression is mediated by long-range interactions with *cis*-regulatory elements

Authors:

J. Cochran, B. B. Rogers, A. G. Anderson, S. N. Lauzon, M. Davis, R. M. Hauser, S. C. Roberts, I. Rodriguez-Nunez, K. Trausch-Lowther, E. A. Barinaga, P. I. Hall, J. W. Taylor, M. Mackiewicz, B. S. Roberts, S. J. Cooper, L. F. Rizzardi, R. M. Myers; HudsonAlpha Inst. for Biotechnology, Huntsville, AL

Abstract:

Background: Tauopathies are a group of neurodegenerative diseases driven by abnormal aggregates of tau, a microtubule associated protein encoded by the *MAPT* gene. *MAPT* expression is absent in neural progenitor cells (NPCs) and increases during differentiation. This temporally dynamic expression pattern suggests that *MAPT* expression is controlled by transcription factors and *cis*-regulatory elements specific to differentiated cell types. Given the relevance of *MAPT* expression to neurodegeneration pathogenesis, identification of such elements is relevant to understanding disease risk and pathogenesis.
Methods: We performed HiC, chromatin conformation capture (Capture-C), single-nucleus multiomics (RNA-seq+ATAC-seq), bulk ATAC-seq, and ChIP-seq for H3K27Ac and CTCF in NPCs and neurons differentiated from human iPSC cultures. We nominated candidate *cis*-regulatory elements (cCREs) for MAPT in human NPCs, differentiated neurons, and pure cultures of inhibitory and excitatory neurons. We then assayed these cCREs using luciferase assays and CRISPR interference (CRISPRi) experiments to measure their effects on *MAPT* expression. Finally, we integrated cCRE annotations into an analysis of genetic variation in AD cases and controls.
Results: Using orthogonal genomics approaches, we nominated 94 cCREs for *MAPT*, including the identification of cCREs specifically active in differentiated neurons. Eleven regions enhanced reporter gene transcription in luciferase assays. Using CRISPRi, 5 of the 94 regions tested were identified as necessary for *MAPT* expression as measured by RT-qPCR and RNA-seq. Rare and predicted damaging genetic variation in both nominated and confirmed CREs was depleted in AD cases relative to controls (OR = 0.40, $p = 0.004$), consistent with the hypothesis that variants that disrupt *MAPT* enhancer activity, and thereby reduce *MAPT* expression, may be protective against neurodegenerative disease.
Conclusions: We identified both proximal and distal regulatory elements for *MAPT* and confirmed the regulatory function for several regions, including three regions centromeric to *MAPT* beyond the well-described H1/H2 haplotype inversion breakpoint. This study provides compelling evidence for pursuing detailed knowledge of CREs for genes of interest to permit better understanding of disease risk.

Title: Trisomy 21 reshapes enhancer-gene maps to influence cellular phenotypes

Authors:

A. Marderstein[1], M. De Zuani[2], H. Xue[2], J. Bezney[1], S. Wong[2], S. Montgomery[1], A. Cvejic[3]; [1]Stanford Univ., Palo Alto, CA, [2]Univ. of Cambridge, Cambridge, United Kingdom, [3]Univ. of Copenhagen, Copenhagen, Denmark

Abstract:

Introduction: While enhancer-gene (E-G) links are critical to cellular phenotype and provide an important link between non-coding variants and causal genes, it is unclear how they are modified by genetic background. Down Syndrome (inborn trisomy 21; Ts21) presents a unique model for understanding the genetic influence over E-G regulation and downstream phenotypic impacts because the genetic "variant" is known (Ts21), the molecular impact is large (fetal hematopoiesis is defective, driving 150x higher risk for childhood leukemia and red blood cell (RBC) overproduction), and the critical cellular context has been identified (fetal liver hematopoietic stem cells (HSCs)).
Methods: We developed a computational framework combining GWAS variants, 10X multiome data of 56,890 hematopoietic CD45+ cells, and 1.1 million scRNA-seq cells from fetal liver in Ts21 and controls to study how Ts21 alters E-G maps and its phenotypic consequences. To identify E-G links, we used a Poisson regression model with bootstrapped errors to correlate enhancer accessibility with gene expression in Ts21 HSCs. At identified E-G links, we tested whether Ts21 modifies the effect of enhancer accessibility on expression by including an accessibility-by-trisomy interaction term. Finally, we intersected fine-mapped RBC GWAS SNPs with E-G links and Ts21 differentially expressed genes to pinpoint key enhancers and genes in differentiation.
Results: We identified 4.1-times more E-G links in Ts21 HSCs compared to control HSCs, with 36.1% of those in controls also found in Ts21. Nearly all E-G links specific to control HSCs had a significant interaction term, indicating widespread loss of E-G links in the Ts21 background. In contrast, only 11.2% of Ts21-specific enhancer-gene links had a significant interaction term that suggest trisomy strengthens an E-G connection. Instead, we found greater activation of regulatory elements in Ts21 HSCs that makes these E-G links discoverable, as the remaining 88.8% of Ts21 links were significantly more accessible (2.1-fold) and upregulated (1.4-fold) in Ts21 compared to control. Finally, we found that RBC GWAS-harboring enhancers were (1) more accessible in a subpopulation of Ts21 HSCs, (2) enriched for association with gene expression, and (3) linked to target genes differentially expressed between Ts21 and control HSCs ($P<0.05$). This suggests a key role for E-G links in biasing differentiation of Ts21 HSCs towards the erythroid lineage, which we validated in our scRNA-seq and using *in vitro* assays.
Conclusion: Integrating single-cell multi-omics and GWAS illuminates how Ts21 reshapes gene regulation by modifying E-G links to impact blood development.

Title: Dissecting the role of pleiotropy in the genetic adaptation of Tibetans to high altitudes.

Authors:

A. Thornburg, S. Park, L. Zhang, I. Aneas, D. R. Sobreira, I. M. Salamone, N. J. Sakabe, K. M. Farris, Z. T. Weber, O. A. Gray, J. Yoo, A. Di Rienzo, M. A. Nobrega; Univ. of Chicago, Chicago, IL

Abstract:

In a well-established example of genetic adaptation in humans to local environments, Tibetans have undergone positive selection due to hypoxic stress at high altitudes. This involved a selective sweep at the *EPAS1* locus, encoding a subunit of the hypoxic inducible factor pathway. We previously identified an *EPAS1* enhancer (ENH5) harboring variants strongly associated with high altitude in Tibetans. We showed that this enhancer is evolutionarily and functionally conserved in mice, and generated ENH5 knockout mice (ENH5-KO) to model the effects of ENH5 *in vivo.* We showed that ENH5 deletion in mice leads to differences in gene expression in several tissues, suggestive of adaptive pleiotropy as a mechanism behind the selection signals. ENH5 is also active in adipose tissue, suggesting a yet to be characterized physiological role in this tissue. To dissect this role, we assayed adipose biology phenotypes in wild-type (WT) and ENH5-KO mice. We initially tested for adiposity-related and metabolic phenotypes, but did not detect any differences between WT and ENH5-KO mice. We next studied the effects of hypoxia on adipocytes from these mice. We found ENH5-KO preadipocytes differentially express genes encoding the mitochondrial electron transport chain subunits in response to hypoxia, a key pathway regulating cellular energetics, suggesting that ENH5 has a role in regulating mitochondrial function under hypoxic conditions. Adipocytes also have a role in thermogenesis, where the electron transport chain is uncoupled from energy production to release energy as heat. Since Tibetans experience extreme cold temperatures at high altitudes, we tested if ENH5 has a role in thermogenic response in adipocytes. We found again differential expression of multiple genes encoding the electron transport chain subunits. Our results raise the possibility that the pleiotropic effects of ENH5 may implicate unforeseen mechanisms, such as cellular energetics and thermogenesis, possibly contributing to the phenotypic adaptation to high altitude in Tibetans.

Title: DNA looping linked GWAS variants to target genes and orthogonal integrative functional analyses substantiated*RAPGEF1*role in melanoma susceptibility

Authors:

R. Thakur[1], M. Xu[1], A. Thornock[2], H. Sowards[1], E. Long[3], T. Rehling[1], R. Hennessey[1], K. Funderburk[1], J. Yin[1], R. Chari[4], T. Zhang[1], L. Jessop[1], T. Myers[1], M. Johnson[5], A. Wells[5], A. Chesi[5], S. Grant[5], M. Iles[6], M. Landi[1], M. Law[7], Melanoma Meta-Analysis Consortium, M. Machiela[1], J. Choi[1], L. Zon[2], K. Brown[1]; [1]Natl. Cancer Inst., Rockville, MD, [2]Harvard Dept. of Stem Cell and Regenerative Biology, Boston, MA, [3]Chinese Academy of Med. Sci. and Peking Union Med. Coll., Beijing, China, [4]Natl. Cancer Inst., Frederick, MD, [5]Children's Hosp. of Philadelphia Res. Inst. and Univ. of Pennsylvania, Philadelphia, PA, [6]Univ. of Leeds, Leeds, United Kingdom, [7]QIMR Berghofer Med. Res. Inst., Brisbane, Australia

Abstract:

Background: DNA looping allows distant regulatory elements such as enhancers to come in proximity with gene promoters, facilitating gene expression regulation. We hypothesized that mapping DNA looping at GWAS loci could link GWAS risk variants to target genes, and integrative bioinformatic analyses and functional experiments can clarify the role of target genes in disease susceptibility.
Methods: We mapped DNA looping through region-focused chromatin conformation capture assays (**CaptureHiC**) in human melanocytes, covering all 68 cutaneous melanoma (**CM**) loci (Landi *et al.* 2020). In addition to DNA looping, we performed integrative analyses with ATAC-seq, massive-parallel reporter assay (**MPRA**), gene expression, and publicly available ROADMAP imputed chromHMM datasets to prioritize candidate causal genes (**CCG**) for functional experiments.
Results: CaptureHiC looping presented significant improvement in CCG identification at the CM GWAS loci (61 of 68 of loci (90%) had a CCG nominated) over previous melanocyte-specific QTL/TWAS/MWAS approaches. In addition, integrative analysis of DNA looping with ATAC-seq and chromHMM datasets showed loops from at least one variant in accessible chromatin or enhancer region to a CCG promoter at 82% of CM GWAS loci. Subsequent analysis of these regulatory variants with the MPRA data revealed 72% of the CM loci harbored at least one variant with significant allele-specific effect on gene expression. Interestingly, at the *9q34.13* locus we observed DNA looping between risk variants located near the three prime end of Rap Guanine Nucleotide Exchange Factor 1 (***RAPGEF1***) gene to the *RAPGEF1* and *UCK1* promoters. The risk allele of the lead variant (rs3780269) variant was associated with increased *RAPGEF1* mRNA expression and not *UCK1* expression in melanocytes, further supporting *RAPGEF1* as the target gene. We performed CRISPRi to validate the regulatory effect of the region containing the lead variant. Repressing the lead variant region led to decreased *RAPGEF1* mRNA expression. We performed CRISPR knockout proliferation screen in melanocytes and identified *RAPGEF1* as a key regulator of melanocyte growth and survival. To confirm these findings with an orthogonal approach, we overexpressed RAPGEF1 in melanocytes and observed increased melanocyte growth. Currently, we are assessing the role of *RAPGEF1* in melanoma tumor development in a zebrafish model.
Conclusion: To summarize, CaptureHiC DNA looping linked GWAS risk variants to target genes, while integrative bioinformatic analyses followed by functional experiments identified *RAPGEF1* as novel candidate melanoma susceptibility gene.

Session 018: Chromatin connection to health and disease

Location: Conv Ctr/Room 202A/Level 2

Session Time: Thursday, November 2, 2023, 10:45 am - 12:15 pm

Title: Enhanced detection of chromatin QTLs by MultiCAQ empowers GWAS variants interpretability.

Authors:

L. Wang, B. Fair, C. Buen Abad Najar, X. Liu; Univ. of Chicago, Chicago, IL

Abstract:

Regulatory variants that modulate chromatin state play an important role in disease variations. They facilitate comprehensive understanding of disease mechanisms through context-specific gene regulations. However, detecting genetic variants associated with chromatin states (cQTLs) is nontrivial due to inaccurate peak calling and limited power at current sample size. Only a low proportion of epigenetic peaks (<20%) have been associated with cQTLs. To address the challenges, we propose MultiCAQ, a multivariate chromatin QTL detection method, that leverages co-regulation of nearby chromatin states to improve power of cQTL detection. MultiCAQ-**P** groups multiple adjacent **p**eaks into a window and employs a powerful PC-based multivariate method, PCO, to test association between each variant and the window. We further develop MultiCAP-**S**, which skips peak calling and divides the genome into fixed-width **s**egments, to directly test association between variants and multiple segments grouped in a window.We found that MultiCAQ significantly outperforms the standard univariate cQTL detection approach. We applied MultiCAQ-P to ChIP-Seq data of three histone marks: H3K27ac, H3K4me1, and H3K4me3 in lymphoblastoid cell lines (LCL) from 75 Yoruba individuals. For H3K27ac, the standard univariate method identified 4967 cPeaks, and 90% were detected by MultiCAQ at the window size of 50kb. MultiCAQ additionally detected 4005 windows with at least one significant cQTL that were missed by the standard univariate method. We found the novel cQTLs detected by MultiCAQ are significantly enriched for eQTLs in blood (1.37-fold-enriched, p-value=1.8e-12, one-sided Fisher's exact test), indicating their role in regulating gene expression variations. To investigate the role of these cQTLs in complex trait genetics, we performed colocalization analyses of cQTLs and GWAS loci from 36 blood traits. We observed that 375 (6%) GWAS loci colocalized with cQTLs, with 43 additionally found by novel cQTLs. Interestingly, 152 colocalized GWAS loci did not overlap with any eQTLs, suggesting the newly discovered cQTLs can offer additional insights other than expression level changes into explaining trait mechanisms. In conclusion, MultiCAQ is a powerful and reliable framework that detects cQTLs and effectively addresses inaccurate peak calling, enhancing our ability to interpret GWAS variants and the underlying molecular mechanism.

Title: Machine learning of open chromatin regions enables prediction of constrained regions and prioritization of putative causal variants in a cell type-specific manner.

Authors:

K. Tomizuka[1], M. Koido[1,2], A. Suzuki[1], S. Yoshino[1], N. Tanaka[1], S. Koyama[1], K. Ishigaki[1], Y. Murakawa[1], Immune transcript/enhancer consortium(ITEC), K. Yamamoto[1], C. Terao[1,3,4]; [1]RIKEN Ctr. for Integrative Med. Sci., Yokohama, Kanagawa, Japan, [2]The Univ. of Tokyo, Tokyo, Japan, [3]Shizuoka Gen. Hosp., Shizuoka, Japan, [4]Univ. of Shizuoka, Shizuoka, Japan

Abstract:

Open chromatin regions (OCRs) mark regulatory regions but are not always active and functional, hampering the prioritization of variants identified in genome-wide association studies. Here we developed a new method, CAMBUS (Chromatin Accessibility Mutation Burden Score), to prioritize functional OCRs and their overlapping variants. First, leveraging median of 253k OCRs in 29 immune cell types from 50 Japanese individuals, we developed machine learning models to predict OCRs in each cell type from reference genome sequence only. These models could accurately predict OCRs (median value of the area under the receiver operating characteristic curve: 0.87) including cell type-specific OCRs and in silico mutagenesis analysis by these models showed > 90% directional concordance of mutation effects to the known chromatin accessible quantitative trait loci. Given the outstanding performance of our models, we derived CAMBUS for each OCR by aggregating in silico mutation effects of all surrounding variants within 1kb. We defined CAMBUS positive OCRs by low CAMBUS score which suggesting intolerant to mutations based on chromatin openness (N=6.6k in total and median 1.6k for each cell-type). We observed significant enrichment of CAMBUS positive OCRs in known constrained regions [gnomAD constrained z-score > 4; Odds Ratio (OR) = 11.45, $P = 4.7\times10^{-68}$]. While CAMBUS positives over all 29 immune cell types were strongly enriched in candidate cis-regulatory element (cCRE)-promoter-like signatures in ENCODE3 (OR = 8.79, $P < 1\times10^{-300}$), cell-type-specific CAMBUS positives were rather enriched in active enhancers like super enhancers (OR = 9.39, $P = 4.9\times10^{-103}$) and FANTOM5 enhancers (OR = 5.46, $P = 3.4\text{x}10^{-106}$). Finally, we prioritized candidates of causal variants for complex traits by exploring overlapping with CAMBUS positives. The prioritized variants were significantly enriched in 95% credible sets for complex traits in UK Biobank (94 traits) and Biobank Japan (62 traits) compared to all 253k OCRs (median OR: 4.1 and 4.4, median P: $8.0\text{x}10^{-5}$ and $9.7\text{x}10^{-5}$ respectively). As an example, rs8009224, a candidate causal variant for the Hemoglobin A1c levels, overlapped with a CAMBUS positive only in plasmablast. This variant and CAMBUS positive were also selectively highlighted as enhancer region by the rule-based activity-by-contact model (ChIP-seq × Hi-C) for immune-related cells. In summary, we demonstrate that CAMBUS positives can mark active, constrained regulatory regions from only cell-type specific ATAC-seq data with machine learning and prioritize causal variants for complex traits.

Title: Genetic mechanisms at scale: Mapping chromatin dysregulation induced by hundreds of protein-coding variants simultaneously

Authors:

M. Frenkel, Z. Morris, S. Raman; Univ. of Wisconsin - Madison, Madison, WI

Abstract:

Advances in DNA sequencing have allowed geneticists to catalog millions of variants and thousands of disease associations throughout the human genome. While tools exist for high-throughput DNA *reading,* we lack methods for high-throughput *understanding* of genetic mechanisms. Existing high-throughput methods for studying genetic variation *in vitro* achieve scale by sacrificing resolution: collapsing variant effects onto single reporters or highly abstract phenotypes. This is inadequate for understanding genetic variants that rewire whole genomes. On the other hand, high-resolution genomic measurements can resolve complex mechanistic differences but are relegated to profiling one variant at a time, which doesn't scale to the number of variants that demand interpretation. Here we create a pipeline to meet both of these needs which can be used to discover how hundreds of causal variants affect genome-wide epigenetic regulation simultaneously.
We sought to create a method for performing hundreds of reverse-genetic experiments with high resolution readouts in a single flask. To do this, we modified 10X Genomics single-cell ATAC sequencing to capture perturbation identity alongside chromatin accessibility thereby resolving the epigenetic consequences of individual variants from a heterogenous library. In one experiment we map how each variant from the pool affects global chromatin state. We applied this method to the currently intractable problem of identifying cell-state dysregulation caused by hundreds of putative oncofusion proteins. Oncofusions are chimeric proteins which contain two fused coding sequences from disparate proteins, and most have no known mechanism. To address this, we expressed a library of ~110 fusions and controls in HEK293T cells and determined how each pioneers chromatin. We recapitulate known mechanisms of oncogenesis across many cancers including EWSR1-FLI1 in Ewing sarcoma, EWSR1-ATF1 in clear cell sarcoma, and PAX3-FOXO1 in rhabdomyosarcoma validating our approach. We reveal groups of fusions with common underlying biology such as ETV6-NTRK3, CCDC6-RET, and FGFR3-TACC3 all creating a similar *de novo* state likely via constitutive kinase activity. We also distinguish fusions with gain of function pioneer abilities not attributable to either component domain alone. Our pipeline is disease- and gene-agnostic and can be used to learn epigenetic dysregulation caused by any set of protein-coding variants, classify variants based on complex mechanisms, and identify possible genotype-informed drug targets. We believe this method can be used to rapidly reveal genetic mechanisms at scale even for rare, orphan variants.

Title: Development of a catalog of structural variants that significantly disrupt TAD boundaries in the human genome and affect gene expression and splicing.

Authors:

C. Li[1], M. J. Bonder[2], S. Syed[3], M. Jensen[4], Human Genome Structural Variation Consortium (HGSVC), HGSVC Functional Analysis Working Group, M. Gerstein[5], M. Zody[6], M. Chaisson[7], M. Talkowski[8], T. Marschall[9], J. Korbel[10], E. Eichler[11], C. Lee[3], X. Shi[1]; [1]Temple Univ., Philadelphia, PA, [2]German Cancer Res. Ctr., Heidelberg, Germany, [3]The Jackson Lab. for Genomic Med., Farmington, CT, [4]Yale Unversity, New Haven, CT, [5]Yale Univ, New Haven, CT, [6]New York Genome Ctr., New York, NY, [7]Univ. of Southern California, Los Angeles, CA, [8]Massachusetts Gen. Hosp., Boston, MA, [9]Heinrich Heine Univ., Düsseldorf, Germany, [10]European Molecular Biology Lab., Heidelberg, Germany, [11]Univ. of Washington, Seattle, WA

Abstract:

Functional annotation of genomes plays a critical role in unraveling the complicated genotype-to-phenotype relationship in humans. A key aspect of this annotation involves understanding how the spatial organization of DNA within the nucleus influences genome functionality and gene regulation. Techniques like Hi-C have been extensively employed to investigate the 3D chromatin structure, including identifying topologically associating domains (TADs) which represent stable genomic regions demarcated by insulating proteins like CCCTC-binding factor (CTCF), to restrict chromatin contacts between regulatory elements and genes. Meanwhile, TAD boundaries, which separate adjacent TADs, exhibit high conservation across cell types and are more evolutionarily constrained than the TADs themselves. Studies have shown that genetic variants like structural variants (SVs) can disrupt TADs, leading to changes in gene expression levels and disease development.
This study presents an integrative Hi-C analysis pipeline that generated a comprehensive catalog of TADs and TAD boundaries in 44 human genomes representing five super populations from the 1000 Genomes Project. This Hi-C data collection consists of 38 billion sequenced read pairs and 23 billion contacts that quantify the intensity of interactions between any two genomic regions. With this data, we have generated an unprecedented high-resolution contact map and a comprehensive TAD catalog in human genomes containing 14,612 TAD boundaries, 18,972 TADs, and 6,819 sub-TADs, where 2,121 TADs and 172 sub-TADs have not been previously reported.
To assess how SVs disrupt the structure of TADs, we identified 430 SVs that significantly disrupted TAD boundaries (TAD-SVs). Out of these 430 TAD-SVs, we discovered that 39 of them overlap with previously reported SV-eQTLs (i.e., SVs that are significantly associated with changes in gene expression profiles) and 19 TAD-SVs overlap with known SV-sQTLs (i.e., SVs that are significantly associated with changes in gene splicing profiles). Furthermore, by leveraging the registry of candidate cis-Regulatory Elements (cCREs) released by the ENCODE Project, we discovered that 71 out of the 430 TAD-SVs overlap with cis-regulatory elements such as promoter-like signature, enhancer-like signature, DNase-H3K4me3, and CTCF.
In summary, this study provides a detailed map of the 3D genome structure in 44 humans and offers an invaluable resource for evaluating the impact of SVs on the 3D chromatin architecture and gene regulation.

Title: Genetic regulation of cell-type specific chromatin accessibility shapes the etiology of brain diseases

Authors:

B. Zeng[1], J. Bendl[2], C. Deng[3], D. Lee[4], R. Misir[1], S. M. Reach[1], S. Kleopoulos[1], P. Auluck[5], S. Marenco[5], D. A. Lewis[6], V. Haroutunian[1], N. Ahituv[7], J. Fullard[1], G. Hoffman[4], P. Roussos[4]; [1]Mount Sinai, New York, NY, [2]Icahn Sch. of Med. at Mount Sinai, New York City, NY, [3]Univ. of California, San Francisco, San Francisco, CA, [4]Icahn Sch. of Med. at Mount Sinai, New York, NY, [5]Natl. Inst. of Mental Hlth.-Intramural Res. Program, Bethesda, MD, [6]Univ. of Pittsburgh, Pittsburgh, PA, [7]UCSF, San Francisco, CA

Abstract:

Hundreds of loci associated with diseases of the human brain have been identified by genome-wide association studies (GWAS), and these associations are enriched in cell type specific regulatory elements thought to drive variation in gene expression. Understanding the contribution of genetic regulation to the quantitative variation in chromatin accessibility in these regulatory elements can elucidate the molecular mechanisms meditating disease risk from genetic variant to disease phenotype. Here we perform ATAC-seq to measure cell type specific chromatin accessibility on 1,932 ATAC-seq samples using sorted neurons and glia from 4 brain regions and 643 individuals. We identified 34,539 open chromatin regions with chromatin accessibility quantitative trait loci (caQTL). Only 10.4% of caQTL are shared between neurons and non-neurons, supporting the cell type specificity of genetic regulation of the brain regulome. Incorporating allele specific chromatin accessibility improves statistical fine-mapping and refines molecular mechanisms underlying disease risk. Colocalization analysis of caQTLs with cell type specific QTLs for histone modification and gene expression identifies a high degree of shared effects across these assays. Genes and open chromatin regions with shared genetic regulation are enriched for enhancer-promoter links identified by activity-by-contact analysis of the 3D genome. Further colocalization with GWAS identifies shared genetic regulation of chromatin accessibility, gene expression and risk for neuropsychiatric and neurodegenerative diseases. Using massively parallel reporter assays in induced excitatory neurons, we screened 19,893 brain QTLs, identifying the functional impact of 476 regulatory variants. Combined, this comprehensive resource captures variation in the human brain regulome and provides novel insights into brain disease etiology.

Title: The topological chromatin remodeling in trisomy 21

Authors:

S. Antonarakis[1], I. Kolpakov[2], B. Ren[3], F. Berzukov[4], C. Borel[1]; [1]Univ. of Geneva, Geneva, Switzerland, [2]Campus Biotech Fndn. of Geneva, Geneva, Switzerland, [3]Univ California San Diego, La Jolla, CA, [4]Sophia Genetics, Lausanne, Switzerland

Abstract:

What happens to chromatin in the cell nucleus when an extra chromosome is added? To answer this question we have used human Trisomy 21 (T21) as a model. We have hypothesized that the architecture/topology of chromatin changes in T21, and these changes contribute to the genome-wide transcriptome dysregulation in T21.We chose an ideal experimental cellular system: fetal fibroblasts from a pair of monozygotic twins discordant for T21 (same genomes; only differences a supernumerary chromosome 21) to eliminate the effect/noise of the variability of human genomes. Using Hi-C along with histone modifications (H3K4me3) and Lamina Associated Domains (from Nature PMID 24740065) in these samples, we performed modeling of chromatin domains in T21 versus the euploid isogenic nucleus. The topological associated domains (TADs) did not show any differences in T21. Remarkably, however, the trans (intrachromosomal) HiC data of all chromosomes showed notable and consistent differences. There were 83 genomic regions dispersed throughout all chromosomes accounting for 235 Mb that change topology in T21 from the B to the A chromatin compartment or vice versa. We termed these regions ΔABt21. Consequently, the gene expression within ΔABt21 changed accordingly; regions that moved from B to A compartment showed increased gene expression, while regions that moved from A to B showed decreased gene expression in T21. In additional preliminarily experiments, a fraction of these ΔABt21 were also observed with fibroblasts from unrelated T21 when compared to euploid individuals.These data provide evidence for substantial chromatin remodeling of the entire genome in T21 that impacts the resulting transcriptome.More broadly, the changes in the chromatin architecture in aneuploidy and the resulting functional genomic dysregulation could cause some of the common features of total or partial chromosomal aneuploidy.

Session 019: Genetic basis of cardiovascular development and disease

Location: Conv Ctr/Room 145A/Level 1

Session Time: Thursday, November 2, 2023, 10:45 am - 12:15 pm

Title: Identification of Novel Candidate Risk Genes Causing Thoracic Aortic Disease

Authors:

B. Ziganshin[1,2], C. Leduc[3], J. Elefteriades[2], Y. Shen[4], W. Chung[3]; [1]Dept. of Genetics and Dev., Columbia Univ. Irving Med. Ctr., New York, NY, [2]Aortic Inst. at Yale-New Haven, Yale Univ. Sch. of Med., New Haven, CT, [3]Dept. of Pediatrics, Columbia Univ. Irving Med. Ctr., New York, NY, [4]Dept. of Systems Biology, Columbia Univ. Irving Med. Ctr., New York, NY

Abstract:

Background:
Diseases of the aorta are the 20th leading cause of mortality in the US contributing to 10,000 deaths annually. Aneurysm of the thoracic aorta are typically asymptomatic and often undetected until life-threatening aortic dissection/rupture ensues. One in five cases of thoracic aortic aneurysm and dissection (TAAD) are familial. We hypothesized that although more than 60 genes (with varying degree of evidence) have already been implicated in causing TAAD, more risk genes are yet to be discovered.
Methods:
A total of 1650 DNA samples (source: 1166 (71%) resected aortic tissue, 345 (21%) saliva samples, 139 (8%) blood) from 1429 patients with TAAD underwent exome sequencing at the Regeneron Genetics Center. After quality control (QC) procedures, data on 1274 unrelated TAAD patients of European ancestry were analyzed for enrichment in a case-control model for rare (AF<10^{-4}) deleterious missense (Dmis) and likely gene disrupting (LGD) variants. For Dmis variants we utilized variable threshold method to define deleteriousness. For the control group, we used 18,213 unaffected parents of European ancestry from the SPARK-autism study (sequenced on the same platform). After matching the cases and controls for the frequency of rare synonymous variation (difference <8%) by adjusting QC filters, we conducted a per-gene burden analysis (for protein coding genes only) using binomial test. To improve the power of detection of novel risk genes, we integrated case-control association of rare damaging variants with cell-type specific gene expression data from single cell RNA-sequencing of the ascending and descending aorta of E15 mouse embryos with the hypothesis that true risk genes are highly expressed early in development.
Results:
Two potential novel candidate genes emerged from this analysis. 16 Dmis variants in *VPS8*, a gene involved in endosomal vesicle fusion, were identified ($p=2.3 \times 10^{-5}$) in the TAAD cohort. 11 of these Dmis variants are localized in the β-propeller domain of VPS8 ($p=1.1 \times 10^{-6}$ for the domain only), with 9 variants appearing to cluster together. Seven Dmis variants in the *CLEC12A* gene showed genome wide significance for patients with ascending TAAD ($p=4.0 \times 10^{-7}$). *CLEC12A* plays a critical role in innate and adaptive immune response, and it is highly expressed in antigen presenting cells in aorta. Its role could help explain the role of inflammation in the pathogenesis of TAAD.
Conclusion:
In a case-control association analysis, we have identified two promising candidate risk genes for TAAD, which merit further investigation and confirmation in other cohorts of patients with aortopathy.

Title: Gene-Environment Interactions and Congenital Heart Defects.

Authors:

I. Zohn; Children's Natl. Hosp., washington, DC

Abstract:

Precise spatial and temporal regulation of retinoic acid (RA) signaling is essential for embryonic development. Increased or decreased RA signaling results in severe birth defects affecting multiple organ systems, such as congenital heart defects (CHDs). Pregnant women ingest varying vitamin A levels, depending on their diet and supplement usage, and the embryo must buffer variations to ensure normal development. Cardiac progenitors are specified within a gradient of RA established by negative and positive feedback loops involving RA synthesis and degradation enzymes. Progenitors from the anterior and posterior second heart fields (aSHF and pSHF, respectively) are derived from a common progenitor. Descendants of the aSHF contribute to the elongation of the outflow tract. In contrast, the pSHF contributes to the endothelium of the posterior pharyngeal arch arteries, which remodel to form segments of the aortic arch. We found distinct genetic mutations interact differentially with maternal diet, preferentially altering aSHF or pSHF development. For instance, 22q11.2 deletion syndrome (22q11DS) is associated with variable penetrance and expressivity of aortic arch and outflow tract defects. Our data studying a mouse model of 22q11DS indicates that variability in CHDs may be due to impaired buffering of RA exposures. We find that 22q11DS embryos from dams fed control levels of vitamin A exhibit aortic arch but not outflow tract defects. But, when dams are fed slightly more vitamin A, 22q11DS embryos, but not their wild-type siblings, exhibit both aortic arch and outflow tract defects. Conversely, mutation of *Hectd1,* a ubiquitin ligase required to fully activate RA signaling, interacts with mild vitamin A deficiency to cause CHDs. Heterozygous *Hectd1* mutant embryos from dams fed a control diet do not exhibit CHDs; however, when dams are fed a vitamin A-deficient diet, outflow tract but not aortic arch defects result. These findings illuminate how genetic mutations can interact with changes in the maternal diet, suggesting differential sensitivity of aSHF and pSHF progenitors to changes in RA gradients.

Title: A Multiomic timecourse atlas of murine vascular disease smooth muscle cell transition phenotypes and their perturbations with coronary disease gene *Tcf21* knockout

Authors:

D. Li[1], S. Kundu[1], D. Sharma[1], W. Gu[1], P. Cheng[1], J. P. Monteiro[1], M. Räsänen[1], T. Nguyen[1], C. Park[1], A. Kundaje[1], R. Wirka[2], T. Quertermous[1]; [1]Stanford Univ., Stanford, CA, [2]Univ. of North Carolina, Durham, NC

Abstract:

Recent focus on the context dependence of coronary artery disease (CAD) genome wide association study (GWAS) risk loci has highlighted vascular smooth muscle cells (SMC) as a key cell type which drives heritable CAD risk. Contribution to risk by this cell type is particularly interesting because of the cell state transitions to fibroblast-like fibromyocyte (FMC) and osteochondrogenic chondromyocyte (CMC) phenotypes. *TCF21* is one SMC CAD causal gene that promotes proliferation, migration and phenotypic transition to a FMC phenotype and is protective toward CAD risk.

To determine how SMC CAD genes like *TCF21* modulate phenotypic transitions and affect CAD pathophysiology, we performed deep multiomic profiling of *Tcf21* knockout and control atherosclerosis model mice with SMC specific Tomato lineage tracing. Aortic roots from control mice ($Myh11Cre^{ERT2}$, $ROSA^{tdT/+}$, $ApoE^{-/-}$) on high fat diet (HFD) were harvested for 7 scRNAseq and 6 scATACseq time points across 16 weeks. Similarly, aortic roots from *Tcf21* knockout ($Myh11Cre^{ERT2}$, $Tcf21^{\Delta SMC/\Delta SMC}$, $ROSA^{tdT/+}$, $ApoE^{-/-}$) mice were collected at 3 timepoints for paired scRNA and scATAC analyses for a total of over 140,000 cells after quality control filtering. Waddington-Optimal Transport (WOT) was applied to map the phenotypic modulation transcriptional trajectory across SMC lineage clusters. Then, utilizing multiomic profiles, we used the PANDO framework to develop regulatory networks conditioned on transcription factor expression and motif accessibility.

These analyses generated a multiomic timecourse atlas demonstrating the lineage traced phenotypic progression of SMC to FMC and CMC. The FMC transition phenotype was found to be significantly changed with *Tcf21* knockout, and this transcriptomic shift found to be enriched for CAD GWAS associated genes. scATAC profiling of *Tcf21* knockout mice demonstrated altered enrichment in epithelial mesenchymal transition pathway and inflammatory signals, along with increased global enrichment of AP-1 transcription factor motif activity. Integrating the single cell data with GWAS data from the Million Veterans Program CAD meta-analysis identified early pseudotime SMC and FMC but not CMC clusters to be the highly enriched for CAD risk. Lastly, we intersected WOT predicted transcription factors enriched in cells fated to become fibromyocytes with regulatory cascades predicted through PANDO to nominate key drivers of the phenotypic modulatory process. From these transcription factor modules, we found significant module score changes with *Tcf21* knockout, demonstrating the importance of these regulons in the phenotypic modulatory process and disease risk.

Title: Mapping the pleiotropic effects of genetic susceptibility loci identifies multi-system involvements in heart failure pathology

Authors:

A. Henry, R. T. Lumbers, HERMES Consortium; Univ. Coll. London, London, United Kingdom

Abstract:

Introduction: Heart failure (HF) is a multifactorial clinical syndrome with complex etiology that is not fully understood. Characterization of pleiotropic effects of HF genomic susceptibility loci on the medical phenome provides an opportunity for a system-agnostic identification of key etiologic mechanisms underlying specific HF subtypes.
Methods: We evaluated the pleiotropic effects of 66 HF genetic susceptibility loci identified through a genome-wide association study in 1.9 million individuals, including 145,795 cases of HF stratified by etiology and function. By integrating functionally-informed gene mapping and enrichment analyses, we identified putative effector genes, pathways, tissues, and cardiac cell types involved in the pathology across HF subtypes. Using graph modeling, we constructed a pleiotropy network for the identified HF loci derived based on a phenome-wide association analysis across 294 disease phenotypes in 400,000 UK Biobank participants. We then undertook a community detection analysis to identify etiological modules based on the network topology. We evaluated the sharing of causal variants and global pleiotropy between 24 implicated traits and HF through colocalization and genetic correlation analysis, and tested their causal inter-relationships using Mendelian randomization (MR).
Results: We identified 142 candidate genes within HF susceptibility loci, which were enriched for pathways including regulation of stress response, cell cycle, and sarcomeric components. Of these, 53 genes were differentially expressed in failing heart cells, predominantly in fibroblasts (28 genes) and cardiomyocytes (21 genes). Heritability of HF were enriched for genes that were more specifically expressed in both cardiac and extra-cardiac tissues, including kidney and pancreas. Pleiotropy network analysis revealed 207 genotype-phenotype associations that were mapped to 18 distinct etiological modules, including those consisting of ischaemic, metabolic syndrome, and cardiac morphofunctional phenotypes. Colocalization analysis demonstrated sharing of causal variants between 22 implicated traits and HF. At trait level, genetic correlation with at least 1 HF subtype was observed for 21 traits; however some, including type 2 diabetes, showed no causal effects on any HF subtype in MR.
Conclusions: Using human genetics, we uncover the involvements of multi-system disorders underlying the phenotypic spectrum of HF, providing insights into disease archetypes that may be amenable to targeted treatments.

Title: Large-scale genome-wide association study of dilated cardiomyopathy identifies genes and pathways contributing to pathology.

Authors:

S. Zheng[1], A. Henry[2], J. Ware[1], T. Lumbers[2], HERMES Consortium; [1]Imperial Coll. London, London, United Kingdom, [2]Univ. Coll. London, London, United Kingdom

Abstract:

IntroductionDilated cardiomyopathy (DCM) is a primary heart muscle disease characterized by ventricular dilatation and impaired contractility. While considered a Mendelian disorder, recent evidence highlights the important role of common genetic variation. Better characterization of the genetic architecture will provide opportunities for improved diagnosis, risk stratification and biological understanding.
MethodsWe performed a GWAS meta-analysis of 14,256 DCM cases and 1,185,671 controls of European ancestry from 16 studies ($GWAS_{DCM}$). To improve discovery power, we performed multi-trait analysis of GWAS ($GWAS_{MTAG}$) with three genetically correlated cardiac imaging traits in 36,203 individuals. To identify effector genes at loci we used a two-step approach. First, the nearest gene and top 3 genes prioritized by PoPS or V2G were selected as candidates. Second, the totality of evidence including 5 additional methods (coding variant, co-localization with expression, TWAS, ABC-model, and known Mendelian cardiomyopathy genes) was summarized in a single score. We performed rare variant burden analysis (MAF<0.1%) for predicted truncating variants (PTV) in prioritized genes against DCM and quantitative CMR traits in UK Biobank (UKB) and 100,000 Genomes Project (GeL), implemented by REGENIE. A polygenic score (PGS) generated using a Bayesian framework (PRS-CS) was assessed for association with the same outcomes in UKB participants.
ResultsWe identified 80 independent loci, of which 65 were novel (62 loci in $GWAS_{DCM}$, and 54 loci in $GWAS_{MTAG}$). We prioritized a single high-confidence gene at 61 loci, including genes with an established role in Mendelian cardiomyopathies (*TTN*, *BAG3*, *FLNC*, *MYBPC3*, *FHOD3*, *ALPK3*). Prioritized genes were enriched in key biological processes: cellular adhesion, junction organization, intracellular signaling, and aggrephagy. Rare variant burden analysis highlighted novel associations with DCM and quantitative traits for PTVs in *NEDD4L* and *MAP3K7* in UKB that were replicated in GeL. PGS was associated with DCM and quantitative CMR traits (OR per PGS SD 1.8, $P<1x10^{-16}$). Individuals in the top centile had 4-fold increased risk of having DCM compared with median risk. PGS stratified penetrance of rare variants in 1,546 carriers of pathogenic variants (top quintile 7.3%, bottom quintile 1.7%, P 0.005).
ConclusionWe have performed the largest DCM GWAS, resulting in 80 significant loci (65 novel), and prioritized genes using a systematic approach. We identify novel DCM-causing genes (*NEDD4L*, *MAP3K7*) and generate a PGS that associates with DCM and modulates penetrance in carriers of Mendelian disease-causing rare variants.

Title: Non-Canonical Splice Variants are an Underrecognized Cause of Acute Aortic Dissections.

Authors:

D. Murdock, D. Guo, A. Cecchi, I. Marin, D. Milewicz; UT McGovern Med. Sch. at Houston, Houston, TX

Abstract:

Background: Thoracic aortic dissections are a significant yet preventable cause of death. Individuals with heritable thoracic aortic disease (HTAD) have pathogenic variants in genes that significantly increase their risk of aortic dissections. We previously identified a splice-altering, deep intronic *FBN1* variant in a 5-generation family with highly variable age of onset of aortic dissection. We recently identified two additional individuals with early onset sporadic aortic dissections with non-canonical *FBN1* variants predicted by SpliceAI to disrupt splicing. We hypothesized that HTAD gene variants typically excluded from genetic analyses outside the canonical ± 1 or 2 splice sites are an underrecognized contributor to acute aortic dissections with reduced penetrance.
Methods: SpliceAI was used to evaluate the impact of non-canonical splice site rare variants in exome sequencing data from cohorts of unsolved thoracic aortic dissection patients. The first cohort included 577 individuals with sporadic aortic dissection (under 60 years of age, no family history, no syndromic features). The second cohort consisted of 395 probands from unsolved families where two or more members had thoracic aortic disease. We focused on variants in HTAD genes in which haploinsufficiency is disease-causing (*COL3A1*, *FBN1*, *LOX*, *MYLK*, *SMAD3*, *TGFB2*).
Results: Five non-canonical splice site variants in six individuals with a history of aortic dissections were found, exclusively in the sporadic dissection cohort and not the familial cohort. Intriguingly, despite these individuals not having syndromic features, all variants occurred in syndromic genes: *FBN1* (n=3, SpliceAI=0.83,0.95,1), *SMAD3* (n=1, SpliceAI=1), and *COL3A1* (n=1, SpliceAI=0.48). Three additional exome/genome sequencing cohorts are currently being interrogated for such variants.
Conclusion: Non-canonical splice site variants may be an underrecognized contributor to aortic dissections, particularly for sporadic and not familial cases. They may also be associated with a reduced penetrance phenotype, given their presence solely in the sporadic cohort lacking syndromic features. We hypothesize these variants may also be more prevalent in isolated thoracic aortic aneurysm patients. Diagnostic laboratories should consider transitioning to genome sequencing-based platforms to capture these clinically significant non-coding regions.

Session 020: Genetic tango: Dance between gene-gene and gene-environment interactions

Location: Conv Ctr/Room 146B/Level 1

Session Time: Thursday, November 2, 2023, 10:45 am - 12:15 pm

Title: Distinct explanations underlie gene-environment interactions in the UK Biobank

Authors:

A. Durvasula[1], A. Price[2]; [1]Harvard Med. Sch., Cambridge, MA, [2]Harvard Sch Pub Hlth, Boston, MA

Abstract:

The role of gene-environment (GxE) interaction in disease and complex trait architectures is widely hypothesized, but currently unknown. Here, we apply three statistical tests to detect and distinguish three different types of GxE interaction for a given disease/trait and E variable. First, we detect locus-specific GxE interaction by testing for genetic correlation < 1 across E bins. Second, we detect genome-wide effects of the E variable on genetic variance by leveraging polygenic risk scores (PRS) to test for significant PRSxE in a regression of phenotype values on PRS, E, and PRSxE, together with differences in SNP-heritability across E bins. Third, we detect genome-wide proportional amplification of genetic and environmental effects as a function of the E variable by testing for significant PRSxE with no differences in SNP-heritability across E bins. Extensive simulations show that these tests achieve high sensitivity and specificity in detecting and distinguishing these three distinct GxE scenarios. We applied our framework to 33 UK Biobank diseases/traits (average N=325K) and 10 E variables spanning lifestyle, diet, and other environmental exposures. First, we identified 20 trait-E pairs with genetic correlation significantly < 1 (FDR < 5%) (average genetic correlation = 0.94); for example, white blood cell count had a genetic correlation of 0.95 (s.e. 0.01) between smokers and non-smokers, implying locus-specific differences in trait effect size with smoking status. Second, we identified 26 trait-E pairs with significant PRSxE and significant differences in SNP-heritability across E bins; for example, type 2 diabetes had a significant PRSxE effect for alcohol consumption (P = 1e-13) with 4.2x larger SNP-heritability in the largest vs. smallest quintiles of alcohol consumption (P<1e-100), implying genome-wide effects of alcohol consumption on genetic variance. Third, we identified 18 trait-E pairs with significant PRSxE with no differences in SNP-heritability across E bins; for example, triglyceride levels had a significant PRSxE effect for composite diet score (P = 4.2e-5) with no differences in SNP-heritability, implying genome-wide proportional amplification as a function of diet. Analyses using biological sex as the E variable produced additional significant findings in each of the three scenarios. We estimate that the GxE (resp. GxSex) effects detected by our framework explain 1.8% ± 0.5% (resp. 2.8% ±0.7%) of trait variance across traits with significant GxE, compared to SNP-heritability of 29% ± 3%. Our estimates account for correlation between G and GxE when genetic variance varies with E, which can lead to bias in other methods.

Title: Transcriptome-level interpretation of gene-by-environment interactions for human complex traits

Authors:

Y. Wu, J. Miao, Q. Lu; Univ. of Wisconsin–Madison, Madison, WI

Abstract:

Genotype-by-environment interaction (GxE) studies have identified numerous genetic variants with heterogeneous phenotypic effects in different environments. However, interpretating the biological mechanisms underlying such interactions remains challenging. While it seems intuitive to leverage expression quantitative trait loci (eQTL) data to reveal the regulatory machinery behind GxE interactions, previous attempts to combine eQTL data with GxE signals have relied on methods designed for post-GWAS analysis which generally focus on genotype-phenotype main effects instead of GxE interactions. Here, we demonstrate that direct application of eQTL-GWAS integration approaches to interaction studies has a shaky statistical foundation and will lead to crucial misinterpretation of interaction findings. In particular, we demonstrate that gene expression can be involved in complex trait GxE through two distinct mechanisms that are often misinterpreted in the literature: environment-biased eQTL and expression-by-environment interaction. Based on this observation, we present a proper statistical framework to guide the integration of eQTL data in GxE research with two complementary analytical strategies. The first strategy leverages genome-wide GxE results and eQTL data to quantify the overall contribution of gene expression in mediating total GxE heritability. Second, we introduce a transcriptome-wide interaction study (TWIS) approach to identify specific genes explaining local genotype-environment interaction effects. We applied these methods to the gene-by-sex interaction (GxSex) analysis of waist-hip ratio (WHR) in UK Biobank using multi-tissue eQTL data from GTEx. We found that 11% (SE = 5.2%) and 20% (SE = 7.7%) of the GxSex component of WHR is explained by sex-biased eQTL and gene expression-by-sex interactions in the adipose tissue, respectively. We further identified 89 genes with sex-differential effects on WHR, including 12 novel genes distant from previously reported loci. The identified genes include *GRB14* (p=5e-41), a weight-loss-responsive gene in skeletal muscle, and *CCDC92* (p=5e-37), whose ablation inhibited adipose tissue dysfunction. In conclusion, our study completely redesigns the statistical approach for integrating transcriptome-level data into GxE analysis. The proposed methodology is broadly useful for future GxE research. The gene discoveries deepen our understanding of sex differences in the genetics of WHR.

Title: A scalable statistical framework for genome-wide interaction testing harnessing cross-trait correlations with an application to lipid analysis in UK Biobank

Authors:

S. Bian[1], A. Bass[2], D. Cutler[3], A. Wingo[4], T. Wingo[4], Y. Liu[4], M. Epstein[3]; [1]Dept. of Biostatistics and Bioinformatics, Emory Univ., Atlanta, GA, [2]Dept.s of Human Genetics, Emory Univ., Atlanta, GA, [3]Dept. of Human Genetics, Sch. of Med., Emory Univ., Atlanta, GA, [4]Dept.s of Neurology and Human Genetics, Emory Univ., Atlanta, GA

Abstract:

Well-powered GWAS have shown that many complex human traits originate from thousands of trait loci. Nevertheless, for many traits, family-based heritability estimates are often considerably larger than the corresponding heritability estimates from GWAS SNP data. While larger GWAS samples may fill this heritability gap, there is interest in exploring whether this missing heritability could be due to non-additive effect of genetic variation including GxG and GxE interaction effects. Exploring such non-additive effects is inherently challenging due to many factors including the modest magnitude of interaction effect sizes, the crippling multiple-testing burden, the complications with determining the interaction exposures, and impractical computational time and cost. To avoid these issues, we propose a scalable method called SCAMPI, a Scalable Cauchy Aggregate test using Multiple Phenotypes to test Interactions, by leveraging the variance-covariance structure of multiple phenotypes for improved performance. Specifically, we can show that SNPs with interactive effects yield differential correlation patterns among traits per genotype category. SCAMPI can be primarily applied for prioritizing SNPs for subsequent GxG and GxE studies. SCAMPI utilizes linear regression to evaluate the impact of genotype category on each pairwise cross-products among the multiple standardized phenotypes. This process yields multiple p-values. We then employ a Cauchy combination test to merge these p-values to obtain the SCAMPI p-value. SCAMPI offers three plausible advantages: First, there is no requirement to specify or measure interaction exposures. Second, it allows for simultaneous testing of multiple dependent or independent phenotypes. Third, SCAMPI is computationally scalable for conducting genome-wide interaction analyses on biobank scale data. Simulations demonstrated that SCAMPI preserved type-1 error and improved power compared to univariate variance-based tests of interactions across various conditions. We applied SCAMPI to UK Biobank (N= 337,422) to investigate the interaction effects for 288,910 SNPs post QC on four phenotypic outcomes, HDL-C, LDL-C, Triglycerides and BMI. Adjusting for confounding and considering SNPs with MAF $\geq$ 0.05, SCAMPI identified four genes, NCAN, TM6SF2, GATAD2A and NECTIN2 that were not identified by existing univariate variance-based tests. Indeed, these genes have been previously implicated with these lipid outcomes. We further discuss the modification of SCAMPI to discover interaction effects among rare variants using a principal-components strategy. SCAMPI is implemented as an R package for public use.

Title: Factorizing polygenic epistasis improves prediction and uncovers biological pathways in complex traits

Authors:

D. Tang[1,2], J. Freudenberg[3,2], A. Dahl[2]; [1]Harvard Univ., Boston, MA, [2]Univ. of Chicago, Chicago, IL, [3]Univ. of California, Los Angeles, Los Angeles, CA

Abstract:

Epistasis is central in many domains of biology, but it has not yet proven useful for complex traits. This is partly because complex trait epistasis involves polygenic interactions that are poorly captured in current models. To address this gap, we have developed a new model called Epistasis Factor Analysis (EFA). EFA assumes that polygenic epistasis can be factorized into interactions between a few Epistasis Factors (EFs) that represent latent polygenic components of the observed complex trait. Statistically, EFA aims to improve polygenic prediction and to increase power to detect epistasis. Biologically, EFA aims to unravel genetic effects into more-homogeneous units. The key idea in EFA is a parsimonious model of epistasis that enables scalable computation and allows sharing of signal between additive and epistasis effects. More concretely, EFA unbiasedly partitions GWAS effects into interacting components that can inform the biology underlying GWAS hits.
We mathematically characterize EFA with optimization algorithms to fit the model, proofs of its identifiability, and formal relationships with standard epistasis models. We use simulations to show that EFA substantially outperforms current epistasis models when its assumptions approximately hold. We apply EFA to yeast growth traits and find that it significantly outperforms the additive model for 6/46 traits, and more broadly that EFA performance increases with the epistasis heritability (Spearman correlation of 0.48 between differential prediction accuracy and differential heritability). Moreover, we show EFA always outperforms the standard epistasis model. We replicate these prediction improvements in a second yeast dataset. Finally, we apply EFA to four previously characterized complex traits in the UK Biobank and find statistically significant epistasis in all four (all empirical $p <= 0.006$), including two that are robust to scale transformation (both empirical $p = 0.002$). We also find that the inferred EFs partly recover pre-defined biological pathways for two of the traits ($p = 0.007$, $p = 0.001$), demonstrating that statistical models of epistasis can have biological utility.
Our results directly challenge the common wisdom that complex trait epistasis is nonexistent, undetectable, and/or useless in complex traits. The successes of EFA demonstrate that more realistic models can substantially improve our understanding of polygenic epistasis. Overall, we conclude that epistasis has potential for precision medicine and characterizing the biology underlying GWAS results, and that better models of complex trait epistasis are needed.

Title: A cross-population atlas of genome-wide gene-environment interactions between the East Asian and European populations

Authors:

S. Namba[1], K. Sonehara[1,2,3], K. Matsuda[4], Y. Okada[1,2,3]; [1]Osaka Univ. Graduate Sch. of Med., Suita, Japan, [2]Graduate Sch. of Med., the Univ. of Tokyo, Tokyo, Japan, [3]RIKEN Ctr. for Integrative Med. Sci., Yokohama, Japan, [4]Graduate Sch. of Frontier Sci., the Univ. of Tokyo, Tokyo, Japan

Abstract:

Genetic associations with human complex traits are influenced by environmental factors, and such gene-environment (GxE) interactions can help us understand the genetic etiology of human health and diseases. Despite the growing number of cross-population meta-analyses of genome-wide association studies (GWAS), large-scale genome-wide scans of GxE interactions have been limited to specific trait categories and have been conducted primarily in European populations, which has hindered the evaluation of the cross-population consistency of the GxE structure. Here, we parallelly conducted genome-wide GxE interaction scans for 39 biomarkers and 7 diseases in Biobank Japan (BBJ, N_{max} = 157,588) and UK Biobank (UKB, N_{max} = 272,757). By jointly testing GxE effects for age, sex, drinking, smoking, physical activities, and dietary traits, we detected 44 and 69 GxE interaction trait-locus pairs ($P < 5.0\times10^{-8}$) for BBJ and UKB, respectively. Six loci with clear etiology of trait-locus associations were shared between the biobanks and mediated by the same environments, supporting the existence of the GxE effects shared across populations. Replication in an independent Japanese cohort (N_{max} = 64,619) confirmed 19 / 44 (43%) of the BBJ GxE interactions with $P < 0.05 / 44$ (= 1.14×10^{-3}), including the highly pleiotropic, multi-environment-mediated GxE interactions of the East Asian-specific variants in the *ALDH2* locus. Meta-analyses across the three cohorts enhanced the evidence of GxE interactions for 65 / 107 (61%) trait-locus pairs, including 11 pairs not detectable by GWAS even after the meta-analyses. When comparing the GxE structure at the genome-wide level, the overall GxE heritability was concordant between BBJ and UKB (Pearson's correlation = 0.45). The relative amount of GxE heritability over marginal effect (i.e., G-only) heritability varied across traits, ranging from 8.9% (S.E. 1.3%) for height to 45.5% (S.E. 2.2%) for body mass index. The GxE heritability for age was relatively high in late-onset diseases and traits, reflecting that the decades of accumulation of disease-promoting effects trigger these diseases. The GxE heritability was not correlated with that of variance quantitative trait loci (vQTL), although vQTL has been used to detect GxE loci without measuring environmental exposures. The number of overlapped loci between GxE and vQTL was also limited, indicating that direct evaluation was important to detect GxE interactions. Our work presents both shared and distinct GxE interactions between the East Asian and European populations, providing a cross-population portrait of the GxE structure.

Title: Ancestry-specific regulatory and disease architectures are likely due to cell-type-specific gene-by-environment interactions

Authors:

J. Wang, S. Gazal; USC, Los Angeles, CA

Abstract:

Multi-ancestry genome-wide association studies (GWAS) highlighted a non-negligible fraction of causal variants with ancestry-specific effect sizes. Understanding why and where these effects occur is fundamental to understand the genetic basis of human diseases and human complex traits. Because these traits are dominated by regulatory variants, we investigated if ancestry-specific gene regulation (e.g. ancestry-specific eQTL effect sizes) due to gene-by-environment (GxE) interactions could lead to ancestry-specific disease genetic effect sizes. Here, we leveraged scRNA-seq data in PBMCs from 44 individuals of East-Asian (EAS) and European (EUR) ancestry (Perez et al. 2022 Science) to detect and characterize genes differentially expressed across ancestries (ancDE genes) within 7 main immune cell types. We observed that ancDE genes tend to be differentially expressed in a single cell type (83%). Half of these ancDE genes have corresponding cell-type-specific eQTLs, which tend to have high Fst (0.21) across EAS and EUR. AncDE genes with eQTLs were enriched in genes involved in immune responses to environments, suggesting that ancDE genes could be driven by recent adaptation (rather than only genetic drift of their eQTLs).Next, we leveraged ancestry-matched GWAS of 31 diseases and complex traits to test if genetic variants surrounding ancDE genes were enriched in ancestry-specific effect sizes using S-LDXR (Shi et al. 2021 Nat Com). We determined that squared multi-ancestry genetic correlation is 0.69 ± 0.04 for SNPs surrounding ancDE genes, representing the lowest correlation reported by S-LDXR; numbers were similar when stratifying genes with and without eQTLs, suggesting that even if genes were differentially expressed due to allele frequency differences of their eQTLs, they are likely enriched in ancestry-specific effect sizes. We observed that these depletions were driven by ancDE genes from B cells (0.35 ± 0.06) and conventional dendritic cells (0.36 ± 0.10).Finally, we illustrated how GxE interactions have led to differential MCL1 expression in B cells, different MCL1 eQTL effect sizes in blood, and different allele effect sizes in Lymphocyte Count GWAS between EAS and EUR.Altogether, we observed that ancDE genes tend to be cell-type-specific, enriched in genes interacting with the environment, and extremely enriched in variants with ancestry-specific disease effect sizes, suggesting the impact of cell-type-specific GxE interaction in both regulatory and disease architectures. Our results imply that large single-cell and GWAS datasets in diverse populations are required to improve our understanding of human diseases.

Session 021: Human genome evolving I

Location: Conv Ctr/Room 147A/Level 1

Session Time: Thursday, November 2, 2023, 10:45 am - 12:15 pm

Title: Quantifying gene loss across 462 mammalian species to capture human gene essentiality

Authors:

C. Liao[1], R. Ye[1], J. Fu[1], F. Ivankovic[1], R. Walters[2], C. Churchhouse[1], E. Karlsson[3], K. Lindblad-Toh[4], M. Hiller[5], M. Talkowski[2], B. Neale[2]; [1]Broad Inst. of MIT and Harvard, Cambridge, MA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]Univ. of Massachusetts Med. Sch., Worcester, MA, [4]Broad Inst., Cambridge, MA, [5]Senckenberg Society for Nature Res. & Goethe Univ., Frankfurt, Germany

Abstract:

Gene orthologs provide valuable insights into the functions and evolutionary history of genes across different species, making them a critical tool for understanding the biological basis of complex traits and gene essentiality. Here, we leveraged millions of years of evolution and developed novel metrics for measuring gene conservation across 462 mammalian species, GISMO (Gene identity score of mammalian orthologs) and GISMO-mis (GISMO-missense), to provide a more comprehensive understanding of gene conservation in mammalian species. GISMO is a measure of gene conservation across mammalia by modelling the number of gene loss events, and GISMO-mis measures the fixed missense differences across mammalian species for a given gene. Our analysis showed that both metrics are strongly correlated with other measures of gene constraint, such as LOEUF from gnomAD ($R^2_{GISMO} = 0.41$, $R^2_{GISMO\text{-}MIS} = 0.66$, $P < 1 \times 10^{-300}$). Importantly, GISMO is not strongly correlated to coding sequence length and can identify constrained and conserved genes that are too small to be captured well by independent constraint metrics such as LOEUF, pTriplo, and pHaplo. We also found that GISMO captures the rare variant association signal across a range of neuropsychiatric phenotypes, such as schizophrenia, autism, and other neurodevelopmental disorders ($R^2 = 0.12\text{-}0.44$, $P < 1 \times 10^{-300}$). Moreover, common variant heritability of neuropsychiatric traits are highly enriched in the most conserved deciles of GISMO, further underscoring the biological relevance of these metrics in identifying functionally important genes. We additionally find that GISMO has the lowest duplication and deletion rate in the most conserved deciles for copy number variants in the UK Biobank, suggesting that it may be an important metric for dosage sensitivity. In conclusion, GISMO and GISMO-mis are novel metrics that leverage millions of years of evolution to provide a valuable tool for studying gene conservation across mammalian species and identifying biologically relevant genes underlying complex traits.

Title: Deep learning to understand the genetic architecture and evolution of the human pelvis

Authors:

L. Xu[1], E. Kun[2], M. Brasil[3], T. Singh[4], V. Narasimhan[2]; [1]Univ. of Texas at Austin, Austin, TX, [2]The Univ. of Texas at Austin, Austin, TX, [3]Univ. of California, Berkeley, Berkeley, CA, [4]Columbia Univ., New York, NY

Abstract:

The human pelvic shape has undergone significant change during evolution from primates. This transformation, reducing pelvic circumference to support bipedal locomotion, however, gives rise to the obstetrical dilemma - a mismatch between the large brain size of infants and the narrowed birth canal. To understand the genetic underpinnings of the human pelvic shape, we applied a deep-learning model to analyze 30,334 whole-body dual-energy X-ray absorptiometry (DXA) images from the UK Biobank (UKB). We extracted 12 different image-derived phenotypes that capture anatomical measurements of hip shape, including the width of the birth canal. Model performance was similar to the inter-rater differences of human annotation, with a difference of <3 pixels (<0.81cm) for each phenotype. We verified the accuracy of our phenotyping through checks of duplicate images taken across two time points, which had correlation rates of greater than 99% across all phenotypes. All pelvic proportions were found to be highly heritable (~25-40%) and a genome wide association study (GWAS) of these traits identified a total of 201 independent loci. Unlike other proportion traits like arm and leg proportions, certain pelvic shape phenotypes important for birth like subpubic angle and acetabular diameter exhibit significantly lower genetic correlation between sexes (rg~0.7) while other pelvic phenotypes such as pelvic width and iliac flare angle/ratio showed correlation consistent with 1. We then conducted phenotypic associations and polygenic risk score analyses of pelvic proportions with a range of musculoskeletal disorders and discovered significant associations particularly with hip and back osteoarthritis, leading causes of adult disability in the United States. As pelvic morphology changed dramatically through hominin development, we performed evolutionary analysis. Loci identified by our GWAS associated with the size of the birth canal and overall pelvic height showed enrichment in regulatory elements of genes that are differentially expressed early on in development between humans and primates. On examining the genetic covariance matrix we found that birth canal width had the lowest conditional evolvability of the phenotypes we examined. Birth canal width was also highly genetically correlated with the width of the skull providing additional evidence of constraint. Taken together, our work validates the use of deep learning models to extract phenotypes from medical images, identifies genetic variants that affect the pelvic form - an important part of our biomechanical shift to bipedalism - and links an evolutionarily novel aspect of human anatomy to its pathogenesis.

Title: Order of magnitude increase in signals of natural selection in ancient DNA data realizes the promise of time transects to provide qualitatively new insights into human adaptation

Authors:

A. Akbari[1,2,3], S. Gazal[4], A. Barton[2,3], S. Mallick[1,2,3], D. Reich[1,2,3,5]; [1]Harvard Med. Sch., Boston, MA, [2]Harvard Univ., Cambridge, MA, [3]Broad Inst., Cambridge, MA, [4]USC, Los Angeles, CA, [5]Howard Hughes Med. Inst., Harvard Med. Sch., Boston, MA

Abstract:

The ancient DNA revolution has transformed our understanding of the past, debunked previous theories and revealed the prevalence of migration and population mixing in ancient times. However, limited sample sizes have hindered the study of natural selection. Here, we analyze 5925 high-quality ancient individuals, mostly unpublished, to study natural selection in Western Eurasia over the past 10,000 years. We employ new statistical methods, including imputation tailored to ancient DNA data and a mixed model approach to account for population structure. Our methodology effectively corrects for population structure and identifies 428 independent loci showing significant signals of selection. We found 17 times more significant loci in our ancient DNA time series compared to previous reports. These loci are highly enriched in variants with known phenotypic effects, suggesting that the majority of them are likely real.

We highlight results on the level of both single alleles and complex traits. First, we detected novel signals of selection affecting particular loci, including an ~20% decrease over the last 8000 years in the frequency of an allele at the TCHH gene that predisposes to balding and straight hair, and a rise from ~0% to ~8% over the last 6000 years in the frequency of the blood type B variant at the expense of the blood type A variant, potentially reflecting balancing selection for two alleles with profoundly different impacts on a range of phenotypes.

Secondly, we combine our selection results with genome-wide association studies to study signals of polygenic adaptation, examining directional and stabilizing selection. There is no evidence of directional selection for stature, but strong evidence of stabilizing selection. We detect clear signs of directional selection on genetic variants that today modulates risk for autoimmune disease, with polygenic scores (PGS) for susceptibility to Celiac Disease and Crohn's Disease increasing over time, and for Rheumatoid Arthritis decreasing. We furthermore observe directional selection for cognitively linked traits, with PGS that today predict obesity decreasing over time and PGS that today predict higher educational attainment increasing.

Overall, we observe extreme enrichment of signals of selection on alleles that today modulate immune traits (3-7 fold), and metabolic traits (3-5 fold), but no signal of enrichment in alleles affecting brain-related traits. This suggests that even though there are individual signals of selection affecting cognitive traits, behavioral evolution may not have been a particularly important focus of adaptation in this period of profound cultural and environmental change.

Title: Examining molecular responses to dietary selection pressures in mammals and humans

Authors:

W. Meyer[1], M. Tene[1], K. Foley[2], M. D. Pollard[3], D. E. Schäffer[4], A. M. Graham[5], E. E. Puckett[3], I. M. Kaplow[6]; [1]Lehigh Univ., Bethlehem, PA, [2]Univ. of Iowa, Iowa City, IA, [3]Univ. of Memphis, Memphis, TN, [4]Massachusetts Inst. of Technology, Cambridge, MA, [5]Univ. of Utah, Salt Lake City, UT, [6]Carnegie Mellon Univ., Pittsburgh, PA

Abstract:

Our mammalian relatives have adapted to a wide array of different environments, inducing changes in selective pressures that have affected their genomes. Some mammals have experienced similar selective pressures to those experienced by the ancestors of modern human populations, including pressures related to obtaining and processing new and different food sources. Have the same molecular networks been subject to selective pressure related to diet in other mammals as those we observe to underlie local adaptation in humans? We here explore this question by comparing the results of phylogenetic comparative analyses across mammals to population and statistical genetic signatures of adaptation and trait association within humans. To identify regions of the genome associated with mammalian dietary specialization, we implement a genome-wide scan for convergent changes in the relative evolutionary rates of genes and regulatory elements across branches of the mammalian phylogenetic tree with similar diets, leveraging newly generated genome-wide alignments for hundreds of mammalian species. We additionally search for changes specifically influencing gene expression in the liver, analyzing changes in expression associated with dietary transitions using publicly available RNA-seq data for over 60 mammals, and identifying enhancers whose activity co-evolves with diet using machine learning techniques. We compare the genes and regulatory elements associated with dietary transitions in mammals identified in this way with the genetic variation previously found to be associated with liver-related disorders in genome-wide association studies in humans, as well as with variants underlying local selective sweeps and polygenic adaptation inferred to be related to dietary changes. By comparing the regions of the genome impacted by natural selection related to dietary traits across mammals with those influenced by similar selective pressures in humans, we generate a deeper understanding of the constraints on genomic evolution across different evolutionary timescales.

Title: Human Accelerated Regions Regulate A Common Set of Target Genes during Human and Chimpanzee Neurodevelopment

Authors:

A. Pal, M. A. Noble, J. P. Noonan; Yale Univ., New Haven, CT

Abstract:

Changes in gene regulation have been implicated in the evolution of uniquely human traits. Human Accelerated Regions (HARs), which are highly constrained regions with numerous sequence changes in humans compared to other primates, have been shown to encode human-specific regulatory functions, notably in the developing brain. However, the mechanisms by which HARs may alter gene expression remain unclear. A recent study proposed that HARs may regulate different genes in humans compared to chimpanzee, a mechanism termed enhancer hijacking. Other studies suggest that HARs alter the expression of common genes targeted by the human and chimpanzee orthologs. Here we comprehensively identified the gene targets of 1,590 HARs and their chimpanzee orthologs using Capture-HiC in iPSC-derived human and chimpanzee neural stem cells (NSCs) and neurons. We found that HARs exhibit highly similar interaction profiles in both species, with ~ 97% of gene targets being conserved between HARs and their chimpanzee orthologs. Notably, HARs targeted genes involved in neurogenesis and axon guidance, including the *SLIT/ROBO* pathway. Integrating our HAR interaction maps with maps of gene expression, histone modifications, looping factors, RNA Pol II, and chromatin accessibility, we discovered that many HAR interactions are maintained in a repressed state in both human and chimpanzee NSCs, particularly involving genes associated with neuronal differentiation and maturation, such as *MEIS2* and *CDKN1C*. These findings indicate that interactions between many HARs and target genes expressed in neurons are pre-established in NSCs prior to neuronal fate commitment. Moreover, we have identified a set of 75 genes contacted by ~100 HARs, where the activating or repressive chromatin environment around the gene target correlates with the changes in expression profiles of these genes in either species during early neurodevelopment. Overall, our results support that HARs primarily act by modulating the expression of ancestral gene targets rather than targeting different genes in humans compared to chimpanzee. HARs may thus contribute to human-specific features of neurodevelopment via modulation of a common regulatory architecture.

Title: Pooling-based phylogenetic methods elucidate accelerated evolution of cis-regulatory elements.

Authors:

X. Zhang[1], Z. Liu[1], B. Fang[2], Y. Huang[1]; [1]Pennsylvania State Univ., University Park, PA, [2]Harvard Univ., Cambridge, MA

Abstract:

Recent comparative genomic studies have identified many human accelerated regions (HARs) that are conserved across mammals but show elevated substitution rates in the human lineage. However, it remains unknown to what extent non-conserved cis-regulatory genetic elements are under accelerated evolution in humans and other primates. To bridge this gap, we first introduce two pooling-based phylogenetic methods (GroupAcc) with dramatically enhanced sensitivity to examine accelerated evolution in cis-regulatory elements. Using these new methods, we show that more than 6,000 transcription factor binding sites (TFBSs) annotated in the human genome have experienced accelerated evolution in hominini, apes, and Old World monkeys. Although these TFBSs individually show relatively weak signals of accelerated evolution, they collectively are more abundant than HARs. Also, we show that accelerated evolution in Pol III binding sites may be driven by lineage-specific positive selection, whereas accelerated evolution in other TFBSs might be driven by nonadaptive evolutionary forces. The accelerated TFBSs are enriched around developmental genes, suggesting that accelerated evolution in TFBSs may drive the divergence of developmental processes between primates. While GroupAcc effectively identifies genomic features associated with even subtle signals of accelerated evolution, they face limitations in disentangling the contributions of correlated genomic features. For example, when accelerated elements are bound by multiple transcription factors, it becomes challenging to determine which specific transcription factor is primarily responsible for this accelerated evolution. To address this problem, we build an evolution-based regression model to infer the independent effects of correlated genomic features on accelerated evolution. We show the power of this new framework in both simulated and empirical data.

Session 022: New treatments for some rare classics

Location: Conv Ctr/Ballroom C/Level 3

Session Time: Thursday, November 2, 2023, 10:45 am - 12:15 pm

Title: A novel, all-in-one adeno-associated vector for epigenome-editing perturbations: *proof-of-concept* and preclinical validation in the CNS

Authors:

B. Kantor[1], B. Odonovan[2], J. Rittiner[1], N. Lindner[1], O. Chiba-Falek[2]; [1]Duke Univ.- Neurobiology, Durham, NC, [2]Duke Univ.- Neurology, Durham, NC

Abstract:

Epigenetic editing is an emerging field in developing safe and effective engineering tools to control gene expression. The system consists of a guide RNA (gRNA), a deactivated-Cas9 nuclease (dCas9) fused to epigenetic effector molecules that can safely and effectively silence or activate a specific gene target. Adeno-associated vector (AAV) is a platform-of-choice for the delivery of therapeutic cargoes; nevertheless, its small packaging capacity is commonly not suitable for the delivery of large cargoes, such as CRISPR/dCas9-effector systems. To circumvent this limitation, most AAV- based CRISPR/Cas tools delivered separately from two viral cassettes. However, this approach requires higher viral payloads and usually is less efficient. Here, we developed a compact CRISPR/dCas9-repressor system packaged within an optimized and single AAV vector. The system uses a smaller dCas9 paired with bimodular, synthetic repressor carrying small transcription repression domain (TRD) of MeCP2 repressor and KRAB. The developed construct can be efficiently packaged into AAV particles at the levels required for its clinical use. Using a reporter assays, we demonstrated that the platform is capable to robustly and sustainably repress the expression of a gene-of-interest in vitro and in vivo. Moreover, we provided an example of silencing APOE, known as the major genetic risk factor for late onset Alzheimer's disease (LOAD). Our findings suggest that the developed platform will broaden the CRISPR/Cas9 toolsets used for transcription control over gene expression in the therapeutic settings.

Title: Single AAV-mediated miniaturized CRISPR activation for the treatment of congenital muscular dystrophy.

Authors:

J. Cheng-Zhang, M. A. Johnson, A. E. Azar, R. D. Nicholls, **D. U. Kemaladewi**; Div. of Genetic and Genomic Med., Dept. of Pediatrics, UPMC Children's Hosp., Univ. of Pittsburgh, Pittsburgh, PA

Abstract:

Programmable recruitment of transcription factors to activate genes in a targeted and specific manner, commonly known as CRISPR activation (CRISPRa), offers unprecedented opportunities for therapeutic interventions. For example, the upregulation of the compensatory gene *LAMA1* using CRISPRa to treat LAMA2-related deficient congenital muscular dystrophies (LAMA2-RD) has emerged as a mutation-independent therapeutic approach for this condition. However, the viral delivery of CRISPRa components has been complicated by their size and the limited packaging capacity of adeno-associated viruses (AAVs), necessitating the use of dual AAV approaches that contribute to a common bottleneck in gene therapy, including high dose, potential toxicity, and production cost. Here, we present a miniaturized CRISPRa system that can be packaged into a single AAV and apply this approach to upregulate a compensatory gene *Lama1* in a mouse model representing a severe form of LAMA2-RD. Single-AAV9 carrying *S. aureus* dCas9, driven by novel mini promoter 4XNRF1, combined with tripartite activators VP64-Δp65-ΔRTA and a single guide RNA targeting the mouse *Lama1* promoter resulted in systemic LAMA1 expression, including critical target tissues for LAMA2-RD such as the skeletal muscles and peripheral nerves. Importantly, the single AAV-mediated *Lama1* upregulation significantly improved neuromuscular functions and extended the lifespan of the severe LAMA2-RD mouse model. Compared to the current state of CRISPRa therapeutics in LAMA2-RD, here, we successfully rescued the disease phenotype in a much more severe mouse model with only half the viral load. These results are crucial to advancing the therapeutic development of the disease modifier gene upregulation for LAMA2-RD. Moreover, the components and engineered cassettes presented here are adaptable for AAV-based gene therapy for different genetic diseases, such as other neuromuscular disorders and haploinsufficiency-related diseases.

Title: Durable *HTT* silencing using non-evolved dCas9 epigenome editors in patient-derived neuronal cells

Authors:

J. Waldo[1,2,3,4,5], J. A. Halmai[1,2,3,4,5], J. L. Carter[1,2,3,4,5], D. Brown[1,2,3,4,5], J. A. Nolta[1,4,5], K. Fink, D[1,2,3,4,5]; [1]UC Davis Med. Ctr., Inst. for Regenerative Cures, Sacramento, CA, [2]Ctr. for Interventional Genetics, Sacramento, CA, [3]MIND Inst., Dept. of Neurology, Sacramento, CA, [4]Stem Cell Program, Sacramento, CA, [5]Gene Therapy Ctr., Sacramento, CA

Abstract:

Huntington's Disease (HD) is an autosomal dominant disorder caused by a trinucleotide repeat in exon 1 of the Huntingtin (*HTT*) gene. This expansion leads to protein misfolding that causes widespread cellular dysfunction and ultimately leads to cell death. The advent of CRISPR-dCas9 gene regulation technologies allow for the targeting of the causative gene and subsequent downregulation via fused effector domains that induce heterochromatin at the epigenetic level through DNA methylation (DNMT3A/L) and H3K9me3 (KRAB), blocking transcription. Therefore, we propose using dCas9 epigenetic editing for the downregulation of *HTT* as a therapeutic approach for HD. Allele specificity is important in HD, as *HTT* has essential functions within the cell and total knockdown could be deleterious. *HTT* has large haplotype blocks that allow for allele specific targeting based upon the presence of heterozygous single nucleotide polymorphisms (SNPs). These SNPs are in genomic regions devoid of NGG PAM sites, a requirement for SpdCas9 binding. To address this bottleneck, we conducted a screen of multiple dCas9 variants fused to KRAB and DNMT3A/L with increasingly expanded PAM targeting to initially assess ability to downregulate total *HTT*. Surprisingly, only SpdCas9 was able to significantly knockdown *HTT*, while expanded PAM site variants dxCas9 and dCas9-VQR were unable to reduce *HTT* expression. ChIP-qPCR was performed to identify differential binding enrichment of dCas9 variants and H3K9me3 enrichment flanking sgRNA target sites in the *HTT* promoter. We further investigated DNA methylation changes through reduced representation bisulfite sequencing. In addition, we demonstrate mitotically stable *HTT* silencing of up to 6 weeks *in vitro* in rapidly dividing cells. We then moved forward to assess total *HTT* knockdown in HD patient-derived neuronal stem cells and healthy controls. *HTT* silencing efficiency was measured by qPCR in our healthy controls and HD patient cells. We identified significant downregulation of *HTT* in our treatment group compared to an unguided control. RNAseq was used to identify differential gene expression between treatment and controls, and gene ontology analysis was performed to identify rescue of biological processes involved in *HTT* molecular pathogenesis. In addition, we demonstrate high target specificity of our lead sgRNA. Additional bottleneck is the delivery of large dCas9 epigenome editors to the CNS. Current studies are addressing novel viral-like particle delivery for our identified epigenome editors *in vivo*. This approach holds great promise for those suffering from HD.

Title: Atropine treatment rescues myopia in a mouse model of Marfan syndrome

Authors:

J. Doyle, J. Johansson, S. Banerjee, Z. Chai, N. Anderson, A. Niknahad, J. Han, H. Dietz; Johns Hopkins Univ., Baltimore, MD

Abstract:

Early-onset extreme near-sightedness (high myopia; HM) is prevalent in many genetic disorders including Marfan syndrome (MFS). HM that results from increased eye size (axial length, AL) predisposes to the development of pathological structural changes in the eye (e.g. retinal detachment, myopic retinoschisis, retinal atrophy) that can cause irreversible visual impairment or even blindness. Intriguingly, the molecular mechanisms driving pathological eye growth in monogenic forms of HM and therapies to combat it remain unexplored. Atropine eyedrops dose-dependently slow myopia progression in the general population, but their effect in any monogenic form of HM has never been reported. And despite being the most widely used therapy for myopia globally, atropine's mechanism of action remains disputed. We show using in-vivo optical biometry that a well-characterized mouse model of MFS (*Fbn1*$^{C1039G/+}$) displays increased AL (60µm, $p<0.0001$) and HM (-9D, $p<0.0001$) compared to WT littermates at 2 months of age. MFS eyes display increased Smad2/3 and Erk1/2 activation ($p<0.001$ for both), although the two pathways are activated in different layers of the retina. Notably, the angiotensin receptor blocker losartan, which markedly inhibits aortic root aneurysm growth in this mouse model, shows no protective effect against eye growth. In contrast, atropine eyedrop treatment shows significant promise. Whilst the highest concentration of atropine studied in humans (1%) non-selectively inhibits eye growth in both MFS and WT mice ($p<0.0001$ for both), lower concentrations (e.g. 0.1%) achieve a selective inhibition of eye growth in myopic MFS eyes ($p<0.01$) without impacting physiological growth in WT eyes. Adjusting the frequency of atropine treatment also improves its selectivity. Whilst daily treatment with 1% atropine markedly inhibits eye growth in both MFS and WT mice, treatment either 2 times ($p<0.05$) or 3 times ($p<0.001$) per week selectively inhibits eye growth in myopic MFS eyes without impacting WT eye growth. Atropine markedly inhibits Smad2/3 activation in the retinas of MFS mice ($p<0.0001$) to levels indistinguishable from WT. Cumulatively, these data demonstrate that: 1. This mouse model of MFS faithfully recapitulates the increased AL and axial myopia seen in humans; 2. Atropine eyedrops show promise for the treatment of monogenic forms of HM such as MFS, which infers that there may be etiological overlap between monogenic and common forms of myopia; 3. Inhibition of retinal Smad2/3 activation represents a novel mechanism of action for atropine in the treatment of myopia.

Title: Development of bone targeted anti-TGFβ antibody for the treatment of osteogenesis imperfecta

Authors:

I-W. Song[1], D. Nguyen[1], A. Tran[1], S. Musaad[1], Y. Mirabile[1], B. Greene[2], J. Cao[2], X. Ying[2], A. Park[2], E. Masterjohn[2], D. Honey[2], S. Liu[2], S. Nagamani[1,3], B. Lee[1,3]; [1]Baylor Coll. of Med., Houston, TX, [2]Sanofi, Waltham, MA, [3]Texas Children's Hosp., Houston, TX

Abstract:

Osteogenesis imperfecta (OI) is a genetic skeletal disorder for which, currently, there are no FDA-approved therapies. We, and others, have previously shown that excessive TGFβ signaling is a key pathogenic driver in OI and that inhibition of TGFβ could be a potential mechanism-specific approach for treatment. In a phase 1 study with fresolimumab, an anti-TGFβ antibody, we found that individuals with moderate OI, i.e., OI type IV had increase in lumbar spine areal bone mineral density whereas individuals with the more severe forms, i.e., OI types III and VIII did not. This suggested a dose-specific effect based on disease severity that correlates with treatment studies in preclinical models of OI of varying severity. Here, we report on the pharmacodynamics and pharmacokinetics of fresolimumab and preliminary results in generating a bone-targeted anti-TGFβ antibody to improve efficacy and safety. We found that peak plasma levels of fresolimumab were higher in individuals with OI who received 4 mg/kg dose compared to 1 mg/kg. The plasma levels dropped to 84% of the peak level by day 15 and returned to baseline by day 90. Overall, an anabolic window, i.e., lower CTX and higher P1NP levels as compared to baseline values, were observed in 4 out of 8 participants. In the 4 mg/kg cohort, suppression in bone turnover marker osteocalcin was observed at day 30 and 180 together with an anabolic window. This alluded to the possible need for a higher anti-TGFβ concentration in OI bone. Therefore, we generated a bone-targeted anti-TGFβ antibody, 1D11-D10. After a single dose administration in WT mice, 1D11-D10 showed increased distribution to bone (lumbar and femur) and reduced concentrations in plasma, kidney and heart as compared to untargeted 1D11. Moreover, 1D11-D10 has a longer half-life of 76 days in bone compared to 16.3 days for 1D11 ($t_{1/2}$ ratio=4.7). Importantly, at 0.3, 1, and 5 mg/kg 3 times/week dosing of *Col1a2*$^{G610C/+}$ OI mice, 1D11-D10 showed comparable effect on bone mass at 1mg/kg and a trend towards higher efficacy at 5mg/kg comparing to untargeted 1D11. These data supported the effective bone targeting and a potential increased efficacy and safety profile for 1D11-D10. As bone targeted anti-TGFβ therapy may offer wider therapeutic index, this may be important for clinical translation in OI and other skeletal diseases with TGFβ dysregulation.

Title: Development of a splicing modulator therapy for frontotemporal dementia

Authors:

E. Morini[1], H. Lindmeier[1], C. Morill[2], M. Arnold[2], S. Barraza[2], Y. Yu[2], J. Laughlin[2], S. Khalil[2], M. Dieterich[2], A. Minnella[2], K. Datta[2], N. Zhang[2], M. Weetall[2], C. Trotta[2], M. Woll[2], E. Welch[2], J. Trimmer[3], M. Silva[1]; [1]Massachusetts Gen. Hosp., Boston, MA, [2]PTC Therapeutics, South Plainfield, NJ, [3]PTC Therapeutics, Soth Plainfield, NJ

Abstract:

Frontotemporal dementia (FTD) is a neurodegenerative disease characterized by the accumulation of abnormal forms of the microtubule-associated protein Tau (MAPT). FTD can manifest as either sporadic or inherited, with the latter resulting from mutations in the *MAPT* gene. Currently, there are no effective disease-modifying therapies available for FTD. The regions encompassing MAPT exon 10 and its intron10-exon10 boundary are hot spot for pathogenic mutations, with many intronic mutations affecting the accessibility of the spliceosome to this region and increasing exon 10 inclusion. Furthermore, missense and deletion gain-of-function mutations in exon 10 enhance the propensity of 4R-Tau to aggregate. A targeted mRNA approach that promotes the exclusion of *MAPT* exon 10 would be advantageous for both classes of disease-causing mutations.Our team has generated a class of molecules known as splicing modulator compounds (SMCs) that promote MAPT exon 10 exclusion. We tested the efficacy of these SMCs in neuronal cell models derived from patients carrying splicing mutations (IVS10+16 and S305N) and gain-of-function mutations (P301L). Through a machine learning approach, we initially identified BPN-15477 as an SMC that promoted exon 10 exclusion in MAPT transcripts. BPN-15477 led to a reduction in the accumulation of total Tau and phosphorylated Tau (pTau) in FTD-induced pluripotent stem cell (iPSC)-derived neurons expressing the 4R-specific Tau-P301L mutation. However, due to its low potency, high concentrations of BPN-15477 were required to achieve modest changes in splicing. Subsequently, we developed next-generation SMCs and evaluated their efficacy using an optimized assay platform. These new SMCs exhibited significantly improved potency and efficacy compared to BPN-15477, enabling us to reduce treatment concentrations to the nanomolar range. The new SMCs demonstrated a dose-dependent decrease in exon 10 inclusion and 4R MAPT expression, resulting in reduced levels of 4R Tau protein. In parallel, we observed reductions in 3R-Tau, total Tau, and pTau, with the latter exhibiting a more pronounced dose-dependent effect. By decreasing the expression of the disease-causing mutant 4R Tau isoform, we observed a global shift in Tau accumulation and oligomerization, which also impacted the steady-state levels of 3R Tau. Our findings highlight that treatment with SMCs led to a significant reduction in 4R Tau expression, thereby influencing the accumulation of total Tau and pTau proteins in FTD neuronal models. These results underscore the remarkable therapeutic potential of this novel class of small molecules as an oral treatment for FTD.

Session 023: Systematic analysis of variant functions

Location: Conv Ctr/Ballroom A/Level 3

Session Time: Thursday, November 2, 2023, 10:45 am - 12:15 pm

Title: iPSC-SGE: assessing variant effects in differentiated cell types at scale

Authors:

S. Fayer[1], C. E. Friedman[1], S. Pendyala[1], M. Dawood[2], D. Yang[1], D. Fowler[3], L. Starita[3]; [1]Univ. of Washington, Seattle, WA, [2]Baylor Coll. of Med., Houston, TX, [3]Univ of Washington, Seattle, WA

Abstract:

Multiplexed Assays of Variant Effect have so far been performed in utilitarian cancer cell lines, excluding the assessment of genes that are expressed in specialized cell types in their correct cell context. For example, many genes associated with cardiomyopathies encode sarcomere structural components that are exclusively expressed in cardiomyocytes and cannot be assayed in utilitarian cancer cell lines. To overcome this limitation, we developed a new saturation genome editing method for introducing single nucleotide variants (SNVs) into a single allele of a diploid induced pluripotent stem cell line (iPSC-SGE). We applied iPSC-SGE to the cardiac myosin binding protein C3 (MYBPC3) gene, in which pathogenic variants are the most common cause of familial hypertrophic cardiomyopathy. We focused on the C10 domain, spanning exons 32-34, which anchors MYBPC3 to myosin heavy chain within the sarcomere and is a hotspot for pathogenic missense variants. We introduced 498 SNVs into exon 32 and the flanking intronic regions of MYBPC3 in iPSCs and differentiated these cells to cardiomyocytes for phenotyping. We used a fluorescence activated cell sorting assay to score variants based on their abundance relative to the wild type allele, a proxy for incorporation into the sarcomere, and recovered 492 of 498 possible variants. 79 variants had reduced function scores, including all 14 nonsense and all 12 canonical splice site variants. All 11 pathogenic controls were functionally abnormal and all 14 benign controls were functionally normal. We are currently applying iPSC-SGE to the remainder of the MYBPC3 C10 domain and plan to use our results to inform clinical variant interpretation of SNVs in this region. iPSC-SGE massively expands the phenotypic space accessible to multiplexed assays of variant effect. For example, the method could be applied to many of the 22 actionable cardiomyopathy risk genes, as well as many other clinically important genes where assessment of variant phenotype in a differentiated cell context is required.

Title: Single cell sequencing as a universal variant interpretation assay

Authors:

F. Aguet, D. Cao, K. Jean-Baptiste, M. Sun, X. Zhou, S. Wong, L. Chen, H. Xu, K-H. Farh; Illumina, Inc., Foster City, CA

Abstract:

The vast majority of all possible ~70 million protein-altering variants in the human genome are of uncertain significance, with only a small fraction annotated in clinical variant databases. Closing this gap is essential to identify clinically relevant variants and understand their mechanism of action. Toward this goal, we transduced all possible coding variants in the TP53, CDKN2A ($p16^{INK4a}$), and SOD1 genes underlying pan-cancer, melanoma, and amyotrophic lateral sclerosis, respectively, and developed an approach based on single-cell RNA sequencing to read out the functional effects of each variant in the global expression signature. We sequenced ~215,000 cells expressing TP53 variants, ~140,000 cells expressing CDKN2A variants, and ~450,000 cells expressing SOD1 variants, with 98.5% (2774), 89.3% (972), and 100% (1113) of all possible amino acid variants observed in at least 20 cells for the three genes, respectively. We used variant enrichment and expression profiles to characterize the gene expression programs that are perturbed by pathogenic variants and quantify proliferation differences, enabling the identification of protein-domain specific axes of pathogenicity and their corresponding functional mechanisms. Using both supervised and unsupervised classification approaches to quantify the pathogenicity of each variant (e.g., with respect to synonymous variants), we show strong concordance with prior experimental measurements of TP53 variant effects as well as computational predictions, and accuracies of >90%-100% for pathogenic ClinVar variants. In summary, we show that it is possible to quantify the effects of almost all variants of a gene with single cell sequencing, demonstrating its use as a universal variant interpretation assay.

Title: Uncovering the functional landscape of *PMS2* variants by deep mutational scanning.

Authors:

S. Vishnopolska, J. Kitzman; Univ. of Michigan, Ann Arbor, MI

Abstract:

The effectiveness of genetic testing is limited by the difficulty of interpretation, particularly for missense variants. A prominent example comes from one of the most widely screened gene families, the DNA mismatch repair (MMR) pathway, which safeguards the genome against replication errors. The human MMR machinery comprises two protein complexes encoded by four genes (*MSH2, MSH6, MLH1*, and *PMS2*). Inherited heterozygous loss of any of these genes causes Lynch Syndrome (LS), which affects ~1:300 individuals worldwide, conferring early-onset colorectal and endometrial cancer risk. *PMS2* deficiency exhibits incomplete penetrance, challenging its diagnose and screening based upon family history. Accordingly, >90% of all *PMS2* missense variants reported in clinical variant databases such as ClinVar remain as variants of uncertain significance (or VUS). We applied deep mutational scanning to systematically and prospectively resolve *PMS2* variant function. Libraries containing every possible *PMS2* missense and nonsense mutation (N=17,240) are introduced, one per cell, into human *PMS2* KO cells. To report on variant function, we use genotoxic selection with 6-thioguanine (6-TG), which selectively kills cells with intact MMR. Cells expressing a dysfunctional missense variant survive 6-TG treatment, become enriched, and are identified by deep sequencing. Overall, 21% of screened missense variants exhibited loss of function, including deeply conserved residues such as Asn45 and Asp48 essential for ATPase activity. We observed high concordance (prAUC=0.92) between our function scores and available clinical variant interpretations in ClinVar (n=12), indicating that these function scores have the potential to assist reclassification of the 1,959 standing missense VUS as well as novel variants discovered in the future. We further leveraged the scale of this approach to explore epistatic interactions within *PMS2* coding sequence. We identified nine common haplotypes with one or more missense variants, segregating with a minor allele frequency between 0.5%-34% in the 1000 Genomes Project. Each of these common variants could modify the effects of other variants *in cis*, potentially explaining some of the incomplete penetrance associated with *PMS2* mutation. We find the common variant p.Arg20Gln (MAF=7.5% in 1KGP; 7.4% in gnomAD), interacts with 29 other nearby missense variants, such that they are changed from neutral to deleterious (or vice versa) when combined *in cis*. Mechanistic studies of these epistatic interactions are underway. Our results suggests haplotype-resolved studies of variant function may play a key role in clinical variant interpretation.

Title: The impact of amino acid insertions and deletions on proteinsolubility and function.

Authors:

M. Topolska[1], A. Beltran Marqués[1], B. Lehner[1,2,3,4]; [1]Ctr. for Genomic Regulation, Barcelona, Spain, [2]Univ. Pompeu Fabra (UPF), Barcelona, Spain, [3]Institució Catalana de Recerca i estudis Avançats (ICREA), Barcelona, Spain, [4]Wellcome Sanger Inst., Wellcome Genome Campus, Hinxton, United Kingdom

Abstract:

Amino acid insertions and deletions (indels) are common mutational events, accounting for a large fraction of polymorphisms in human genomes (Mullaney et al. in 2010). However, despite this there is limited information about indel effects on protein stability and function, hence their effects remain less well understood compared to substitution mutations. Whereas substitutions only alter the amino acid side chains of proteins, indels are 'backbone mutations' that modify protein length, giving them the potential to cause major phenotypical changes in a single mutational step. Generating a "leap" across sequence space, small indels can result in dramatic changes in protein structure, which makes it difficult to predict their effects. Since indels are known to be involved in genetic diseases like cystic fibrosis (Wang et al., 2000) and numerous types of cancer (Falini et al., 2005; Ye et al., 2015), developing models to predict their effects would be useful for clinical diagnosis as well as in the evolving field of protein engineering and personalised medicine. To address the lack of data on the impact of indels, we are systematically quantifying their effects on *in vivo* protein solubility of a diverse set of human protein domains using deep mutational scanning (DMS). DMS is an experimental approach which allows quantification of the effects of genetic variants on biological functions using selection assays coupled with next-generation sequencing. Here, as the selection read-out, we use a simple protein fragment complementation assay, which allows us to measure indel tolerance across multiple proteins in bulk (Faure et al., 2022). We use this approach to explore indel tolerance across different domains and secondary structure elements and how these patterns of tolerance differ from substitutions. Our mutational libraries include short indels of 1-3 amino acids in length, which reflects the most abundant natural variation and allows us to capture length-dependant effects on protein stability. We have also used our dataset to evaluate variant effect predictors. While the effects of substitutions are quite well predicted, existing algorithms perform less well on indels. However we show that indel tolerance can be predicted using simple models that combine the measured (or predicted) substitution tolerance with secondary structure features. Our results provide a first systematic overview of *in vivo* indel tolerance and predictive methods for exploring how non-missense mutations contribute to evolution and disease.

Title: Mapping the genetic contribution to neuroinflammatory cell traits using QTL mapping of iPSC-derived microglia

Authors:

M. Perez-Alcantara[1,2], Y. Chen[1,2], S. Washer[3,1,2], J. Steer[1], J. McWilliam[1], D. Ebner[4], S. Cowley[3], G. Trynka[1,2], A. Bassett[1,2]; [1]Wellcome Sanger Inst., Wellcome Genome Campus, Hinxton, United Kingdom, [2]Open Targets, Wellcome Genome Campus, Hinxton, United Kingdom, [3]James and Lillian Martin Ctr. for Stem Cell Res., Sir William Dunn Sch. of Pathology, Univ. of Oxford, Oxford, United Kingdom, [4]Target Discovery Inst., Univ. of Oxford, Oxford, United Kingdom

Abstract:

Mounting genetic evidence, from both common and rare variants, implicates microglia in neurodegenerative disease (ND). Microglia perform various immunological roles in the brain as sentinels, housekeepers and mediators of neuroinflammatory response via cellular signalling, migration and phagocytosis. We aim to identify regulators of these cellular functions to provide insights into molecular mechanisms of ND (Alzheimer's and Parkinson's). To achieve this, we have examined the role of neurodegeneration-associated variants and loci on gene expression of induced pluripotent stem cell (iPSC)-derived microglia via expression quantitative trait locus (eQTL) mapping.

We have developed an improved monoculture differentiation protocol that produces cells transcriptionally closer to primary microglia and with improved viability. We have performed pooled differentiations (19-72 donors in a pool) of iPSCs from 227 donors from the human pluripotent stem cell initiative (HipSci) to map the effects of common genetic variation. We also stimulated terminally differentiated cells with lipopolysaccharide (LPS) or interferon gamma (IFNγ) to capture transcriptomic changes that occur during immune response. Single-cell RNA-seq (scRNA-seq) was performed at the terminal point of differentiation, followed by eQTL mapping using tensorQTL and colocalization with common variants to link ND-associated risk variants to their target genes.

Our strategy leveraged the expression signatures of donors shared across pools and allowed for seamless integration of the data without retaining batch effects. We identified 1,685 differentially expressed genes (FDR < 0.05 and |log2 fold change| > 1)) across the three conditions. Mapping cis-eQTLs for each condition identified 2,143 genes in total with at least one significant (FDR < 0.05) associated variant (eGenes), of which 58% were treatment-specific. eGenes were significantly enriched in GWAS candidate genes for Alzheimer's disease, particularly in the LPS and IFNγ-treated samples (hypergeometric test q-value = 0.002).

Our findings highlight candidate genes that could be driving neuroinflammatory disease aetiology through microglial phenotypes. To link identified disease-eGenes to cell effector functions we are undertaking phenotypic QTL analysis of migration and phagocytosis, using the same pools of iPSC-derived microglia as well as CRISPR knockout screens of phagocytic function. This will enable us to map the full mechanisms of disease, from genetic variants to gene expression regulation through to misregulated cellular processes.

Title: SIGMA leverages protein structural information to predict the pathogenicity of missense variants

Authors:

H. Zhao[1], H. Du[1], S. Zhao[2], Z. Chen[1], Y. Li[1], K. Xu[1], B. Liu[1], X. Cheng[1], W. Wen[1], G. Li[1], G. Chen[1], Z. Zhao[1], G. Qiu[1], P. Liu[2], T. Zhang[1], Z. Wu[1], N. Wu[1]; [1]Chinese Academy of Med. Sci. & Peking Union Med. Coll., Beijing, China, [2]Baylor Coll. of Med., Houston, TX

Abstract:

Assessing the pathogenicity of missense variants is vital for interpreting genetic data. Given the close relation between the structure and function of proteins, it is promising to evaluate the effect of a missense variant in the context of the protein structure. However, the scarcity of known 3D protein structures has hindered the exploitation of the structural information. Recently, the launch of AlphaFold2 provided extensive and accurate 3D protein structure predictions. Based on these Alphafold2 structures, we extracted 57 features for 27,165 benign and 22,957 pathogenic variants from the ClinVar and gnomAD databases. Using the gradient boosting machine algorithm, we developed the Structure-Informed Germline Missense mutation Assessor (SIGMA) that predicts the pathogenicity of missense variants. SIGMA outperformed the state-of-the-art computational predictors on both the labeled variant dataset and experimental dataset. We found that the relative solvent accessibility of the mutated residue contributed greatly to the predictive ability of SIGMA. We next explored the potential application of combining SIGMA with other state-of-the-art computational predictors (SIGMA+) and demonstrated its high predictive ability for the variant pathogenicity (AUC = 0.966). To facilitate the application of SIGMA, we have pre-computed SIGMA scores for all possible missense variants (more than 48 million variants) across the disease-associated genes (n = 3,454), and developed an interactive online platform (https://www.sigma-pred.org/).

Session 024: All about mom: Genetics and genomics of maternal outcomes

Location: Conv Ctr/Room 146B/Level 1

Session Time: Thursday, November 2, 2023, 1:45 pm - 2:45 pm

Title: Exploring the Genetic Landscape of Pregnancy-related Phenotypes and Complications through Low-Pass Whole Genome Sequencing in Over 100,000 Chinese Pregnancies

Authors:

X. Jin, **H. Zhu**; BGI-Shenzhen, Shenzhen, China

Abstract:

Over the last decade, significant advancements in Genome-wide Association Studies (GWAS) have been made, albeit with considerable diversity in cohort composition, genotyping strategy, and sample size. A primary challenge is the underrepresentation of non-European populations and limited availability of pregnancy and newborn phenotypic studies. Non-Invasive Prenatal Testing (NIPT) for fetal trisomies, utilizing maternal plasma cell-free DNA (cfDNA) sequencing, has emerged as a globally adopted molecular test with over 10 million participants. Our previous research demonstrated the potential of low-pass Whole Genome Sequencing (WGS) data derived from NIPT for human genetic studies, identifying novel signals for traits such as height, weight, fertility, and maternal age. In this study, we broadened this approach to encompass a range of maternal and neonatal phenotypes and common pregnancy complications. We conducted a genetic analysis of 104 maternal and neonatal phenotypes from over 100,000 Chinese women, marking this as the largest such genetic study in the Asian population. Our GWAS revealed 407 genome-wide trait-locus associations, with 75.18% previously cataloged in the GWAS Database. An independent replication study validated 70 out of the 88 (79.5%) novel hits. We discovered a significant association of ESR1 (estrogen receptor) with fasting glucose, hemoglobin, hematocrit, and leukocyte counts. Considering the dramatic fluctuations in estrogen levels during pregnancy, we propose these ESR1-associations may be pregnancy-specific and merit further exploration. We investigated the genetic background of common pregnancy complications, including Gestational Diabetes Mellitus (GDM) and Intrahepatic Cholestasis of Pregnancy (ICP). For GDM, our GWAS identified two candidate genes, and an ancient DNA analysis unveiled a mutation initially present in an ancient Chinese individual during the Holocene period, potentially providing insight into the genetic diversity of GDM across populations. For ICP, we replicated a known signal in the ABCG8 gene and discovered a novel association in SLC39A9. To facilitate further research, we've established a website for visualizing, sharing, and downloading the GWAS summary statistics of pregnancy phenotypes and complications. In conclusion, our findings underscore the power of leveraging NIPT data to gain genetic insights into maternal and neonatal phenotypes, accelerating future mechanistic studies into complex pregnancy traits and diseases.

Title: Investigating the genetics of perinatal depression via polygenic scores

Authors:

A. Pimplaskar, K. Gelev, V. Sarwal, J. Chiang, L. Olde Loohuis; Univ. of California, Los Angeles, Los Angeles, CA

Abstract:

Perinatal depression (PND) is a depressive illness affecting mothers around pregnancy. PND heritability estimates (40-55%) have been shown to be higher than non-PND depression (28-44%), but, due to a lack of sufficiently powered genome-wide association studies (GWAS) for PND, its genetic structure remains understudied. We use data from the UCLA-ATLAS Biobank to investigate the genetics of PND.

We define PND cases as women who develop depression or receive antidepressants during or up to 1 year after pregnancy, without prior history of depression. Polygenic scores (PGS) were generated for existing GWAS of psychiatric, hormonal, pregnancy-related, and inflammation traits using SBayesR. Using logistic regression, we compare PGS between PND cases to mothers with non-PND pregnancies, PND cases to depression unrelated to pregnancy, and stratifying ante- (APD) and post-partum (PPD) depression in PND cases.

Within the 36778 patients in ATLAS, we identify 417 PND cases, 986 non-PND pregnancies, and 7141 non-PND depression cases. 848 non-PND pregnancies remained after omitting patients with a history of depression. The majority of PND cases (79.7%) were postpartum.PND patients compared to controls (without history of depression) showed a nominally increased PGS for bipolar disorder (OR = 1.14; p = 0.026). PND cases compared to non-PND depression patients have lower PGS estimates for major depression (MDD) (OR = 0.87; p = 0.009) and ADHD (OR = 0.88; p = 0.009), but higher estimates for age at menarche (OR = 1.14; p = 0.012).

Comparing onset of depression in the perinatal period, we observe a stronger association between PGS for MDD and APD vs. controls (OR = 1.25; p = 0.05) compared to PPD vs. controls (OR = 1.07; p=0.29). This association in APD is also stronger than the overall observed association between the MDD PGS and PND case-control status.

While a majority of current literature focuses on postpartum depression, this work considers depression both before and after birth. Differences between genetic risk for MDD between PND and non-PND depression, as well as based on time of onset, can point to differential disease etiology.

We are currently assessing the relationship between genetic risk estimates and clinical predictors of PND risk, while also expanding to a larger sample of ~60000 ATLAS patients. Due to limited sample size, we aim to replicate our results in external data. While in the future, we aim to perform well-powered GWAS for PND, polygenic scores facilitate a cursory investigation into its genetic basis.

Title: Genome-wide polygenic risk score predicts future type 2 diabetes in women with history of gestational diabetes

Authors:

J. Choi[1], H. Lee[2], A. Kuang[3], A. Huerta[4], W. Lowe[5], H. Jang[6], K. Park[1], S. Kwak[1]; [1]Seoul Natl. Univ. Hosp., Seoul, Korea, Republic of, [2]Seoul Natl. Univ., Seoul, Korea, Republic of, [3]Northwestern Univ., Chicago, IL, [4]Broad Inst., Cambridge, MA, [5]NU Feinberg Sch Med, Chicago, IL, [6]Seoul Natl. Univ. Bundang Hosp., Bundang-gu, Korea, Republic of

Abstract:

Around 30-50% of women with a history of gestational diabetes (GDM) develop type 2 diabetes (T2D) within 10 years after parturition. While prediction of incident T2D prediction with conventional clinical risk factors has been actively investigated, it is hard to apply to GDM patients who are relatively young because they do not fully express the endophenotype as much as other T2D patients. In addition, the role of genetic information in prediction of incident T2D in women with history of GDM is still unclear. We aimed to investigate the utility of the polygenic risk score (PRS) in predicting incident T2D using five different GDM cohorts: 1) UK Biobank (UKBB, N=941), 2) Seoul National University Hospital (SNUH, N=357), 3) Korean Genome and Epidemiology Study (KoGES, N=379), 4) Mexican GDM cohort (MXGDM, N=67) and 5) Hyperglycemia and Adverse Pregnancy Outcome cohort (HAPO, N=418). We calculated the PRS for T2D by Bayesian regression and continuous shrinkage priors, using summary statistics from representative genome-wide association studies in corresponding ancestries. We conducted case-control test and time-to-event analysis for incident T2D with PRS to evaluate the association between PRS and incident T2D. Moreover, we assessed the improvement of the C-statistic when PRS was added to clinical risk factors (family history of diabetes, age at GDM diagnosis, body mass index and systolic blood pressure) for the diagnostic implications of PRS. Additionally, the risk of T2D was estimated for those who were in the top 10 percentile of PRS compared to all other remainders. The incidence of T2D was 19.8% in UKBB, 28.3% in SNUH, 23.5% in KoGES, 20.9% in MXGDM and 8.8% in the HAPO cohort. The mean standardized PRS for those who developed T2D was higher when the results from five cohorts were meta-analyzed: 0.31±1.04 in women who developed T2D and -0.08±0.98 in women who did not develop T2D (P<0.001). Adding PRS to clinical risk factors improved the C statistic from 0.71 (95% confidence interval [CI] 0.68-0.73) to 0.74 (95% CI 0.71-0.76) when meta-analyzed. PRS was associated with prevalent T2D in GDM women when meta-analyzed: odds ratio 1.44 (95% CI 1.29-1.60). Furthermore, those who were in the top 10 percentile had 2.59-fold increased risk of T2D: (95% CI 1.88-3.57). The result was consistent in time-to-event analysis. The hazard ratio was 1.68 (95% CI 1.52-1.85, P<0.001) when meta-analyzed and top 10 percentile PRS group had a 3.43-fold increased risk of T2D: (95% CI 2.68-4.48, P<0.001). In summary, genetic information in women with previous GDM pregnancy could be used to identify those who are at high risk of future development of T2D and provide the opportunity for tailored prevention.

Title: A multi-ancestry genome-wide association study identifies a novel candidate locus near the *RARB* gene associated with hypertensive disorders of pregnancy in the Personalized Environment and Genes Study (PEGS) cohort.

Authors:

J. A. Mack[1,2], A. Burkholder[1], F. Akhtari[1], J. S. House[1], U. Sovio[2], G. C. Smith[2], C. Schmitt[1], D. Fargo[1], J. Hall[1], A. A. Motsinger-Reif[1]; [1]Natl. Inst. of Environmental Hlth.Sci., Durham, NC, [2]Univ. of Cambridge, Cambridge, United Kingdom

Abstract:

Hypertensive disorders of pregnancy (HDP) are a collection of multisystem disorders defined as experiencing hypertension during pregnancy, pre-eclampsia, and/or eclampsia. HDP is one of the leading causes of maternal morbidity and mortality. Significant racial and ethnic disparities exist in prevalence and severity of HDP, where non-Hispanic Black people are dying at more than three times the mortality rate for non-Hispanic White people in the United States. Genetic factors related to pregnancy-related traits are understudied, especially among ancestrally diverse cohorts. This study assessed genetic contributions to HDP in multi-ancestral cohorts. Genetic variation associated with self-reported HDP was estimated using whole-genome sequencing data in 202 HDP case participants and 1,569 control participants in the Personalized Environment and Genes Study (PEGS). Using Firth-corrected logistic regression with an additive-allele model, we performed a genome-wide association study for previously pregnant case and control groups, where 25% of the sample were of African, South Asian, East Asian, or Admixed American ancestry. We adjusted for the first 10 principal components in addition to self-identified race/ethnicity, gravidity, age, and BMI at study enrollment. We identified two genome-wide significant associations with HDP in the maternal genome. The most significant region is near the *RARB* gene (retinoic acid receptor beta: embryogenesis, trophoblast differentiation) on chromosome 3. The lead variant was *rs61176331* (3p24.2; 3:24930847_C/A [hg38]; OR (95% CI): 3.09 (2.11, 4.53); $p=7.97\times10^{-9}$). This candidate locus overlaps with an open chromatin region (ENSR00000680404). In ancestry-stratified analyses for participants of predominant African ancestry and predominant European ancestry, this signal was attenuated. Validation analysis conducted in the multi-ancestral UK Biobank cohort demonstrated association of a proximal variant to *rs61176331* near *RARB* with HDP (p=0.03). To our knowledge, this is the first association at this locus discovered on the maternal genome. To further explore the genomic landscape as it relates to HDP, we performed pathway enrichment analysis to identify biological pathways and conditions that are enriched for genes associated with HDP. The most significant pathways were related to immunological conditions and processes, such as autoimmune thyroid disease, and allograft rejection. These findings demonstrate the power of multi-ancestral studies for genetic discovery, while highlighting the relationship between immune response and HDP.

Session 025: ARG! Ancestry-led reanalysis of genomes

Location: Conv Ctr/Room 147A/Level 1

Session Time: Thursday, November 2, 2023, 1:45 pm - 2:45 pm

Title: Leveraging ancestral recombination graphs for fast and accurate estimation of selection coefficients and allele histories from ancient DNA

Authors:

A. Vaughn; UC Berkeley, Berkeley, CA

Abstract:

The inference of natural selection from sequence data is one of the classic questions in population genetics. Many current methods rely on summary statistics that do not take full advantage of all the information present in the data. Furthermore, existing machine learning approaches lack the flexibility to extend to contexts for which they were not trained. However, there has recently been increased interest in using ancestral recombination graphs (ARGs) for population genetic inference, as these methods do not rely on summary statistics. This means that methods utilizing ARGs can be considered full-likelihood methods which will have dramatically increased accuracy. Here, we present an accurate ARG-based method for inferring selection coefficients from genetic data, based on the existing CLUES method. A particular advantage of this method is its flexibility, as we are able to estimate changing selection coefficients through time in a rigorous hypothesis testing framework as well as incorporate changes in population size. We also make significant changes to the underlying hidden Markov model of CLUES to make it much dramatically faster, allowing it to be applied genome-wide and on larger datasets. Finally, we extend this framework to apply to ARGs sampled on ancient data, allowing us to leverage the increasing number of high-quality ancient genomes available. This allows us to use both time-series information and linkage information around the locus of interest to estimate selection. We apply this new method, which we call CLUES2, to a set of recently published high-quality ancient genomes from Stone Age Eurasia to interrogate questions surrounding the timing of selection in these populations and how it relates to demographic and ecological selective pressures present during some of the major societal upheavals of this period. We present CLUES2 as a well-documented Python package that can be readily used by population genetics researchers.

Title: Improved imputation and association of rare variants using ultra-fast inference of ancestral recombination graphs

Authors:

A. Gunnarsson[1,2], P. Palamara[2,1]; [1]Wellcome Ctr. for Human Genetics, Oxford, United Kingdom, [2]Dept. of Statistics, Univ. of Oxford, Oxford, United Kingdom

Abstract:

The evolutionary history of a set of genomes can be represented using an ancestral recombination graph (ARG). These genealogical graphs may be leveraged in a wide range of applications, including association studies of rare variation in under-sequenced populations and genotype imputation and storage, but such applications require methods that scale to biobank-sized, genome-wide data sets. We present Threads, a new approach to ARG inference that combines efficient haplotype matching with likelihood-based modeling of genealogical relationships. Threads runs one or more orders of magnitude faster than current scalable methods such as ARG-needle (Zhang et al, Nat Genet 2023), Relate (Speidel et al, Nat Genet 2019), and tsdate/tsinfer (Kelleher et al, Nat Genet 2019, Wohns et al, Science 2022), consuming less memory and disk space while achieving equal or higher accuracy across a range of metrics. Threads is robust to genotyping errors and can be applied to genotyping arrays and sequencing data, phased and unphased.

We used Threads to infer genome-wide genealogies using sequencing data for 2,251 samples from 26 populations in the 1000 Genomes Project data set (1KGP) and genotyping arrays for 358,592 individuals from 6 ancestry groups within the UK Biobank (UKBB), and performed three analyses to demonstrate the usefulness of these inferred genealogies. First, we show that inferred ARGs can be used to achieve greater genotype imputation accuracy than current algorithms for ultra-rare variants (minor allele count ≤5). In simulated data sets, we observe gains in r^2 equivalent to a 50% increase in reference panel size for ultra-rare variants compared with IMPUTE5 and Beagle 5.4. Using the 1KGP as a reference panel, we obtain a 2-3% increase in r^2 for ultra-rare variants. Second, we performed genealogy-wide association on height and six blood-related traits using ARGs from the UKBB, assessing the power to detect association with unobserved genetic variation. In these analyses, Threads runs 7.5 times faster than ARG-needle while finding a comparable number of independent associations. Lastly, ARGs inferred by Threads achieve greater compression rates than traditional genotype formats, while encoding not only genomic variants but also their genealogical history. ARGs inferred using Threads reduce the disk space required to store 1KGP sequencing data by 60%, UKBB genotyping array data by 56%, and 1KGP imputed genotype dosages by 20%. Together, these results demonstrate that accurate biobank-scale, genome-wide genealogical inference may be used to effectively complement or improve upon existing methods for the analysis of rare variation in complex traits.

Title: A Litmus Test for Confounding in Polygenic Scores

Authors:

S. P. Smith[1], H. Mostafavi[2], J. Berg[3], D. Peng[4], G. Coop[5], M. D. Edge[4], A. Harpak[1]; [1]The Univ. of Texas at Austin, Austin, TX, [2]Stanford Univ., Stanford, CA, [3]The Univ. of Chicago, Chicago, IL, [4]The Univ. of Southern California, Los Angeles, CA, [5]The Univ. of California, Davis, Davis, CA

Abstract:

Polygenic scores (PGS) are being rapidly adopted for trait prediction in the clinic and beyond. In addition to “direct” effects of one’s genotype on one’s trait value, PGS and their predictive accuracy can depend on environmentally or culturally-mediated factors such as Stratification, Assortative mating, and Dynastic (indirect) genetic effects (“SAD effects”). Despite the accumulating qualitative evidence, there is currently no estimation method to quantify the relative contribution of SAD effects to variation in a given PGS. We therefore developed a method that measures how much of the variance in a PGS (in a given sample) is driven by direct effects, SAD effects and their covariance. We leverage a comparison of a standard-GWAS based PGS and a PGS based on a sib-GWAS—which is immune to SAD effects. Furthermore, the method breaks down effect variance components into principal axes of genetic ancestry, allowing for a nuanced interpretation of SAD effects. We applied our method to a range of PGS from the UK Biobank and large meta-analyses for anthropometric, biomedical and behavioral traits and found large and interpretable SAD signals in some of them. For example, we estimate that most of the variance in height PGS based on the GIANT study is appropriated to SAD effects and nearly 40% of the variance aligns specifically with the main North-South genetic ancestry axis, highly suggestive of stratification effects. Our method shows how the power of large GWAS can be married with the articulation of family designs to aid in the interpretation of PGS.

Title: A probabilistic graphical model for estimating selection coefficient of nonsynonymous variants from human population sequence data

Authors:

Y. Shen, Y. Zhao; Columbia Univ., New York, NY

Abstract:

Predicting the effect of missense variants is critically important in disease risk gene discovery and clinical interpretation of genetic variants in population screening and diagnosis. Most of published computational methods rely on pathogenic labels from curated databases. These databases have uneven quality and uncertain bias across genes. Additionally, machine learning methods trained on such data tend to combine gene and variant properties implicitly, leading to inconsistent scale of scores across genes. To address these issues, we developed a method, FMVP (Fitness effect of Missense Variants by Protein embedding), to explicitly model the effect at molecular level (*D*) and population level (selection coefficient, *S*) separately, and estimate them jointly in a probabilistic graphical model. *D* is a non-linear function (*F_d*) of protein sequence context, *S* is a function (*F_s*) of *D* with gene-specific parameters, and *S* determines the distribution of allele counts in a population. We use embeddings from a protein language model pre-trained on millions of protein sequences to represent protein sequence context as predictive features for *D*, and use attention neural networks to approximate *F_d*. Given *D*, we approximate *F_s* by a Gaussian mixture model with gene-specific parameters. We use a Poisson-Inverse-Gaussian distribution to approximate distribution of allele counts conditioned on *S* and mutation rate, with parameters based on simulated European population prior to training. This model does not use pathogenicity labels. Instead, we trained the model using all possible missense variants in genes intolerant of loss of function variants to estimate parameters of *F_d* and *F_s*, based on likelihood of observed allele counts in large population genome data sets (145K UKBB and 90K gnomAD NFE), then performed variational inference to estimate posterior *S* for all possible missense variants in all genes. FMVP estimated *D* reached the state-of-the-art performance in classification of ClinVar pathogenic variants and in identifying damaging variants in deep mutational scanning assays with the most consistent distribution across genes. Finally, based on precision and recall, FMVP-estimated *S* achieves much better prioritization of *de novo* missense variants in neurodevelopmental disorders compared to previous methods, especially among the variants with estimated $S > 0.01$, a threshold that have a direct interpretation of genetic effect size in conditions that are often the primary driver of negative selection in humans. To our knowledge, this is the first attempt to estimate selection coefficient of individual missense variants using the latest genome data sets.

Session 026: CNVs in large-scale studies

Location: Conv Ctr/Room 202A/Level 2

Session Time: Thursday, November 2, 2023, 1:45 pm - 2:45 pm

Title: Copy number variant discovery from exome data in 637,229 individuals of diverse ancestries informs patterns of mutational constraint in the Genome Aggregation Database (gnomAD)

Authors:

J. Fu[1], R. Collins[2], I. Wong[3], C. Liao[1], L. Wang[4], S. Baxter[3], S. Chapman[3], D. Ben-Isvy[4], C. Stevens[3], C. Cusick[3], A. Sanchis-Juan[3], X. Zhao[1], M. Walker[3], G. Tiao[3], K. Chao[3], S. Gabriel[3], E. Banks[3], A. O'Donnell-Luria[3], D. MacArthur[5], H. Rehm[1], B. Neale[1], M. Daly[1], K. Samocha[1], K. Karczewski[3], H. Brand[1], M. Talkowski[1]; [1]Massachusetts Gen. Hosp., Boston, MA, [2]Dana Farber Cancer Inst., Boston, MA, [3]The Broad Inst., Cambridge, MA, [4]Harvard Med. Sch., Boston, MA, [5]Garvan Inst., Darlinghurst, Australia

Abstract:

The Genome Aggregation Database (gnomAD) has been an essential resource for the study of population diversity, models of natural selection, and variant interpretation in clinical genetics. However, while the largest and most diverse datasets currently accessible in the field involve single nucleotide variants (SNV) and indels, few reference resources at this scale have enabled accurate discovery and curation of individual gene- and exon-resolution copy number variants (CNVs). Here, we describe the completion and open release of rare CNV discovery from 637,229 exomes across diverse genetic ancestry groups in gnomAD v4, including 134,968 samples from non-European genetic ancestry groups.

We identified a median of 1 rare coding CNV (<1% site frequency) per person among 84,060 unique rare CNV sites in these individuals. We observed strong correlations between rare coding CNVs and metrics of loss of function (LoF) constraint (LOEUF) and dosage sensitivity (pHaplo and pTriplo), where deletions affecting a gene in the most constrained decile of LOEUF were 15x less common than those affecting the least constrained decile, with a similar 6x ratio for duplications, and alike results when compared to dosage sensitivity metrics.

Of the 84,060 unique rare CNV sites identified in gnomAD v4, 5,104 displayed significantly different carrier proportions by genetic ancestry group after multiple-correction. Consistent with previous studies of SNVs uncovering higher genetic diversity within African genetic ancestry groups, we observed 58% of individuals with African ancestries harboring at least one rare coding deletion (74% a duplication) compared to 35% of individuals with European ancestries harboring a rare coding deletion (53% a duplication). Importantly, corroborating previous work on SNV data (Simons et al. Nat Genet 2014), we observed no significant differences between genetic ancestry groups in the number of likely deleterious deletions overlapping constrained genes (LOEUF<0.4, average 0.02 per individual); while among the least constrained LOEUF decile, we observed a 4.3x difference in average number of genes deleted across groups, signaling the utility of cross-ancestry CNV rates in refining constraint metrics. Finally, we find that partial gene duplications detectable with exome sequencing were similarly depleted over genes that are estimated to be intolerant to LoF.

These data highlight the unique patterns of rare coding CNVs across population-scale cohorts and the critical utility of diverse ancestral reference populations for interpretation of deleterious variation in the human genome.

Title: Identification of structural and copy number variations linked to cancer susceptibility in a pan-cancer cohort from the 100k Genomes Project.

Authors:

R. Scott[1], G. Contino[1,2,3]; [1]Inst. of Cancer and Genomic Sci., Univ. of Birmingham, Birmingham, United Kingdom, [2]Univ. Hosp. of Birmingham Univ., Birmingham, United Kingdom, [3]Von Hugel Inst., Univ. of Cambridge, Cambridge, United Kingdom

Abstract:

Background: Understanding the heritability of cancer has significant implications for disease prevention, early detection, and personalized treatment. Genome-wide association studies (GWAS) based solely on single nucleotide polymorphisms explain only a modest fraction (2-15%) of the heritable variation in cancer risk. The missing heritability may reside in understudied non-coding variability and structural variations (SVs), including copy number variations (CNVs). However, the analysis of these features has been limited due to the lack of large-scale whole-genome sequencing datasets and the unavailability of suitable methods for detecting and investigating structural variations for association purposes. In this study, we leveraged the pan-cancer data from the 100k Genome Project of Genomics England (GEL) to develop computational methods for analysing the impact of structural variations and copy number variations on cancer heritability.
Methods: We developed an in-house pipeline to accomplish the following tasks: 1) filtering and constructing matched case-control cohorts for different cancer types within the cancer cohort and the rare diseases cohort in the Genomics England Research Environment (GEL); 2) identifying germline CNVs and SVs across the masked GRCh38 genome assembly with statistically significant differences between cases and controls; 3) annotating candidate loci.
Results: We selected 11,676 patients from 14 cancer types and matched them with 24,743 control participants from the rare disease cohort, after filtering for phenotypes with increased cancer risks, known CNVs or SVs, and proband status. We generated case-control cohorts for different cancer types, ensuring matching for sex, age, and ethnicity. Our analysis revealed several recurrent polymorphic CNVs and SVs that exhibited significant associations with specific cancer types. Some loci were associated in different tumor types, suggesting an implication in multiple cancer susceptibility. Notably, we found a statistically significant association between cancer risk and deletions in 1q24.2 and 12p13.31, and we identified novel candidate tumor-suppressor genes. Further validation using large independent datasets is currently underway.
Conclusions: This study introduces novel approaches to identify previously unknown candidate regions associated with cancer heritability.

Title: Association of mitochondrial DNA copy number with Blood Pressure Responses to Dietary Sodium and Potassium: The GenSalt Study (Genetic Epidemiology Network of Salt Sensitivity).

Authors:

X. Sun[1], Y. Pan[1], Z. Huang[2], R. Zhang[2], C. Li[2], J. He[2], T. Kelly[1]; [1]Univ. of Illinois at Chicago, Chicago, IL, [2]Tulane Univ., New Orleans, LA

Abstract:

Background: Decreased mitochondrial DNA copy number (mtDNA-CN) has been associated with an increased risk for hypertension in previous observational studies. However, whether decreased mtDNA-CN is associated with hypertension endophenotypes, including blood pressure (BP) responses to sodium, potassium, and cold pressor test, is unknown. **Methods**: We examined the associations between mtDNA-CN and BP responses to interventions among 1,855 participants from the Genetic Epidemiology Network of Salt-Sensitivity (GenSalt) study. After a 3-day baseline observation, participants underwent a 7-day low-sodium (51.3 mmol/day), 7-day high-sodium (307.8 mmol/day), and 7-day high-sodium plus potassium-supplementation (60 mmol/day). BP was assessed nine times during baseline and at the end of each intervention period. The cold pressor test was conducted during baseline. Linear mixed regression models were employed to assess the associations of mtDNA-CN with systolic BP (SBP), diastolic BP (DBP) and mean arterial pressure (MAP) responses to the interventions, after accounting for relatedness and adjusting for age, sex, body mass index (BMI), field center (FC), smoking and drinking status, physical activity, estimated glomerular filtration rate (eGFR), and baseline BP. After excluding those with hypertension at baseline, generalized linear mixed effect models were also used to test the associations of mtDNA-CN with development of hypertension and stage-2 hypertension over 8 years follow-up after adjusting for age, sex, BMI, FC, smoking, drinking, physical activity, baseline eGFR, and follow-up time. **Results**: Each standard deviation (SD) decrease in mtDNA-CN was associated with decreased responses of BP to the interventions. For example, each SD decrease in mtDNA-CN associated with a 0.5 mmHg decreased absolute SBP responses to the low-sodium intervention ($P=1.3\times10^{-3}$), respective 0.6 mmHg and 0.5 mmHg decreased DBP and MAP responses to the high sodium intervention ($P=2\times10^{-6}$ and 3.1×10^{-6}, respectively), respective 0.5 mmHg and 0.4 mmHg decreased absolute DBP and MAP responses to the high sodium plus potassium intervention ($P=1.4\times10^{-5}$ and 2.8×10^{-5}, respectively), and 0.5 mmHg decreased MAP response to the cold pressor test ($P=3.9\times10^{-3}$). As expected, each SD decrease in mtDNA-CN was associated with 41% increased odds of developing stage-2 hypertension ($P=4.9\times10^{-2}$). **Conclusions**: MtDNA-CN was associated with decreased BP response (or resistance of BP) to interventions. These data are congruent with recent research linking salt-resistance of BP to increased risk for hypertension development.

Title: Phenome-wide association studies of copy number variations in UK Biobank whole genomes

Authors:

Z. Zou[1], F. Hu[1], O. S. Burren[1], X. Jiang[1], S. Atanur[1], C. Salvoro[1], E. Oerton[1], A. Nag[1], S. H. Lewis[1], S. V. V. Deevi[1], S. O'Dell[1], M. Fabre[1], AstraZeneca Genomics Initiative, K. R. Smith[1], Q. Wang[2], S. Petrovski[1], K. Carss[1]; [1]Discovery Sci., R&D, AstraZeneca, Cambridge, United Kingdom, [2]Discovery Sci., R&D, AstraZeneca, Waltham, MA

Abstract:

DNA copy number gains and losses have profound functional consequences in human population diversity and genetic disorders. Previous phenome-wide association studies (PheWAS) using microarray and whole exome sequencing (WES) copy number variant (CNV) calls have revealed many clinically important genotype-phenotype associations. However, due to technology-specific limitations (e.g., microarray-based CNVs have poor resolution and sensitivity, whilst WES-based CNVs are biased towards protein-coding genomic regions), the full potential of CNV analysis in understanding the causes of human diseases is yet to be explored. To advance human genomics research, UK Biobank conducted whole genome sequencing (WGS) of approximately 500K participants (~5% non-European), which is currently the world's largest whole genome sequencing project. WGS enables much better coverage than array genotyping and WES, and thereby outperforming these technologies in CNV detection with unprecedented higher accuracy. In the present study, we called CNVs (>=10Kb) from ~500K UKB genomes and performed bespoke post-hoc filtering and re-genotyping to generate a high-quality CNV call set including ~70K deletions and ~60K duplications. By characterising the genomic CNV landscape, we found that >70% of CNVs are relatively small with sizes ranging from 10Kb to 50Kb, >95% of CNVs are rare (AF<1%) including ~50% of CNVs that are singletons (observed in only one sample), and >80% of CNVs do not span any protein coding genes. Employing different genetic models, we detected ~3000 known and novel associations (P<1*10-6) in coding and noncoding regions. Novel signals included a rare coding deletion at 15q13.3 associated with increased risk for asthma (P=1.25*10-9, binary odds ratio (OR)=10.11, 95% confidence interval (CI): 4.73-21.60) and a rare non-coding deletion at 5q34 associated with increased risk for heart failure (P=7.38*10-8, binary OR=194.89, 95% CI: 43.59-871.40). Our study not only generates and characterises the largest WGS-based CNV call set, a rich resource for further functional and mechanistic investigations on CNVs, but also presents the largest phenome-wide survey of dosage-sensitive regions in the human genome.

Session 027: Enemy within: Genetics of human autoimmune disorders

Location: Conv Ctr/Ballroom B/Level 3

Session Time: Thursday, November 2, 2023, 1:45 pm - 2:45 pm

Title: Unraveling the genetic basis of autoimmune hypothyroidism

Authors:

M. Bujnis[1], R. Sterenborg[2,3], L. Jorde[1], M. Medici[2], A. Teumer[4], ThyroidOmics Consortium; [1]Univ. of Utah, Salt Lake City, UT, [2]Erasmus Med. Ctr., Rotterdam, Netherlands, [3]Radboud Univ. Med. Ctr., Nijmegen, Netherlands, [4]Univ. Med. Greifswald, Greifswald, Germany

Abstract:

Introduction: Autoimmune hypothyroidism, also known as Hashimoto's thyroiditis (HT), is a common autoimmune thyroid disorder with a prevalence of 3-5% in the United States and Europe. Despite its high prevalence and high heritability (~70%), very little is known about the genetic basis of autoimmune hypothyroidism. To improve our understanding of the genetic signatures associated with HT development, we conducted a genome-wide association (GWAS) meta-analysis using ten large, independent case-control cohorts of over >1.2M unrelated individuals of multiple ethnicities. This represents the largest genetic study to date on autoimmune hypothyroidism. Methods: We conducted a GWAS meta-analysis including 43,113 cases and 1,196,957 controls. Cases were defined as individuals with an ICD9/10 code for autoimmune thyroiditis or unspecified hypothyroidism. Individuals without any thyroid disease were used as controls. The analyses were adjusted for age, sex, population structure, and related individuals. Secondary analyses included colocalization with mRNA levels using GTEx and DICE (Database of immune cell expression, epigenomics, eQTLs), functional enrichment and pathway analyses, protein-protein interaction networks, and genetic correlations. Results: We identified 143 independent SNPs passing the GWAS significance threshold ($p<5e-8$) associated with HT, five are novel associations with HT. The significantly associated loci explain 8.8% of the genetic variance. The genetic correlation between HT and TSH, FT4, and FT3 highlights the shared genetic component between HT diagnosis and variability in thyroid hormone measurements within the clinically normal range. We identified 1003 significant colocalizations (posterior probability > 0.85), with a majority found in brain and thyroid. Tissue enrichment and pathway analyses have revealed associations within the immune system, encompassing not only immune cells, such as T-cells, but also the spleen. We also found an enrichment of pathways involved in chromatin structure regulation. Conclusion: In performing a GWAS meta-analysis using the largest, multi-ethnic, and most precise case definition on HT, we have identified five novel genetic determinants and replicated many previous findings. Our analyses also highlight potential shared mechanisms underlying HT and other autoimmune diseases. These results improve our understanding regarding the genetic basis of HT and have the potential to help predict disease onset and personalize its treatment.

Title: Genome-wide association studies uncover novel genes, diverse disease mechanisms and druggable targets for juvenile idiopathic arthritis.

Authors:

J. Li, H-Q. Qu, J. Glessner, H. Hakonarson; the Children's Hosp. of Philadelphia, Philadelphia, PA

Abstract:

Juvenile idiopathic arthritis (JIA) is the most common rheumatologic disease among children. However, due to clinical heterogeneity, the genetic loci identified remain limited and only explain a small fraction of the genetic heritability. To identify novel loci and investigate underlying biological mechanism under JIA pathogenesis, we conducted a JIA GWAS meta-analysis on 4,550 JIA cases and 18,446 controls. We identified four novel genome-wide significant loci, each showing pleiotropic associations with several other autoimmune diseases and musculoskeletal traits. We prioritized target genes at each novel locus via *in silico* approaches. The targeted candidate genes, including *CD247*, *RHOH*, *COLEC10* and *IRF8*, at these four novel loci each may contribute to disease risk of JIA. We conducted genetic correlation analysis between JIA and the trait of bone mineral density (BMD) establishing local genetic correlation at six genomic regions, including two genome-wide significant loci for JIA and BMD. We employed genome-editing followed by RNA-seq to identify target genes at the novel locus on chromosome 4 and uncovered additional candidate genes at this locus. The analysis also suggested the importance of T-cell signaling underlying JIA pathogenesis. We further carried out cell-type enrichment analyses by integrating bulk and single-cell sequencing data. This analysis identified specific T-cell types as the critical cell types underlying JIA pathogenesis. Based on the GWAS data, we prioritized 24 candidate druggable target genes associated with JIA, providing several novel drug identifications or potential repurposing opportunities for JIA.

Title: Impact of parental autoimmune diseases on type 1 diabetes in offspring can be partially explained by genetic variants from HLA and non-HLA regions: a nationwide registry and biobank study

Authors:

F. Wang[1], A. Liu[1,2], FinnGen, T. Tuomi[1,3,4,5], A. Ganna[1,2,6,7]; [1]Inst. for Molecular Med. Finland, Helsinki, Finland, [2]Program in Med. and Population Genetics, Broad Inst. of Harvard and MIT, Cambridge, MA, [3]Abdominal Ctr., Endocrinology, Helsinki Univ. Hosp., Helsinki, Finland, [4]Folkhalsan Res. Ctr., Helsinki, Finland, [5]Lund Univ. Diabetes Ctr., Malmo, Sweden, [6]Stanley Ctr. for Psychiatric Res., Broad Inst. of Harvard and MIT, Cambridge, MA, [7]Analytic and Translational Genetics Unit, Massachusetts Gen. Hosp., Boston, MA

Abstract:

Patients with type 1 diabetes (T1D) and their relatives are often diagnosed with one or more autoimmune diseases (AIDs). However, it remains unsolved to which degree the shared genetic architecture contributes to the observed familial aggregation. We comprehensively investigated the relationship between T1D and 24 common AIDs by leveraging multi-generational health registers (N≈7.2 million) and genomic information (N≈430,000) from the Finnish population which has the highest T1D prevalence globally. Epidemiological analyses on 58,284 trios shows an increased risk of T1D among individuals whose parents were ever diagnosed with AIDs, six of the parental AID associations were significant after multi-test correction. The results are supported by genetic correlations and novel polygenic scores constructed from human leukocyte antigens (HLA) genotypes. For example, parents with coeliac disease are more likely to have children with T1D (OR=1.97 [1.75, 2.21) and we found a significant correlation between these two diseases both within HLA ($Coef_{HLA}$=0.61 [0.58,0.63], mostly driven by DRB1*03:01-DQA1*05:01-DQB1*02:01) and among non-HLA variants. Similarly, epidemiological evidences (OR=1.62 [1.42,1.85]) as well as non-HLA ($Rg_{non\text{-}HLA}$=0.53 [0.35,0.70]) and HLA variants ($Coef_{HLA}$=0.48 [0.43,0.53]) point to a share effect between parental vitamin B12 deficiency anemia and T1D in children. On the contrary, parental inflammatory bowel disease and multiple sclerosis are not associated with T1D in children in the registry-based analysis, but they have significantly negative associations with T1D in HLA regions. In conclusion, our results implicate that familial aggregation of AIDs can be caused by shared genetic mechanisms in both HLA regions and non-HLA regions. Our design can be extended to other diseases with shared genetic architecture.

Title: Autoimmunity risk variants exert dynamic allelic effects on the chromatin accessibility of stimulated CD4$^+$ T cells.

Authors:

M. Kono[1], H. Hatano[1], A. Suzuki[2], M. Nakano[1], A. Oguchi[3], Y. Murakawa[3], C. Terao[4], K. Yamamoto[2], K. Ishigaki[1]; [1]Lab. for Human Immunogenetics, RIKEN Ctr. for Integrative Med. Sci., Yokohama, Japan, [2]Lab. for Autoimmune Diseases, RIKEN Ctr. for Integrative Med. Sci., Yokohama, Japan, [3]RIKEN-IFOM Joint Lab. for Cancer Genomics, RIKEN Ctr. for Integrative Med. Sci., Yokohama, Japan, [4]Lab. for Statistical and Translational Genetics, RIKEN Ctr. for Integrative Med. Sci., Yokohama, Japan

Abstract:

Many autoimmune risk variants locate within accessible chromatin regions (ACRs) specific to CD4$^+$ T cells. However, previous studies conducted functional assays using CD4$^+$ T cells in limited cellular states and failed to detect the allelic effects of many risk variants. Here, to overcome these limitations, we purified CD4$^+$ T cells from 69 healthy donors and investigated the regulatory landscape of stimulated CD4$^+$ T cells by conducting an assay for transposase-accessible chromatin with sequencing (ATAC-seq) in four stimulus conditions at 0, 4, and 24hr time points (828 samples in total). The stimulus conditions included i) anti-CD3 and anti-CD28, ii) anti-CD3, anti-CD28, and anti-CD2, iii) anti-CD3, anti-CD28, and IL2, and iv) no stimulation. The ACRs detected by ATAC-seq explained the substantial fraction of the autoimmunity's heritability, especially for rheumatoid arthritis (RA). We tested the associations of the each ACR's read counts and its cis-variant genotype using a linear regression model and applied mashr software to jointly evaluate the association significance (lfsr < 0.05). Among 134,294 ACRs in total, 23,906 (17.8%) had at least one significant variant (chromatin accessibility quantitative trait locus: caQTL). In addition, we successfully detected 5,244 dynamic caQTLs specific to one stimulus condition. Integrative analyses with previous genetic study results highlighted 35 ACRs whose caQTLs colocalized with RA risk variants. Among colocalizing caQTLs, rs58107865 at the LEF1 intronic region is the fine-mapped RA risk variant with the highest accuracy (> 99% posterior probability) and showed dynamic caQTL effects specific to non-stimulation condition. We finally sought to experimentally confirm the causal role of rs58107865 on the chromatin accessibility at the LEF1 region and artificially induced both alleles of rs58107865 into human CD4$^+$ T cells by CRIPSR prime editing and achieved 25% editing efficiency on average. To evaluate the edited allele's impact on the chromatin accessibility, we developed a new ATAC-seq experimental platform that efficiently targets specific ACRs, accurately quantifies the allelic imbalance using unique molecular indexes, and enables us to pool multiple libraries using transposase with custom barcode sequences. By applying our platform to the edited alleles, we demonstrated that rs58107865 causally regulate chromatin accessibility of the LEF1 locus, which probably mediates the RA genetic risk. We thus elucidated molecular mechanism underlying multiple risk alleles of autoimmunity and provided novel insight into the autoimmunity pathology.

Session 028: Insights from systematic perturbation of the genome

Location: Conv Ctr/Ballroom A/Level 3

Session Time: Thursday, November 2, 2023, 1:45 pm - 2:45 pm

Title: Discovery of novel signaling-to-transcription networks in T cell activation coupling base editor screens with single cell multiomics

Authors:

S. Myers[1,2,3,4], **P. Kennedy**[1,5,6], A. Albozarian[1,2,6], M. Matias[1,2,6], M. Hedge[7], M. Balakar[7], M. Olive[7], N. Pirete[7], A. C. Jones[8], R. Burt[7], J. Levy[9], S. A. Carr[7], D. Liu[7], J. Doench[10], G. Newby[7], S. Alarcon[1,11]; [1]La Jolla Inst. for Immunology, San Diego, CA, [2]Ctr. for Autoimmunity and Inflammation, La Jolla Inst. of Immunology, La Jolla, CA, [3]Lab. for Immunochemical Circuits, La Jolla Inst. of Immunology,, La Jolla, CA, [4]Div. of Signaling and Gene Expression, La Jolla Inst. of Immunology, La Jolla, CA, [5]Ctr. for Autoimmunity and Inflammation, La Jolla Inst. of Immunology,, La Jolla, CA, [6]Lab. for Immunochemical Circuits, La Jolla Inst. of Immunology, La Jolla, CA, [7]Broad Inst. of MIT and Harvard, Cambridge, MA, [8]Dept. of Pharmacology, Univ. of California, San Diego, San Diego, CA, [9]Broad Inst. of MIT and Harvard, Cambridge, CA, [10]Broad Inst. of MIT and Harvard, San Diego, CA, [11]AUGenomics, San Diego, CA

Abstract:

The study of human genetics has been revolutionized by forward, genome-wide mutational screens, where individual gene knockouts and their contribution to a specific phenotype is examined in high throughput. However, the signal transduction pathways that drive these genetic circuits remain understudied, largely due to the lack of comparable methods to interrogate the vast number of protein phosphorylation events. While mass spectrometry is capable of profiling the dynamics of hundreds to thousands of phosphorylation sites that occur upon cell stimulation, functionally evaluating that scale of signaling events with practical throughput has remained elusive. Here, we present CRISPR/Cas9-mediated base editor screens to functionally assess tens of thousands of phosphorylation events and their contribution to specific gene expression programs. Using T cell activation as a model system, we demonstrate a “proteome-wide” functional phosphorylation screen to identify specific signaling events that inhibit or promote cell proliferation and survival, or specific gene expression programs. Combining cellular indexing of transcriptomes and epitopes (CITE-seq) and base editor screens, we present a novel approach to multiomic characterization of single cells to elucidate genotype-proteotype-phenotype relationships and untangle the complexities of the network of biochemical signaling reactions and how they lead to the control of specific cellular functions. We used a targeted CRISPR/Cas9-based method to selectively remove uninformative molecules upstream of Next Generation Sequencing (NGS) coupled with the AVITI System to enhance resolution in multimodal analyses. We show that mutation of phosphorylation sites previously unknown to be involved in T cell activation can influence the expression of genes associated with T cell exhaustion such as *PD1* and *NR4A*. This new experimental and computational framework will transform the way we study the biochemical networks that drive genetic circuits, and identify novel points of intervention to control specific cellular behaviors.

Title: Massively parallel CRISPR-based perturbation of a COPD risk-enriched region on chromosome 4q.

Authors:

V. Malik[1], L. Gong[1], Y. Yu[1], L. Dasilva[2], Q. Yao[3], D. Bauer[3], L. Pinello[4], E. Silverman[1], X. Zhou[1], M. Cho[1]; [1]Brigham and Women's Hosp., Boston, MA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]Boston Children's Hosp., Dana-Farber Cancer Inst., Harvard Med. Sch., Boston, MA, [4]MGH/Harvard/BROAD, Charlestown, MA

Abstract:

Background: Genome-wide association studies have identified more than 80 genomic regions associated with chronic obstructive pulmonary disease (COPD), including five loci clustered within 70 Mb in the 4q chromosomal region bounded by *BTC* and *HHIP*. However, the function of most of these genes in this region in lung-relevant cells is unknown.
Methods: We performed Perturb-Seq, i.e. CRISPRi-based perturbations followed by single cell RNA sequencing (scRNA-seq) targeting genes in the 4q COPD GWAS region in human bronchial epithelial cells (16HBE). We analyzed data from two libraries and used Sceptre to analyze the cis- and trans- effects.
Results: We detected expression of 194/221 and 181/373 targeted genes in our two libraries and identified statistically significantly decreased cis-expression of 133 (89%) and 122 (67%) targeted genes, respectively. We also identified 10 and 15 significant trans-effects, respectively. Among the genes showing trans-effects, *AIMP1* and *ABCE1* were located within the COPD GWAS locus. Significant trans-effects included *RPL34* (60S ribosomal protein L34) perturbation associated with increased level of expression of *MRPL1*, *MRPS18C* and *CHIC2*. In turn, reduced *RPL34* expression was observed by perturbation of *ABCE1* and *POLR2B*. *POLR2B* perturbation reduced expression of *MRPL1* and *HNRNPDL*. Trans-interactions when combined in a network were enriched in the ribonucleoprotein compartment; GO functions like nucleic acid and heterocyclic compound binding; and GO processes like RNA metabolic processes, gene expression and translation. **Conclusion**: In a COPD susceptibility gene-enriched region, we were able to perturb expression of 255 genes in human bronchial epithelial cells. These perturbations had several identifiable trans-effects. Future work will perturb COPD-associated cis-regulatory regions and the role of likely effector genes on the whole transcriptome of these cell lines.
Keywords: Perturb-seq, COPD, scRNA-seq, CRISPR interference, Human bronchial epithelial cells

Title: CRISPR-CLEAR - In-Situ Investigation of Genotype-to-Phenotype Relationship with Nucleotide Level Resolution CRISPR saturation mutagenesis screens

Authors:

L. Pinello[1], B. Becerra[2], M. Jankowiak[3], S. Wittibschlager[4], A. Karjalainen[4], A. P. Kutschat[4], T. Wu[5], M. Starrs[5], Z. Patel[1], D. Bauer[6], D. Seruggia[4]; [1]Massachusetts Gen. Hosp./Harvard Med. Sch./BROAD, Boston, MA, [2]Massachusetts Gen. Hosp./Harvard Med. Sch./Boston Children's Hosp., Boston, MA, [3]BROAD, Boston, MA, [4]St. Anna Children's Cancer Res. Inst. (CCRI) & CeMM Res. Ctr. for Molecular Med. of the Austrian Academy of Sci., Vienna, Austria, [5]Boston Children's Hosp., Boston, MA, [6]Boston Children s Hosp., Harvard Med. Sch., BROAD, Boston, MA

Abstract:

In this study, we present the development of a genome editing tool called CRISPR-CLEAR for the in-situ investigation of genotype-phenotype relationships at nucleotide and variant-level resolution. This tool combines CRISPR-Cas9 technologies with novel computational strategies to enable high-resolution interrogation of sequence variants, to ultimately link phenotypes to their causal mutations.

As a proof of concept, we investigated a regulatory element upstream of CD19 using NALM-6 cells, a model of B-cell leukemia, using a library of 200 sgRNAs that tiled a candidate enhancer, and followed by FACS sorting cells into CD19+ or CD19- populations according to their CD19 expression levels based on a fluorescent CD19-conjugated antibody. We have additionally investigated three non-coding regulatory regions of fetal hemoglobin (HbF) induction using HUDEP-2 erythroid progenitor cells, with densely tiled sgRNA libraries for each region, and followed by FACS sorting into HbF high, medium, and low populations. All screens were performed using nucleases and state-of-the-art base-editors (TadA-derived adenine, cytosine, and dual base-editors) with extended PAM recognition (SpG, SpRY). For each sample, we sequenced both the sgRNA (for perturbation counts) and the endogenous targeted region (for direct allele readout) all in tandem to compare each approach. By comparing the sgRNA count, or direct alleles between sorted populations, we are able to identify functional sub-regions at variant-level resolution.

To examine the relationship between the genotype and phenotype at nucleotide level resolution, we developed a Bayesian linear regression model based on a recent statistical method called millipede, which assigns importance scores to each mutation based on its abundance in either sorted population. Our model shows superior signal detection and functional resolution as compared to current count-based methods that only assess sgRNA enrichment. Using this approach, we uncovered and validated in NALM-6 cells a hotspot of mutations at a 20 bp region corresponding to relevant transcription factor binding sites and novel HbF regulatory regions in HUDEP-2 cells, which were consistent across biological replicates and perturbation used.

In conclusion, CRISPR-CLEAR offers a powerful tool for investigating genotype-phenotype relationships by directly observing edits at the endogenous locus with higher resolution compared to count-based methods. We envision that this framework will provide a comprehensive solution for the classification of non-coding variants of uncertain significance and the discovery of causal regulatory elements.

Title: Gene regulatory networks constructed from CRISPR perturbations in primary CD4+ T cells elucidate the genomic basis of autoimmune disease

Authors:

J. Weinstock[1,2], M. Arce[3], J. W. Freimer[2,3], M. Ota[2,3], A. Marson[3], A. Battle[1,4], J. K. Pritchard[2,5]; [1]Dept. of BioMed. Engineering, Johns Hopkins Univ., Baltimore, MD, [2]Dept. of Genetics, Stanford Univ., Stanford, CA, [3]Gladstone-UCSF Inst. of Genomic Immunology, San Francisco, CA, [4]Malone Ctr. for Engineering in Hlth.care, Johns Hopkins Univ., Baltimore, MD, [5]Dept. of Biology, Stanford Univ., Stanford, CA

Abstract:

The effects of genetic variation on complex traits manifest predominately through the contribution of regulatory variation. Although we have mapped many genetic variants to molecular phenotypes in cis, the trans-regulatory cascade mediating their effects on complex traits remains largely uncharacterized. Despite the importance of trans-regulatory variation, mapping trans-regulators based on natural genetic variation, including eQTL mapping, has been challenging due to small effects and limited sample size. Experimental perturbation approaches, which are unconstrained by natural selection, offer a complementary and powerful approach to mapping trans-regulators. We used arrayed CRISPR knockouts (KOs) of 84 genes in primary CD4+ T cells to experimentally perturb an immune cell gene network and assessed effects with bulk RNA-seq. Our KO gene set included three groups: 1) 24 upstream regulators of IL2RA; 2) 30 transcription factors (TFs) associated with inborn errors of immunity (IEI); 3) 30 background TFs that are matched in constraint and expression level to the IEI TFs but do not have a known immune disease association. We then developed a novel Bayesian structure learning method to estimate the gene regulatory network, which in contrast to many causal inference approaches, is not constrained to directed acyclic graphs, enabling us to capture feedback loops. We systematically characterized the differences between the IEI and background TFs, finding that the gene groups were highly interconnected, but that IEI TFs were much more likely to regulate immune cell specific machinery and genes proximal to autoimmune GWAS loci. We observed 211 directed edges among the 84 genes, which were validated using orthogonal data modalities, and could not be detected in existing CD4+ trans-eQTL data.From our network, we characterized nine coherent gene programs. One program included core JAK-STAT signaling regulators and KMT2A, a global epigenetic regulator that was upstream of IL2RA, IRF4, and STAT5A. KMT2A cooperated with IL2 signaling TFs to regulate downstream effector pro-inflammatory cytokines, including IL17. Heritability analyses indicated that the SNPs linked to this module were strongly enriched for heritability of immune phenotypes. These analyses reveal the trans-regulatory cascade from upstream epigenetic regulator to intermediate TFs to downstream effector cytokines and elucidate the logic linking autoimmune GWAS loci to key signaling pathways. We anticipate that our approach can be used to characterize transcriptional logic in other cell types and contexts.

Session 029: Life in the fast lane: Epigenetic aging and age acceleration in disease

Location: Conv Ctr/Room 207A/Level 2

Session Time: Thursday, November 2, 2023, 1:45 pm - 2:45 pm

Title: Longitudinal changes in intron retention in a twin cohort are widespread and highlight mechanisms of ageing

Authors:

D. Wang[1], Y. Raza[1], A. Ramisch[2,3], A. Roberts[1], G. Leday[4], Y. Jiao[5], G. Nicholson[5], M. Stevens[2,3], M. Abdalla[5], C. Menni[1], C. Holmes[5], T. Spector[1], M. McCarthy[5,6], S. Richardson[4], E. Dermitzakis[2,3], J. El-Sayed Moustafa[1], K. Small[1]; [1]King's Coll. London, London, United Kingdom, [2]Univ. of Geneva, Geneva, Switzerland, [3]Inst. of Genetics and Genomics in Geneva, Geneva, Switzerland, [4]Med. Res. Council Biostatistics Unit, Cambridge, United Kingdom, [5]Univ. of Oxford, Oxford, United Kingdom, [6]Oxford Univ. Hosp. Trust, Oxford, United Kingdom

Abstract:

Ageing is a primary risk factor for chronic diseases and is characterised by cellular and molecular dysregulation, including broad changes to gene expression. Alternative splicing, including mis-splicing, is highly responsive to the environment and is more predictive of age than total gene expression levels. To investigate age-related changes in mis-splicing, we quantified intron retention in longitudinal RNA-seq data from the MultiMuTHER study, examining blood samples collected from 335 female twins at three time points over nine years, with an average age of 60 at the first visit. We employed SUPPA to quantify the usage of individual retained introns and summarised the median level of all retained introns to give a global estimate of mis-splicing per sample. Linear mixed-effect models were fitted to test whether individual introns and per sample median intron retention changed over time. Our findings revealed a significant increase in the median level of intron retention over time (P = 1.8e-4), indicating a global increase in mis-splicing within an individual. At individual introns, we observed longitudinal changes in retained introns at 466 genes (FDR <0.05). Intriguingly, 45% of genes with a time-associated retained intron did not show longitudinal changes in overall gene expression, highlighting the importance of investigating changes in splicing alongside overall gene expression in ageing studies. Notably, 88% of time-associated retained introns showed an increase with time, suggesting a striking increase in the prevalence of mis-splicing over time. In contrast, time-associated gene expression levels were relatively evenly split between increasing and decreasing. To explore the genetic regulation of intron retention, we utilised 54 dizygotic and 44 monozygotic twin pairs to calculate the heritability of retained introns at each of the three time points with twin models. Additionally, to investigate if genetic effects on intron retention change over time, we constructed linear mixed-effect models including an interaction term between SNPs and time from baseline. Though the heritability of retained introns ranged from 0 to 0.94, ~70% of intron retention events were lowly heritable (h^2<0.1). While 1,987 (61.6%) retained introns were associated with at least one sQTL, none of them showed an interaction with time. Collectively, our results suggest that environmental factors might be the primary contributors to the longitudinal changes in intron retention and that investigating longitudinal changes in intron retention could provide unique insights into the molecular mechanisms of ageing and age-related diseases.

Title: Development of the first epigenetic clock in tooth tissues: a better understanding of tooth aging.

Authors:

S. C. Zapico[1], A. Kulpa[2], W. M. Freeman[2]; [1]New Jersey Inst. of Technology, Newark, NJ, [2]Oklahoma Med. Res. Fndn., Oklahoma City, OK

Abstract:

Despite the general knowledge that our dental tissues undergo different alterations as a result of aging, there are limited studies trying to understand the molecular mechanisms of tooth aging. Additionally, there are a few works focused on one of the trends in aging research right now, epigenetics. Previous studies in other tissues pointed out that epigenomic patterns, particularly DNA methylation, the addition of a methyl group to a cytosine nucleotide in cytosine-guanine dinucleotides, plays a key role in aging and the development of aging-related diseases. Many attempts are underway to apply epigenetic-based approaches in reducing the deteriorative effects of aging and prevent or treat aging-related diseases, leading to a "healthy aging". However, few studies have been focused on dental tissues and those that have used pyrosequencing technologies to examine small regions of the genome. Nonetheless, these reports demonstrate methylation differences among different dental tissues. The present study went beyond these previous works developing for the first time an epigenetic clock for tooth tissues from a human cohort (18 to 90 years old), through the application of Illumina Infinium MethylationEPIC BeadChip interrogation of 850,000 CpG sites in dentin and pulp tissues. Age-related changes in methylation, both hyper- and hypo-methylation from young to older human adults were identified. Existing epigenetic ‘clocks’ demonstrated moderately accurate age predictions and confirming differences in methylation between these two tissues. This proof-of-concept study is the first step to gain a better understanding of the molecular mechanisms of tooth aging. Further research will be able to expand these results, increasing the number of subjects, and studying the impact of these methylation changes on the transcriptome and proteome.

Title: Epigenetic Age Acceleration and Midlife Cognitive Function: Evidence from Observational Study and Mendelian Randomization.

Authors:

Z. Huang[1], Y. Pan[2], X. Sun[3], I. D. Anda-Duran[1], R. Zhang[1], W. Chen[1], C. Li[4], D. Bennett[5], O. T. Carmichael[6], L. Bazzano[1], T. Kelly[7]; [1]Tulane Univ., New Orleans, LA, [2]UIC, Chicago, IL, [3]Tulane Univ., Chicago, IL, [4]Tulane Univ. Sch. of Publ. Hlth.and Tropical Med., New Orleans, LA, [5]Rush Univ. Med. Ctr., Chicago, IL, [6]Pennington BioMed. Res. Ctr., New Orleans, LA, [7]Univ. of Illinois at Chicago, Chicago, IL

Abstract:

Background and Objectives The early detection of pre-clinical cognitive decline in midlife poses a major challenge and opportunity for timely and targeted interventions for the prevention of Alzheimer's disease (AD) and related dementia syndromes. Epigenetic age acceleration (EAA) is a powerful indicator of biological aging that has emerged as a biomarker for numerous chronic diseases. In this study, we examined the association between EAA and cognitive function within a biracial cohort of middle-aged participants enrolled in the Bogalusa Heart Study (BHS), followed by a two-sample Mendelian randomization (MR) study to investigate potential causal effects of EAA on AD in later life. **Methods** EAA measures, including intrinsic EAA (IEAA), Hannum Age acceleration (HannumAA), PhenoAge acceleration (PhenoAA), and GrimAge acceleration (GrimAA), were derived from genome-wide DNA methylation profiling data collected within the Bogalusa Heart Study (BHS) using the Illumina EPIC array. Linear regression models were employed to examine the associations between EAA and cognitive function, using normalized scores from individual cognitive tests, domain-based scores, and a global cognitive score (GCS). A two-sample MR further investigated possible causal effects of EAA on the risk of AD in later life, leveraging data from a GWAS meta-analysis of EAA(N=34,710) and two independent GWAS studies of AD (N=63,926 and 309,154). **Results** Among 1,252 middle-aged BHS participants, greater HannumAA, GrimAA and PhenoAA were significantly associated with slower processing speed ($p=0.017$, $p<0.001$, and $p=0.002$ respectively), after adjusting for demographic, educational, behavior and clinical covariates. Additionally, increased GrimAA was associated with lower GCS ($p=0.001$) after accounting for known risk factors. Sex modified the effects of GrimAA and PhenoAA on executive function ($P_{interaction}=0.004$ and 0.002, respectively), with inverse associations observed among females ($P=0.003$ and 0.034, respectively) but not males. IEAA was not associated with any cognitive domains or tests. MR study demonstrated a 29% elevated risk of AD ($p = 0.003$) conferred by each five-year increase of genetically predicted HannumAA. **Discussion** EAA measures, including HannumAA, GrimAA and PhenoAA, were associated with lower midlife cognitive function independent of known risk factors. The MR study further supported a causal relationship between EAA and AD.

Title: Epigenetic age prediction using the Asthma&Allergy DNA methylation array reveals enrichments for immune pathways and accelerated epigenetic age among children with asthma

Authors:

E. Thompson[1], E. Zhong[1], A. A. Eapen[2], K. Rivera-Spoljaric[3], R. L. Miller[4], G. K. Khurana Hershey[5], E. Zoratti[2], D. R. Gold[6], C. C. Johnson[7], D. J. Jackson[8], J. E. Gern[8], C. G. McKennan[9], C. Ober[1]; [1]The Univ. of Chicago, Chicago, IL, [2]Henry Ford Hlth., Detroit, MI, [3]Washington Univ., St Louis, MO, [4]Icahn Sch. of Med. at Mount Sinai, New York, NY, [5]Cincinnati Children's Med. Ctr., Cincinnati, OH, [6]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA, [7]Henry Ford Hlth.System, Detroit, MI, [8]Univ. of Wisconsin Sch. of Med. and Publ. Hlth., Madison, WI, [9]Univ. of Pittsburgh, Pittsburgh, PA

Abstract:

DNA methylation (DNAm) is a widespread epigenetic mark that has emerged as one of the most effective biomarkers to predict chronological age, with DNAm-based prediction models of age referred to as epigenetic clocks (ECs). Several ECs have been published using CpGs on the Illumina 450k or EPIC arrays. The goal of this study was to build an EC using CpGs on the Asthma&Allergy DNAm array, which consists of ~40k CpGs that are not on the EPIC array and are enriched for functional regions of the genome and proximity to immune-related genes (Morin et al. *JACI* June 2023). DNAm was assayed in nasal lavage (mostly immune) cells from 960 subjects ranging in age from 11-33 years (median = 14.8 years; 25% with asthma, defined as ever having received a diagnosis of asthma from a doctor). We built an EC using elastic net with a training set of 525 individuals, including sex, white blood cell count, and sample plate as non-penalized covariates (Pearson correlation between chronological and predicted age = 0.97, median error = 0.68 years). This provided an accurate prediction of age in the remaining 312 test samples (Pearson correlation = 0.87, median error = 1.31 years). The 217 CpGs selected by elastic net were characterized by intermediate methylation levels (% DNAm ranging from 0.2-0.8), compared to the most widely used EC CpGs from the 450k and EPIC arrays that are characterized by low methylation levels (% DNAm <0.10) across tissues. The 186 genes nearest the 217 selected CpGs on the Asthma&Allergy array were enriched for multiple pathways relative to genes nearest to all CpGs on the array (FDR<0.05; iPathway Guide), including Th17, Th1, and Th2 cell differentiation, MAPK signaling, and inflammatory bowel disease pathways, highlighting immune processes and immune-mediated diseases. Finally, we tested for associations between asthma and deviations between predicted and chronological age. Accelerated aging was not associated with asthma in the full sample but was associated with increased asthma among the youngest half of the sample (<14.8 years; OR = 1.28, 95% CI [1.07-1.55]; *P*=0.007). Our study revealed that CpGs on the Asthma&Allergy DNAm array provide excellent prediction of epigenetic age. Given that previous ECs were comprised primarily of hypomethylated CpGs, the Asthma&Allergy array may capture different aspects of aging. Our findings suggest that the relationship between advanced epigenetic age and asthma in children may be driven by differential DNA methylation of immune response pathways.

Session 030: New genetic insights into long-standing congenital phenotypes

Location: Conv Ctr/Room 145A/Level 1

Session Time: Thursday, November 2, 2023, 1:45 pm - 2:45 pm

Title: Cleft palate probands are enriched for protein-altering *de novo* variants

Authors:

K. Robinson[1], S. Curtis[1], T. Beaty[2], A. Butali[3], C. Buxo[4], D. Cutler[1], M. Epstein[1], J. Hecht[5], G. Shaw[6], L. Moreno[3], J. Murray[3], H. Brand[7], S. Weinberg[8], M. Marazita[9], E. Leslie[1]; [1]Emory Univ., Atlanta, GA, [2]Johns Hopkins Univ, Sch PubHlth, Baltimore, MD, [3]Univ. of Iowa, Iowa City, IA, [4]Univ. of Puerto Rico, San Juan, PR, [5]Univ Texas McGovern Med Sch, Houston, TX, [6]Stanford Univ., Stanford, CA, [7]MGH, Boston, MA, [8]Univ of Pittsburgh, Pittsburgh, PA, [9]Univ Pittsburgh, Ctr. for Craniofacial and Dental Genetics, Pittsburgh, PA

Abstract:

Cleft palate (CP) is a common craniofacial anomaly, occurring in 1 in 1700 live births. Although CP is highly heritable, relatively little is known about genetic risks for the condition. Protein-altering *de novo* variants (DNs) are associated with multiple structural birth defects, but there has been no large-scale investigation of their role in CP to date. Whole genome sequencing of 488 case-parent trios was used to identify protein-altering DNs on an exome-wide and per-gene basis using denovolyzeR. There were 597 protein-coding DNs in 572 genes with 70% of probands harboring at least one variant. There was significant enrichment for missense (1.26, $p=6.90x10^{-6}$), predicted loss-of-function (pLOF, 1.59, $p=2.12x10^{-4}$), and all protein-altering (i.e., combined missense and pLOF) variants (1.30, $p=1.55x10^{-8}$). Genes were then investigated for individual enrichment and were considered statistically significant if they reached a multiple-testing threshold correcting for 572 genes with at least one DN ($p<8.7x10^{-5}$), or a more conservative exome-wide threshold ($p<1.3x10^{-6}$). Two genes reached exome-wide significance: *COL2A1* for pLOF ($p=2.18x10^{-13}$) and protein-altering ($p=2.73x10^{-9}$) DNs and *IRF6* for protein-altering DNs ($p=6.71x10^{-7}$). Five additional genes were significantly enriched: pLOF variants in *MYH3* ($p=1.77x10^{-5}$) and protein-altering variants in *PRKCI* ($p=1.85x10^{-6}$), *SAT2B* ($p=3.04x10^{-6}$), *POLR1F* ($p=5.65x10^{-5}$), and *SLC25A41* ($p=7.98x10^{-5}$). No difference in DN enrichment was found for male versus female probands. However, stratification by CP subtype revealed differing patterns. Clefts of the hard palate only (n=47) were enriched for missense (1.55, p=0.003) but not pLOF variants (1.46, p=0.232), whereas clefts of the soft palate only (n=172) were enriched for pLOF variants (1.86, p=0.002) but not missense variants (1.050, p=0.306). DNMs were not significantly enriched for clefts affecting both the hard and soft palate (n=101) or submucous CP (n=29). Lastly, isolated CP (n=453) was compared to syndromic cases (n=35). Syndromic cases were enriched for missense (1.54, p=0.011) and protein-altering (1.55, p=0.006) variants, with *PRKCI* reaching exome-wide significance for protein-altering variants ($p=1.29x10^{-6}$). Interestingly, all individuals with DNs in *COL2A1*, *MYH3*, or *SAT2B* were in the nonsyndromic group indicating these individuals were not known to have any of the features associated with syndromes caused by mutations in these genes. Ongoing analysis of DNs combined with rare variants within these same genes will contribute additional knowledge to our understanding of the etiology and pathogenesis of CP and its subtypes.

Title: Novel molecular diagnoses in individuals with holoprosencephaly and prior negative sequencing

Authors:

A. Cohen[1,2], J. J. Hughes[3], C. L. Dalgard[4], P. Kruszka[5], E. Delot[2,6], V. Fusaro[7], E. Vilain[7], M. Muenke[8], S. I. Berger[2,6]; [1]NHGRI/NIH, Bethesda, MD, [2]Ctr. for Genetic Med. Res., Children's Natl. Hosp., Washington, DC, [3]NIH, Bethesda, MD, [4]Dept. of Anatomy, Physiology & Genetics and The American Genome Ctr., Uniformed Services Univ. of the Hlth.Sci., Bethesda, MD, [5]GeneDx, Gaithersburg, MD, [6]Dept. of Genomics and Precision Med., George Washington Univ. Sch. of Med. and Hlth.Sci., Washington, DC, [7]Inst. for Clinical and Translational Sci., Univ. of California, Irvine, Irvine, CA, [8]American Coll. of Med. Genetics & Genomics (ret.), Bethesda, MD

Abstract:

Holoprosencephaly (HPE) is a congenital disorder characterized by failure of forebrain division during early fetal development and subsequent structural and functional defects. The causes of HPE have proven difficult to elucidate. HPE is associated with environmental and genetic factors and the phenotype shows highly variable expressivity within and across families. Syndromic forms of HPE account for roughly 45-75% of HPE cases and are associated with chromosomal anomalies, copy number variants, and monogenic syndromes. Isolated HPE is associated with heterozygous variants at least 15 genes. In addition, oligogenic inheritance has been reported in a small number of HPE families. The genetic heterogeneity and variable expressivity of HPE complicate identification and classification of new HPE variants. Current estimates of diagnostic yield for HPE gene sequencing panels in individuals with isolated HPE are around 20%, leaving a significant number of affected families without a molecular diagnosis that would enable risk assessment and family planning.

We analyzed genome sequencing from a holoprosencephaly cohort consisting of over 420 individuals in more than 160 families. The majority of probands in this cohort had negative HPE gene panel, exome sequencing, or both, prior to genome sequencing. For this analysis, we identified small variants (up to 50bp) and structural variants (over 50bp) predicted to affect known HPE genes.

To date, we have identified likely or candidate genetic diagnoses for more than 20% of the families in this HPE cohort. This included variants that would not be detectable on clinical HPE gene panels or exome, such as deep intronic splicing variants, repeat tract insertions, and both coding and non-coding region structural variants. Many of these variants have not previously been reported in HPE. These findings highlight the value of genome sequencing to identify novel, putative pathogenic variants that are not detectable by available clinical testing and improve diagnostic yield in genetically heterogeneous disorders such as HPE.

Title: Core Planar Cell Polarity Genes VANGL1 and VANGL2 in Predisposition to Congenital Scoliosis

Authors:

Y. Ye[1,2,3], X. Feng[4], Z. Zhao[3], G. Li[3], K. Xu[3], J. Cai[3], Q. Li[3], J. P. Cheung[4], T. Zhang[3], B. Gao[5], N. Wu[3]; [1]Guangdong Provincial People's Hosp., Guangzhou, China, [2]Southern Med. Univ., Guangzhou, China, [3]Peking Union Med. Coll. Hosp., Beijing, China, [4]The Univ. of Hong Kong, Hong Kong, China, [5]The Chinese Univ. of Hong Kong, Hong Kong, China

Abstract:

Introduction: Impaired somitogenesis leads to CVMs, which clinically manifest as congenital scoliosis (CS), a birth defect that affects 0.5-1 of 1,000 live births. Wnt/ß-catenin signaling is central to somitogenesis, whereas the role of Wnt/PCP signaling in somite development remains unclear.
Methods: First, we evaluated somitogenesis defects and vertebral malformations in Vangl mutant mouse embryos. Second, we analyzed exome sequencing data from multi-center and multi-ethnic CS patients and confirmed their loss-of-function and dominant-negative effects in cellular assays. Further, we addressed the in vivo functional significance of the most deleterious variants in both zebrafish and mouse models.
Results: Here, we show that deletion of two core PCP components, Vangl1 and Vangl2, leads to defective somitogenesis and spinal malformation that mimics the conditions of human CVMs. We identified a number of rare variants of VANGL1 and VANGL2 in CS patients and observed loss-of-function and dominant-negative effects among these variant alleles. The failure of mutant VANGL mRNA to rescue convergent extension defects in zebrafish models confirmed the variants' pathogenicity. Moreover, Vangl1 knock-in (p.R258H) mice exhibited vertebral malformations in a Vangl gene dose- and environment-dependent manner. Removal of one Vangl2 allele or perinatal hypoxia treatment significantly increased the penetrance of CVMs in Vangl1-R258H mice. A Vangl gene-environment interaction model that controls the formation of the axial skeleton was proposed.
Conclusion: Our studies revealed new critical roles for PCP signaling in somitogenesis and predisposition to CVMs in CS patients.

Title: Transcriptome phenotype analysis based on RNA-Seq increases the diagnostic rate and resolves variants of uncertain significance in patients with undiagnosed infantile spasms

Authors:

H. Tang[1], Z. Pan[2,3], J. Peng[2,3], C. Chen[1]; [1]Ctr. for Med. Genetics & Hunan Key Lab. of Med. Genetics, Sch. of Life Sci., and Dept. of Psychiatry, The Second Xiangya Hosp., Central South Univ., Changsha, China, [2]Dept. of Pediatrics, Xiangya Hosp. of Central South Univ., Changsha, China, [3]Clinical Res. Ctr. for Children Neurodevelopmental Disabilities of Hunan Province, Xiangya Hosp., Central South Univ., Changsha, China

Abstract:

Infantile spasms (IS) or West syndrome is a severe epileptic condition that is commonly associated with epileptic encephalopathy and has unique clinical and electrographic features, and affects children in the middle of the first year of life. While a handful of genes have been linked to IS, many patients are left without molecular diagnosis due to the limitations of whole exome sequencing (WES). Even whole-genome sequencing (WGS), which offers the most extensive genomic examination, including non-coding regions containing regulatory elements, generates variants that are oftentimes difficult to interpret. However, a large portion of these methodological limitations may be mitigated by the functional information that RNA sequencing offers.This pilot study aimed to develop an integrative analysis method using DNA and RNA sequencing to increase the diagnostic success rate of rare genetic mutations in IS. Moreover, we sought to identify novel genes associated with IS with a more robust variant examination pipeline. One hundred and twelve IS families lacking a molecular diagnosis despite trios WES were recruited. Trio whole-genome and RNA-sequencing were performed on peripheral blood mononuclear cells extracted from blood samples to identify variants and outliers in expression, splicing, and alternative polyadenylation events. Integration of RNA-seq analysis allowed us to identify and validate the outlier genes in six of 110 WGS inconclusive individuals with disease (5.5%, including one expression outlier, three splicing outliers, and two alternative polyadenylation outliers). We also identified candidate genes potentially linked to the disease phenotype in 12 cases (11.3%). Additionally, multiple candidate variants were discovered in genes that are not yet reported pathogenic with IS, and further investigation is needed to determine the pathogenicity potential in these variants.In conclusion, this pilot study demonstrates that analysis of aberrant transcriptome phenotypes can increase the diagnostic rate and resolves variants of uncertain significance in patients with undiagnosed infantile spasms.

Session 031: Somatic mutation in health and disease

Location: Conv Ctr/Ballroom C/Level 3

Session Time: Thursday, November 2, 2023, 1:45 pm - 2:45 pm

Title: The landscape of human somatic variability across genomes, tissues, and individuals

Authors:

H. Xu, R. Bierman, D. Akey, J. Akey; Princeton Univ., Princeton, NJ

Abstract:

A human zygote undergoes trillions of cell divisions to result in a fully differentiated multicellular organism. This intricate process creates numerous opportunities for somatic mutations to occur. Even terminally differentiated cells continue to acquire somatic lesions in their genomes through non-replicative processes, such as spontaneous deamination of methylated cytosines in the context of CpG sites. As a result, humans are expected to harbor a significant burden of somatic variation, which contributes to various traits and diseases, including cancer, aging, neurodegeneration, and an expanding list of additional human diseases. Although somatic mutations are fundamentally important to human biology and disease, many outstanding questions remain about their rates, spectrum, and determinants. The two main impediments to a more comprehensive understanding of human somatic mosaicism are the lack of appropriate data sets where multiple tissues from multiple individuals have been sampled and powerful statistical methods tailored to the unique structure of this data. Here, we describe the analysis of somatic mutations in a data set comprised of high-coverage whole-genome sequencing of 300 tissue samples (spanning on average six tissues from 50 donors) using a novel probabilistic approach to identify somatic variants. We first show that our method has higher power and lower false positive rates compared to existing approaches. Next, we apply our method to the whole-genome sequence data and identify a compendium of somatic mutations that we leverage to quantify the rates and mutational spectra of somatic variants across tissues and individuals. By integrating functional genomics data with our compendium of somatic mutations, we show that the burden of deleterious somatic variants varies widely across individuals. We also show that age is significantly related to the burden of somatic mutations, but only in a subset of tissues. Finally, we show that the rates and spectrum of somatic mutations are significantly different between coding and noncoding regions, and how chromatin structure contributes to this difference. In summary, our data provides a high-resolution compendium of somatic mutations across genomes, tissues, and individuals and provides novel biological insights into mutational mechanisms.

Title: Genome-wide detection of somatic mosaicism at short tandem repeats

Authors:

A. Sehgal, M. Gymrek; Univ. of California, San Diego, La Jolla, CA

Abstract:

Somatic mosaicism, in which a mutation occurs post-zygotically, has been implicated in several developmental disorders, diseases and cancers. Short tandem repeats (STRs) consist of repeated sequences of 1-6bp and comprise more than 1 million loci in the human genome. Somatic mosaicism at STRs is known to play a key role in the pathogenicity of loci implicated in repeat expansion disorders, and is highly prevalent in cancers exhibiting microsatellite instability. While a variety of tools have been developed to genotype germline variation at STRs, **no method currently exists** for systematically identifying mosaic STRs (mSTRs). Here, we introduce **DeSASTR**, a novel method for detecting mSTRs from individual next-generation sequencing datasets. **DeSASTR** models the observed distribution of copy numbers in reads at a particular locus as a mixture of germline and somatic alleles. It applies a statistical test to identify STRs exhibiting somatic mosaicism, and additionally infers the maximum likelihood mosaic fraction,size of the mosaic allele, and haplotype on which the mosaic allele likely arose . Unlike many existing mosaicism detection methods for other variant types, **DeSASTR** does not require a matched control sample as input. We validated **DeSASTR** using simulated read data and found that we could accurately detect mosaicism down to *2%* variant allele fraction at *50x* coverage. To validate the entire **DeSASTR** pipeline starting from raw reads, we also developed a new simulation framework, **simTR,** which simulates raw sequencing reads according to a specified coverage level and stutter error model using user-defined repeat alleles. Our analysis confirmed that *accuracy* in *f (mosaic fraction)* detection and *power of detection* improves with *increase in coverage and f value.* We applied **DeSASTR** to detect mosaic STRs in whole genome sequencing (WGS) from *NA12878* (derived from a lymphoblastoid cell line) and identified a total of *1725* significant mSTRs out of *1301421* loci. We validated these results using *haplotagged Pacbio Hifi* data, which confirmed *713/1534* (*163/262* non-homopolymer STRs), given some mosaic sites lacked enough reads in the *PacBio HiFI* data to be processed by **HipSTR**. The majority of **mSTRs (81.04%)** resulted in insertions or deletions of a single repeat unit. Finally, we analyzed mSTRs using *WGS* from *179, 178, 93* samples across *CEU, YRI and CDX* populations from the *1000 Genomes Project*, which *revealed variation* in the *number of mosaic sites* across these populations. Overall, **DeSASTR** provides a powerful method for detecting mSTRs from individual whole genome sequencing datasets.

Title: SoMoSeq- A novel method for genotype informed single cell RNA sequencing of mosaic brain tissue

Authors:

M. Gade[1], D. Lai[2], B. Porter[3], A. Poduri[4], H. Won[1], W. Ma[1], E. Heinzen[1]; [1]UNC Chapel Hill, Chapel Hill, NC, [2]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [3]Stanford Univ., Palo Alto, CA, [4]Boston Children's Hosp, Boston, MA

Abstract:

Post-zygotically acquired somatic variants are contributors to a range of neurological disorders. The mosaic brain tissue in individuals harboring these variants allows for the study the cell type involvement and cell type-specific effects of pathogenic variants on transcription. To leverage mosaic brain tissue to understand how somatic variants contribute to disease, we developed Somatic Mosaic single-cell RNAseq (SoMoSeq), for genotyping pathogenic variant loci and sequencing full-length cDNA in single nuclei isolated from human brain tissue. SoMoSeq is a modified version of G&Tseq which uses biotinylated-oligodT-conjugated streptavidin coupled magnetic beads to physically separate DNA and RNA. The resulting DNA is amplified and genotyped with TaqMan assay, while the RNA is utilized to generate full length cDNA using SMARTseq2. We evaluated the efficacy of our approach using mosaic brain tissue samples harboring known pathogenic somatic variants: *PIK3CA* (E545K, VAF=28%) and *SLC35A2* (Y145X, VAF=10%) and three age matched autopsy/stroke controls. After nuclei isolation and fluorescence-activated nuclear sorting with neuronal marker (NeuN), we used SoMoSeq to genotype and generate cDNA from 3072 nuclei. Genotyping analysis accurately reflected the variant allele frequencies (VAF) observed in bulk nuclei in neuronal(NeuN+) and non-neuronal(NeuN-) cells. In parallel, we generated high-quality cDNA from single nuclei (average yield of 1ng/uL). From the NeuN+ and NeuN- populations, we selected 96 cells positive/negative for the variants (N=384/sample) and sequenced them at a depth of 8-10M reads/cell. Approximately 10,000 genes/nucleus were detected with an average mitochondrial read percentage of 3%, indicating the high quality of the data. We also observed significant upregulation of the mTOR signaling pathway in the *PIK3CA* sample compared to controls and evidence of the enrichment of mutations in NeuN- cell types in the mosaic *SLC35A2* brain tissue. Cell type involvement results align with those reported in literature and further validate the accuracy and significance of our method for investigating the effects of pathogenic somatic mutations on cell type-specific transcriptional profiles. Data integration of full length RNAseq with publicly available datasets from human cortex allowing for the classification of specific cell types harboring variant alleles are ongoing. SoMoSeq is a powerful approach that allows for precise identification of specific cell types harboring somatic variants and the ability to study cell type-specific autonomous and non-cell autonomous transcriptional effects of the variant in human brain tissue.

Title: Deep whole-genome sequencing of GTEx tissues reveals developmental patterns and somatic evolution

Authors:

T. Coorens[1], D. Firer[1], O. Priebe[1], J. Hess[1], F. Aguet[2], G. Getz[3], K. Ardlie[4]; [1]Broad Inst. of MIT and Harvard, Cambridge, MA, [2]Illumina, Inc., Foster City, CA, [3]Broad Inst MIT & Harvard, Cambridge, MA, [4]Broad Inst., Cambridge, MA

Abstract:

From fertilization onwards, the cells of the human body continuously experience DNA damage and accumulate somatic mutations. The somatic genome of a cell is a record of its life history: mutations shared with other cells indicate a shared ancestry and can be used to retrace early development. Early somatic mutations have also been linked to various disease phenotypes, such as childhood cancer and developmental disorders.
Recent studies on a few donors have used somatic mutations to show that embryonic cells contribute asymmetrically to the adult body, such that one daughter cell of the zygote has twice as many descendant cells as the other, likely due to cellular bottlenecks in embryogenesis. It is unclear how variable this pattern is across the human population. Here, we sequenced 308 whole genomes derived from six tissue types of 55 donors within the GTEx cohort to a mean depth of 195x. These tissue types span the embryonic germ layers: ectoderm (brain and skin), mesoderm (heart) and endoderm (esophagus, thyroid and lung).
Facilitated by the breadth in sampling and depth of sequencing, we can efficiently detect embryonic mutations in all donors and estimate the contribution of embryonic progenitor cells to the donors in our cohort. We observe a highly variable asymmetry in zygotic daughter cell contribution, hinting at a stochastic bottleneck during early embryogenesis. We model that the major source of this asymmetry is the early split between trophectoderm and the inner cell mass, but further bottlenecks during gastrulation may modulate contributions to specific germ layers. Our analysis also detects large clonal expansions in thyroid, esophagus and skin, each harboring distinct imprints of known mutagenic processes. The severity of these expansions is highly variable, with one notable clone in skin carrying over 20,000 somatic single-base substitutions. Most of these expansions can be explained by somatic mutations in genes known to be under selection, such as NOTCH1.
Taken together, this study reveals patterns of embryonic development and later somatic evolution from deep whole-genome sequencing data of normal tissues and will serve to provide context to diseases rooted in abnormal development. As costs for whole-genome sequencing continually decrease and technologies advance, the approach outlined here will substantially increase our understanding of human cellular evolution in health and disease.

Session 057: Comparative omics to explore tumor landscapes

Location: Conv Ctr/Ballroom B/Level 3

Session Time: Friday, November 3, 2023, 10:45 am - 12:15 pm

Title: Utilizing nullomers in cell-free RNA for cancer detection.

Authors:

A. Montgomery[1], G. Tsiatsianis[2], I. Mouratidis[1], C. Chan[3], M. Athanasiou[2], V. Kantere[2], N. Yee[1], I. Georgakopoulos-Soares[1]; [1]Penn State Coll. of Med., Hershey, PA, [2]Natl. Technical Univ. of Athens, Athens, Greece, [3]UCSF, San Francisco, CA

Abstract:

Early diagnosis of cancer can significantly improve survival of cancer patients; however, most cancer types still lack sensitive and highly specific non-invasive biomarkers needed for detection. cell-free RNA (cfRNA) may improve upon current biomarkers due to over-representation of highly expressed tumor-associated genes compared to its cell-free DNA (cfDNA) counterpart. Nullomers are short sequences absent from the human genome. As nullomers may resurface due to somatic mutations, they could provide more sensitive and specific biomarkers for cancer detection. Here, we examine over 10,000 whole exome sequencing matched tumor-normal samples to characterize nullomer resurfacing across exonic regions. We find that 29,774,302 different sixteen base-pair nullomers appear in this cohort with ~80% of somatic mutations causing a nullomer to resurface. We also identify nullomer resurfacing mutational hotspots within cancer genes and report that certain mutational signatures are more likely to cause nullomer resurfacing than others. We provide evidence that nullomers can be used to identify neoepitopes and other precision oncology targets. We identify the most frequent 100,000 resurfacing nullomers (for each of 14-16 bp) as a feature space for classifying hepatocellular carcinoma cancer (HCC) samples from cfRNA. We use an L1 regularized logistic regression model with 10 fold cross-validation repeated 100 times as a model. We achieve AUROC scores of 0.998, 0.999, and 1.000 for the models made of 14 bp, 15 bp, and 16 bp nullomers. Each model also shows accurate probabilistic predictions with Brier scores less than or equal to 0.02. We examine the nullomers which occur in over 90% of the repeated cross-validated models and annotate many of them to liver cancer associated genes including *FTH1*, *EEF2*, *TMSB10*, *ACTB* and the long non-coding RNA *MALAT*. We then use a separate dataset to create lasso logistic regression models to classify liver (AUC=0.922), stomach (AUC=0.927), and lung (AUC=0.877) cancer samples against healthy samples. We believe these results show the utility of nullomers within cfRNA as a method of detecting cancer non-invasively.

Title: Single-cell transcriptomics reveals interactions of the tumor microenvironment and infiltrating immune cells in high-grade glioma

Authors:

H. Natri[1], S. Gholamin[2,3], L. Peter[1], M. Aftabizadeh[2], M-i. Chung[1], B. Badie[2], C. Brown[2], N. Banovich, E[1]; [1]Translational Genomics Res. Inst., Phoenix, AZ, [2]City of Hope Beckman Res. Inst. and Med. Ctr., Duarte, CA, [3]California Inst. of Technology, Div. of Biology and Biological Engineering, Pasadena, CA

Abstract:

Glioblastoma multiforme (GBM) is an aggressive form of brain cancer with a post-diagnosis life expectancy of only 14 to 16 months. Chimeric Antigen Receptor (CAR) T cell therapy has emerged as a promising approach to treating cancer, including GBM and other gliomas. However, the tumor microenvironment (TME) poses a challenge for CAR T therapy for the treatment of solid tumors. Anti-tumour immunity relies on cell-cell interactions within the TME, with tumor and stromal cells as well as tumor-associated macrophages (TAMs) playing important roles in mediating T cell recruitment and function. Cell type and single cell level profiling of the TME may uncover new modifiable targets to improve the efficacy of CAR T and other immune therapies. To this end, leveraging single cell RNA sequencing (sc-RNAseq), we have produced transcriptomic profiles of 30,503 cells from baseline tumors from 32 patients with recurrent or refractory malignant glioma who participated in a Phase I study on locoregionally-delivered CAR T cells targeting a glioma-associated antigen IL13Rα2. We find that a higher proportion of fibroblast-like astrocytes and suppressive macrophages is associated with lower proportions of tumor-infiltrating lymphocytes (TILs). A ligand-receptor analysis reveals differences in cell-cell signaling associated with the abundance of TILs: tumors with low numbers of TILs were characterized by a downregulation of the JAK/STAT3 pathway, mediated by upregulation of laminin, collagen, and osteopontin (SPP1) signaling between fibroblasts and suppressive macrophages. Further, upregulation of osteopontin, as well as ligands associated with stem-like properties, was observed in tumors from patients who did not achieve benefit from CAR T therapy. The upregulation of osteopontin in progressive disease was further verified by immunofluorescent staining. This study demonstrates the role of osteopontin as a regulator of tumor-associated macrophages and cancer-associated fibroblasts in GBM and highlights its role in shaping the TME and modulating lymphocyte infiltration. Targeting osteopontin in conjugation with CAR T and other immunotherapies has the potential to result in better outcomes across GBM and other cancers.

Title: Utilizing TimiGP for in-depth analysis of the tumor immune microenvironment and its association with clinical outcomes of various cancers after immunotherapy

Authors:

C. Li[1], W. Hong[2], B. Zhang[3], A. Reuben[1], L. Wang[1], J. Zhang[1], C. Cheng[2]; [1]UT MD Anderson Cancer Ctr., Houston, TX, [2]Baylor Coll. of Med., Houston, TX, [3]Rice Univ., Houston, TX

Abstract:

Cancer remains a formidable global health adversary, with prognostic outcomes hinging on the complex and dynamic interplay between neoplastic cells and the host's anti-tumor immunity. The pivotal role of the tumor immune microenvironment (TIME) in these interactions has been scrutinized in recent years, which wields considerable impact on both the onset and progression of malignancies, as well as the potency of various therapeutic modalities, most notably, immunotherapy. In an endeavor to unravel the complex roles and interactions of immune cells within the TIME, we developed TimiGP (Tumor Immune Microenvironment Illustration based on Gene Pairing), which draws inspiration from the dynamic equilibrium within the immune system. TimiGP is a computational method designed to infer cell-cell interaction networks and estimate associations between various cellular contexts and clinical outcomes using transcriptomics and relevant clinical data. Initially developed and evaluated in a prognostic context, we have optimized this methodological framework to discern associations between immune cell interactions and responses to immunotherapy. We employed this optimized methodology in cohorts of 3,072 patients with 5 distinct cancer types (melanoma, non-small cell lung cancer, renal cell carcinoma, metastatic urothelial carcinoma and hepatocellular carcinoma), treated with 9 different immunotherapies. We elucidated the complexities of the TIME at various resolutions, encompassing immune-tumor interactions, the interplay among immune cells, and interactions within T cell subtypes. Of note, we observed a heightened similarity across various cancer types in interaction networks at the T cell level, in line with the fact that the majority of immunotherapy regimens in this study targeted T cells. Furthermore, we identified *CD4*+ *GZMK*+ effector memory T cells as one of the critical cell types across 5 cancer types and 9 immunotherapies, which may be a potential drug target and biomarker to improve the immunotherapy efficacy. Besides, we compared interaction networks related to the cancer vaccine, single and combined immune checkpoint inhibitors (anti-CTLA-4, anti-PD-1, anti-PD-L1), and their combinatory treatments. Across different types of immunotherapies, we not only identified consistent interactions within T cell subtypes but similar interactions between T cells and innate immune cells (e.g., NK Cells, Neutrophils). This broad application has enabled a thorough exploration of the intricacies of immune cell interactions and their associations with immunotherapy responses across different cancer types and therapeutic interventions.

Title: Identification and analysis of cell-specific expressed genetic variants from scRNA-seq data

Authors:

A. Horvath[1], A. Kim[1], K. Saito[2], Z. Yu[1], H. Arestakesyan[1], E. Ulianova[1], N. Edwards[3]; [1]GWU, Washington, DC, [2]GWU, Northeastern University, MA, [3]Georgetown Univ., Washington, DC

Abstract:

Low cellular frequency variants may indicate pre- or early-somatic clonality in cancer and normal tissues, or cell-specific RNA variance. Currently, most genetic variation is analyzed from bulk sequencing datasets, where insufficient sensitivity prevents the identification of variants in low numbers of cells.To address challenges posed by low frequency variation events, we have developed a computational framework for identification and analysis of Single Cell-specific Expressed Single Nucleotide Variants (sceSNVs) from single cell RNA-sequencing (scRNA-seq) data. Central for the framework is our new tool SCExecute, which enables the execution of various software designed for bulk sequencing data on barcode-stratified, extracted on-the-fly, single-cell alignments. Applying SCExecute in conjunction with tools for analysis of bulk sequencing data, we explored, for the first-time, cell-level expressed genetic variation across 28 publicly available cancer and normal datasets, including prostate cancer, non-small cell lung carcinoma, cholangiocarcinoma, neuroblastoma, normal fetal adrenal and normal embryo. This analysis identified more than 100,000 previously unreported SNVs, over 70% of which could not be called with any other approach. To assess the origin of the novel variants - DNA or RNA - we applied an innovative machine learning algorithm, which classified the majority of them as either posttranscriptional modifications, or, notably, non-random transcriptional infidelity. Approximately 10% of the novel sceSNVs show preferential non-random expression in particular cell clusters, where they correlate with significantly increased expression of their harboring gene and the related gene-regulatory networks. We find preferentially expressed sceSNVs to be frequent in genes involved in the DNA-repair, replication, and cell cycle. We exemplify our analyses using a novel missense substitution - 6:26104128_G>T, located in a gene encoding a core histone (*HIST1H4C*^{V61F}). We demonstrate that the V61F expression transcript is correlated to high expression of the gene *HIST1H4C*, and a deregulation of the *HIST1H4C*-related gene network; this observation is consistent across multiple cancer samples.Our findings suggest that there is an unappreciated repertoire of cell-level expressed nucleotide variation, strikingly recurrent and common across samples, that participates in transcriptome function and dynamics. Their appearance and, for some, relationship to certain gene-sets and cell types, suggests novel mechanisms and function for the expressed genetic variation, including in cancer progression and cell fate.

Title: Clonal selection and diversification of somatic mutations in normal human endometrial epithelium

Authors:

H. Nakaoka[1], M. Yamaguchi[2], K. Suda[2], K. Yoshihara[2]; [1]Sasaki Inst., Chiyoda-ku, Japan, [2]Niigata Univ., Niigata, Japan

Abstract:

It has become evident that somatic mutations in cancer-associated genes accumulate in the normal endometrium, but spatiotemporal understanding of the evolution and expansion of mutant clones is limited. To elucidate the timing and mechanism of the clonal expansion of somatic mutations in cancer-associated genes in the normal endometrium, we sequenced 1,311 endometrial glands from 37 women. We showed that the burdens of somatic mutations with single base substitution signatures (SBS1, SBS5, and SBS18) increased with age and the mutations in cancer-associated genes were under strong positive selection in the normal human endometrium. By collecting endometrial glands from different parts of the endometrium, we showed that multiple glands with the same somatic mutations occupied substantial areas of the endometrium. By using three-dimensional imaging analysis, we demonstrated that “rhizome structures”, in which the basal glands ran horizontally along the muscular layer and multiple vertical glands rose from the basal gland, originated from the same ancestral clone. Moreover, mutant clones detected in the vertical glands diversified by acquiring additional mutations. These results suggest that clonal expansions through the rhizome structures are involved in the mechanism by which mutant clones extend their territories. Furthermore, we showed clonal expansions and copy neutral loss-of-heterozygosity events occurred early in life, suggesting such events can be tolerated many years in the normal endometrium. Based on the result of the target gene sequencing, we examined whole genome sequencing for the endometrial glands with *KRA*S mutations exhibiting allelic imbalance, tumor suppressor gene mutations, and microsatellite instability. The analyses of mutational signatures and the temporal order of these genomic events highlighted the inter-gland heterogeneity in the acquisition of the genomic alterations even within the endometrium of the same individual. Our results of the evolutionary dynamics of mutant clones in the human endometrium will lead to a better understanding of the mechanisms of endometrial regeneration during the menstrual cycle and the development of therapies for the prevention and treatment of endometrium-related diseases.

Title: Genomes and epigenomes of matched normal and tumor breast tissue reveal diverse evolutionary trajectories and tumor-host interactions

Authors:

R. Yang[1], B. Zhu[1], A. Tapinos[2], H. Koka[1], P. Lee[3], T. Zhang[1], W. Zhu[1], K. Wang[1], A. Klein[1], D. Lee[1], M. Hua[1], D. Wang[1], W. Luo[1], K. Jones[1], A. Hutchinson[4], B. Hicks[5], M. Garcia-Closas[1], S. Chanock[6], L. Tse[3], **D. Wedge**[2]; [1]NCI, NIH, Bethesda, MD, [2]The Univ. of Manchester, Manchester, United Kingdom, [3]The Chinese Univ. of Hong Kong, Hong Kong, Hong Kong, [4]NCI Core Genotyping Facility, Bethesda, MD, [5]Leidos Biomed/DCEG/NCI/NIH, Bethesda, MD, [6]Natl. Cancer Inst, Rockville, MD

Abstract:

Normal tissues adjacent to the tumor (NAT), which harbor genomic and epigenomic alterations that likely represent early events in breast carcinogenesis, may provide valuable insights into cancer initiation and evolution. Although previous studies have characterized copy number and transcriptomic alterations, the evolutionary history of NAT in breast cancer (BC) remains unclear. In this study, we conducted comprehensive whole-genome sequencing (WGS), methylation profiling, and RNA sequencing analysis of paired germline, NAT, and tumor samples from 43 Chinese BC patients in Hong Kong (HK). The mean age at BC diagnosis was 59.2 years, with 43.9% luminal A, 26.8% luminal B, 14.6% HER2-enriched, 12.2% basal-like, and 2.5% normal-like subtypes based on PAM50 classification. The most frequently mutated genes in both tumor and NAT samples were *PIK3CA*, *TP53*, *GATA3*, and *AKT1*. Single nucleotide variants (SNVs) were common in NAT, with one-third of NAT samples exhibiting SNVs in driver genes, many of which were also present in paired tumor samples. The most recurrent mutation in NAT samples was *PIK3CA* H1047R, detected in 3 of 43 WGS samples and 18 of 214 NAT samples in subsequent ultra-deep panel sequencing of HKBC patients. In contrast, large-scale aberrations such as somatic copy number alterations (SCNAs) and structural variants (SVs) were rarely detected in NAT samples. To investigate the evolutionary history, we generated phylogenetic trees of paired NAT and tumor samples. Our analysis revealed eight phylogenetic tree patterns, which could be broadly categorized into three groups: present in tumor-only, shared between tumor and NAT, and multiple trees that are specific to tumor or NAT. These groups exhibited distinct genomic and epigenomic characteristics in both NAT and tumor samples. NAT samples belonging to the multiple tree group showed a less inflammatory tumor microenvironment (TME), characterized by lower presence of CD14 cell populations and decreased expression of CD14 downstream markers such as CD163 and IL-8. In tumor samples, the multiple tree group had higher tumor mutational burden, SV count, higher percentage genome with SCNAs, and a higher frequency of luminal B and basal-like tumors compared to the shared tree group. Mutations in known BC driver genes such as *GATA3* were exclusively found in the shared tree group, while mutations in others such as *ARID1A* were unique to the multiple tree group. Additionally, specific SCNA and SV signatures exhibited distinct patterns in different tree groups. Our findings highlight the diverse evolutionary history in BC NAT/tumor pairs and their potential impact in shaping the genomic landscapes of tumors and the TME.

Session 058: Mixtape of statistical genetics greatest hits

Location: Conv Ctr/Ballroom A/Level 3

Session Time: Friday, November 3, 2023, 10:45 am - 12:15 pm

Title: Detecting and adjusting for hidden biases due to phenotype misclassification in genome-wide association studies.

Authors:

D. Burstein[1], G. Hoffman[1], D. Mathur[1], S. Venkatesh[1], K. Therrien[1], A. H. Fanous[2], T. Bigdeli[3], P. Harvey[4], P. Roussos[1], G. Voloudakis[1]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Dept. of Psychiatry, Univ. of Arizona Coll. of Med.-Phoenix, Phoenix, AZ, [3]SUNY Downstate Hlth.Sci. Univ., Brooklyn, NY, [4]Miami VA Hlth.care System, Miami, FL

Abstract:

Background: With the advent of healthcare-based genotyped biobanks, genome-wide association studies (GWAS) leverage larger sample sizes and introduce noisier phenotypic definitions. For example, phenotypic misclassification dilutes the effect size estimates from GWA studies, as the mislabeled phenotype inflates the similarity between the cases and controls, resulting in diluted effect sizes. Yet the extent and impact of cross-study dilution on large-scale datasets is not currently well understood due to a lack of statistical methods to estimate relevant parameters from empirical data. **Methods:** We develop a statistical method and scalable software, PheMED, Phenotypic Measurement of Effective Dilution, to quantify effective phenotypic dilution across GWASs using only summary statistics. We apply our methodology to detect multiple instances of statistically significant dilution in real-world data. We then propose a methodological extension of inverse-variance weights meta-analysis using dilution-adjusted weights (DAW), to adjust the weight of each study according to its effective dilution and benchmark it against alternative methodologies. **Results:** We observe multiple instances of statistically significant dilution across our use-cases (median $p=1.59x10^{-9}$). For example, when comparing the dilution between GWASs using the presence of one (inclusive) or at least two (strict) phecode counts for bipolar disorder in the Million Veteran Program, the effect sizes in the more inclusive GWAS were on average 1.52 times smaller than the strict GWAS, $p =1.35x10^{-13}$. We find that our DAW meta-analysis yields a 25% median increase in FDR-significant loci compared to competing methodologies across five different use-cases. Furthermore, we note that DAW is competitive or outperforms existing methodologies when comparing the number of significant hits that are validated through an independent cohort as well. **Conclusions:** In consideration of our findings, we anticipate that PheMED will serve as a valuable approach for integrating GWA studies with phenotypic quality heterogeneity to guide algorithmic phenotyping and correct for diluted gene-trait associations that can further bias downstream meta-analyses.

Title: Accurate rare variant phasing of 200,031 UK Biobank whole-genomes

Authors:

R. Hofmeister[1], D. Ribeiro[1], S. Rubinacci[2], O. Delaneau[3]; [1]Univ. of Lausanne, Lausanne, Switzerland, [2]Brigham and Women's Hosp., Boston, MA, [3]Univ. of Lausanne, Lausanne-Dorigny, Switzerland

Abstract:

Background: Investigating the impact of rare variations on complex traits requires large collections of samples with whole-genome or whole-exome sequencing (WGS/WES). Recently, the UK Biobank released sequence data for over 200,000 samples, with more than 750 million variant sites. The availability of haplotypes for this dataset offers vast opportunities for disease and population genetics, but estimating the haplotypes (i.e phasing) for this large amount of data implies significant computational and statistical challenges, especially for the large proportion of rare variants found is such data collections (~97% have a minor allele frequency (MAF) lower than 0.1%).
Results: To address these challenges, we present SHAPEIT5, to efficiently and accurately phase large-scale sequencing datasets, accounting for both family data when available and variable chromosome X ploidy. On the UK Biobank's WGS and WES data, we phase variants found in one individual out of 100,000 with 95% accuracy. Using the resulting haplotypes as a reference panel for genotype imputation, we greatly enhance imputation accuracy compared to the Haplotype Reference Consortium reference panel. Variants with a minor allele frequency of 0.01% see a four-fold increase in imputation accuracy, which in turn enhances the power of downstream GWAS. In addition, these haplotypes allow for the detection of 549 genes with loss-of-function compound heterozygotes (i.e, complete gene knockouts), which we show to complement current knowledge of gene essentiality in the human genome.
Conclusion: We showcase how to efficiently infer and leverage haplotypes in large sequencing datasets. We estimated the haplotypes for the 200,031 UKB WGS samples, which will be included in an upcoming UK Biobank data release. In addition, we provide pipelines to impute SNP arrays and low-coverage sequencing data on the UK Biobank Research and Analysis Platform (RAP) using our haplotypes as a reference panel.
Grant references: Swiss National Science Foundation PP00P3_176977

Title: StocSum: stochastic summary statistics for whole genome sequencing studies.

Authors:

H. Chen[1], N. WANG[1], B. Yu[1], G. Jun[1], Q. Qi[2], R. A. Durazo-Arvizu[3], S. Lindstrom[4], A. Morrison[1], R. Kaplan[2], E. Boerwinkle[1]; [1]The Univ. of Texas Hlth.Sci. Ctr. at Houston, Houston, TX, [2]Albert Einstein Coll. of Med., Bronx, NY, [3]Children's Hosp. Los Angeles, Los Angeles, CA, [4]Univ. of Washington, Seattle, WA

Abstract:

Genomic summary statistics, usually defined as single-variant test results from genome-wide association studies, have been widely used to address different scientific questions in genetic and genomic research, such as meta-analysis, heritability estimation, conditional analysis, variant set and gene-based tests, multiple phenotype analysis, genetic correlation or co-heritability estimation. Applications that involve multiple genetic variants also require their correlations or linkage disequilibrium (LD) information, often obtained from an external reference panel. While these methods usually have good performance for common variants in populations of European ancestry, in practice, it is usually difficult to find suitable external reference panels that represent the LD structure for isolated, underrepresented and admixed populations, or rare genetic variants from whole genome sequencing (WGS) studies, limiting the scope of applications for genomic summary statistics. We have developed StocSum, a novel reference-panel-free statistical framework for generating, managing, and analyzing stochastic summary statistics using random vector algorithms. Regardless of the complex sample correlation structure, StocSum always scales linearly with both the sample size and the number of genetic variants in computing stochastic summary statistics from individual-level data. We develop various downstream applications using StocSum including single-variant tests, conditional association tests, gene-environment interaction tests, variant set tests, as well as meta-analysis and LD score regression tools. The complexity of all these downstream applications does not depend on the sample size. We demonstrate the accuracy and computational efficiency of StocSum using two cohorts from the Trans-Omics for Precision Medicine Program. Specifically, we show that StocSum can be used to perform long-range variant set tests, expanding the aggregation units beyond genes or genomic regions in close proximity. We also show that for admixed populations, LD scores estimated by StocSum are much more accurate compared to those from external reference panels, even if all ancestry populations are included in those reference panels. In summary, as a reference-panel-free framework, StocSum will facilitate sharing and utilization of genomic summary statistics from WGS studies, especially for isolated, underrepresented and admixed populations.

Title: MultiSTAAR: A statistical framework for powerful rare variant multi-trait analysis in biobank-scale sequencing studies

Authors:

X. Li[1], H. Chen[2], M. Selvaraj[3], E. Van Buren[4], K. Rice[5], J. Rotter[6], G. Peloso[7], P. Natarajan[8], Z. Li[9], Z. Liu[10], X. Lin[11], TOPMed Lipids Working Group; [1]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [2]The Univ of Texas Hlth.Sci. Ctr. at Houston, Houston, TX, [3]MGH, Boston, MA, [4]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA, [5]Univ of Washington, Seattle, WA, [6]Lundquist Inst., Harbor-UCLA Med Ctr, Torrance, CA, [7]Boston Univ., Boston, MA, [8]Massachusetts Gen. Hosp., Boston, MA, [9]Indiana Univ. Sch. of Med., Indianapolis, IN, [10]Columbia Univ., New York, NY, [11]Harvard T.H. Chan Sch Pub Hlth, Boston, MA

Abstract:

Introduction
Biobank-scale sequencing studies have made it feasible for better understanding rare variant contributions to complex human traits and diseases. Leveraging association strengths across multiple traits in rare variant association analysis of sequencing studies can improve statistical power over single-trait analysis and detect pleiotropic genes or noncoding regions. Existing methods have limited ability to perform rare variant multi-trait analysis when applied to biobank-scale sequencing data.
Methods
We propose MultiSTAAR, a powerful statistical framework and computationally scalable analytical pipeline for functionally-informed rare variant multi-trait analysis in biobank-scale sequencing studies. As a statistical framework, MultiSTAAR accounts for relatedness, population structure and correlation between phenotypes by jointly analyzing multiple traits, and further empowers rare variant association analysis by incorporating multiple functional annotations. As a comprehensive and robust analytical pipeline, MultiSTAAR facilitates functionally-informed multi-trait analysis of both coding and noncoding rare variants by incorporating multiple variant functional annotations for grouping and weighting. MultiSTAAR also provides conditional multi-trait analysis to dissect rare variant association signals independent of known variants.
Results
We applied MultiSTAAR to perform whole-genome sequencing rare variant analysis of 61,838 ancestrally diverse participants from 20 studies by jointly analyzing three quantitative lipid traits from the NHLBI TOPMed consortium: LDL-C, HDL-C and TG. In gene-centric multi-trait analysis of rare variants, MultiSTAAR identified 43 conditionally significant associations with lipid traits, including 4 noncoding associations (enhancer DHS rare variants in *NIPSNAP3A* and *LIPC*; ncRNA rare variants in *RP11-310H4.2* and *MIR4497*) that were missed by any of the three single-trait functionally-informed analysis using STAARpipeline. In genetic region multi-trait analysis of rare variants, MultiSTAAR identified 7 conditionally significant 2-kb sliding windows associated with lipid traits, including two sliding windows in *DOCK7* (chromosome 1: 62,651,447 - 62,653,446 bp; chromosome 1: 62,652,447 - 62,654,446 bp) and an intergenic sliding window (chromosome 1: 145,530,447 - 145,532,446 bp) that were missed by single-trait analysis using STAARpipeline.
Summary
In summary, MultiSTAAR provides a powerful statistical framework and a computationally scalable analytical pipeline for multi-trait analysis of biobank-scale sequencing studies with complex study samples.

Title: An integrative approach with Perturb-seq, eQTL, and GWAS data identifies mediating genes in regulatory networks of complex traits

Authors:

Z. Lu[1], D. Yao[2], B. Cleary[2], N. Mancuso[3], A. Gusev[2]; [1]Univ. of Southern California, Los Angeles, CA, [2]Dana-Farber Cancer Inst., Boston, MA, [3]Univ. of Southern California, South Pasadena, CA

Abstract:

Complex diseases are highly polygenic with numerous causal variants, making it challenging to connect these variants to downstream genes and pathways. One hypothesis, the “omnigenic model,” suggests that causal variants influence regulatory networks mediated by "core" genes. However, identifying these mediating connections and evaluating genes’ importance is difficult due to limited power in identifying trans-eQTLs. Alternatively, Perturb-seq combines CRISPR and scRNA-seq to measure perturbation effects on the full transcriptome, enabling causal inference of regulatory networks.

Here, we propose a new framework that performs *M*endelian *R*andomization integrating the *P*erturb-seq, *e*QTL, and *G*WAS summary data (Mr PEG) to identify mediating genes in regulatory networks. Mr PEG models GWAS signals as a linear combination of mediating genes’ expressions, with each modeled as the product of perturbed genes' *cis*-eQTL effects and perturbation effects estimated from Perturb-seq. Mr PEG tests each gene’s mediating effect and provides a trait-specific “mediating-relevance” ranking. In simulations, Mr PEG is unbiased under the null (i.e., no mediation through perturbation), and power increases with GWAS/eQTL sample sizes. Beyond mediating genes, we extend Mr PEG to identify latent expression factors that capture the perturbation effects on multiple genes.

We apply Mr PEG to GWAS data from 136 complex traits (avg N=221k) with *cis*-eQTL effects estimated in eQTLGen (N=30k). We use summary data from three Perturb-seq: in macrophage (600 knock-outs by 16k tested genes), myeloid leukemia and pigment epithelial (both 2k by 8k) cell lines. Mr PEG identifies 22,545 (4,714 unique) mediating genes for 115 traits and 146 (21 unique) latent factors for 93 traits. We observe significant correlations in Mr PEG estimates across different Perturb-seq, suggesting replication of our results.

To validate the genes identified by Mr PEG, we conduct enrichment analyses against established gene-based methods of evolutionary constraint: EDS, pLI, s_het, and RVIS. Mr PEG genes were significantly enriched for constrained genes for all measures ($P<0.05$), suggesting Mr PEG prioritizes constrained genes that are intolerant to mutations. Focusing on 104 genes causing inborn errors of immunity, we find significantly higher Mr PEG signals ($P=0.02$) compared to other Mr PEG significant genes, suggesting Mr PEG can identify potential disease targets.

Overall, Mr PEG shows a new direction to explore regulatory networks with the integration of experimental and population-scale data and sheds light on the genetic architecture of complex traits.

Title: C-STEM: A fast and powerful method for robust context-specific trans-eQTL mapping in multi context studies.

Authors:

L. Krockenberger[1], M. Thompson[2], N. Zaitlen[1], X. Liu[3], B. Balliu[1]; [1]UCLA, Los Angeles, CA, [2]Ctr. for Genomic Regulation (CRG), Barcelona, Spain, [3]Univ. of Chicago, Chicago, IL

Abstract:

The discovery that most disease-associated genetic variants lie outside exons has fueled extensive research into genetic mechanisms governing transcriptional regulation. While identifying associations between gene expression levels and proximal SNPs (cis-eQTLs) has become more feasible, identification of high-quality distal associations (trans-eQTLs) has been challenging. Trans effects are typically weaker and more context-specific than cis effects, making them harder to detect. In addition, existing methods for trans-eQTL mapping often incur a heavy multiple testing burden and overlook the inherent intra-individual correlation of gene expression found in multi context studies with repeated sampling, e.g., GTEx and single-cell RNA-Seq studies. These oversights can significantly diminish the power to detect context-specific trans-eQTLs. To address these issues, we develop C-STEM, a fast and powerful method for context-specific trans-eQTL mapping. C-STEM first accounts for intra-individual correlation by decomposing the expression of a gene into context shared and context specific components, builds cross-validated cis-genetic genetic predictors (CVGP) for each component, and creates a final predictor using both components. C-STEM then tests for association between all CVGP-trans gene pairs, significantly reducing the number of tests and improving power to detect trans-eQTLs that act through cis effects on a gene. Finally, C-STEM employs a hierarchical testing procedure to control FDR across and within contexts and boost power when a significant CVGP-trans gene pair association exists in multiple contexts. Through simulations, we demonstrate C-STEM's increased power in detecting trans-gene regulation compared to other methods. We apply C-STEM to bulk multi-tissue RNA-seq data from the GTEx consortium (N=948) and peripheral-blood single-cell RNA-Seq data from the CLUES (N=234) and OneK1K (N=982) cohorts, generating a comprehensive tissue and peripheral blood cell type-specific trans-eQTL map. C-STEM identifies 89% of trans-eQTLs mapped by existing approaches, while providing a 65% increase in the number of trans-gene regulations identified. Existing approaches overestimate specificity of trans-eQTL effects across contexts; 12% of trans-eQTLs appear unique to a single context using existing methods, compared to only 6% using C-STEM. In summary, C-STEM enables construction of context-specific trans-eQTL maps, aiding in understanding context-specific gene regulatory networks underlying complex human traits.

Session 059: New mechanistic insights via diverse knockout models

Location: Conv Ctr/Room 145A/Level 1

Session Time: Friday, November 3, 2023, 10:45 am - 12:15 pm

Title: CRISPR-engineered disruption of *POGZ* in neuronal lines alters synaptic pathways through indirect epigenetic effects.

Authors:

Y. Liu[1], M. Moyses-Oliveira[1,2,3], S. Erdin[1,4], D. Gao[1,4,2], R. Bhavsar[1], K. O'Keefe[1], K. Mohajeri[1,4,2,3,5], P. Boone[1,4,2,6], G. Xavier[1,2,3], C. Liao[1,4,2,3], A. Li[7,8], R. Yadav[1,4,2,3], M. Salani[1], B. Currall[1], C. De Esch[1], D. Tai[1,4,2,3], J. Gusella[1,4,5,8,9], D. Ruderfer[10,11], K. Brennand[12,7,13], M. Talkowski[1,4,2,3]; [1]Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [2]Dept. of Neurology, Massachusetts Gen. Hosp. and Harvard Med. Sch., Boston, MA, [3]Stanley Ctr. for Psychiatric Res., Broad Inst. of MIT and Harvard, Cambridge, MA, [4]Program in Med. and Population Genetics, Broad Inst. of MIT and Harvard, Cambridge, MA, [5]Program in Biological and BioMed. Sci., Harvard Med. Sch., Boston, MA, [6]Div. of Genetics and Genomics, Boston Children's Hosp., Boston, MA, [7]Nash Family Dept. of NeuroSci., Icahn Sch. of Med. at Mount Sinai, New York, NY, [8]Dept. of Genetics, Blavatnik Inst., Harvard Med. Sch., Boston, MA, [9]Harvard Stem Cell Inst., Cambridge, MA, [10]Div. of Genetic Med., Dept. of Med., Vanderbilt Genetics Inst., Vanderbilt Univ. Med. Ctr., Nashville, TN, [11]Dept. of BioMed. Informatics and Dept. of Psychiatry and Behavioral Sci., Vanderbilt Univ. Med. Ctr., Nashville, TN, [12]Dept. of Genetics and Genomic Sci., Icahn Sch. of Med. at Mount Sinai, New York, NY, [13]Dept. of Psychiatry, Yale Univ., New Haven, CT

Abstract:

POGZ (Pogo transposable element-derived protein with ZNF domain) is strongly associated with autism spectrum disorder (ASD) and related neurodevelopmental disorders (NDDs), yet little is known about its direct and indirect regulatory mechanisms. On two independent human induced pluripotent stem cell (hiPSC) backgrounds, we created an extensive allelic series of 36 CRISPR-engineered clones harboring *POGZ* loss of function (LoF) variants and 30 unaltered clones exposed to identical editing conditions. All 66 hiPSC models were differentiated into neural stem cells (NSCs) and Ngn2-induced glutamatergic neurons (iNs), followed by RNA-sequencing and ATAC-sequencing. *POGZ* disruption resulted in both reduced and increased expression of its regulatory targets, with the direction of effect dependent on cell-type and predicted POGZ interactors. Differential gene expression analyses indicated that heterozygous and homozygous *POGZ* LoF disturbed genes associated with synaptic function in iNs, and extracellular matrix organization in NSCs and iNs. In heterozygous models, ATAC-seq revealed unchanged chromatin accessibility for POGZ regulatory targets, but footprinting analysis revealed disruption of other transcription factors that occurred more often at promoters of differentially expressed genes associated with synaptic function. Neurite arborization and synaptic activity were increased in iNs with *POGZ* homozygous LoF mutations but unaltered in heterozygous models. In iNs derived from the same hiPSC background, we observed significant convergence on disruption of cytokine-related pathways by *POGZ, MEF2C*, and *SCN2A* heterozygous LoF mutations. Overall, the signatures observed in these isogenic allelic series provide insights into cell-type-specific POGZ regulatory mechanisms and shared molecular consequences between different ASD/NDD-associated genes, suggesting key points of convergence on neurodevelopmental pathologies.

Title: H3K27 demethylase inhibition reverses osteoblast transcriptomic signature and excessive osteogenesis in a mouse model of Weaver syndrome.

Authors:

C. Gao[1], W. Lin[1], R. C. Riddle[2], P. Kushwaha[3], L. Boukas[4], C. Lopez[1], H. T. Bjornsson[5], K. D. Hansen[6], J. A. Fahrner[1]; [1]Johns Hopkins Univ. Sch. of Med., Baltimore, MD, [2]Univ. of Maryland Sch. of Med., Baltimore, MD, [3]Brigham and Women's Hosp., Boston, MA, [4]Children's Natl. Hosp., Washington, DC, [5]Univ. of Iceland, Reykjavik, Iceland, [6]Johns Hopkins Univ. Sch. of Publ. Hlth., Baltimore, MD

Abstract:

Weaver syndrome (WS) is a Mendelian disorder of the epigenetic machinery caused by heterozygous germline pathogenic *EZH2* variants. EZH2 is the primary H3K27 methyltransferase of the Polycomb Repressive Complex 2 (PRC2), which also includes EED and SUZ12. Pathogenic variants in the genes encoding for all three subunits result in shared features of overgrowth and intellectual disability. We generated a mouse model for the most common missense variant in WS, *EZH2* p.R684C. *Ezh2*$^{R684C/+}$ mice exhibit skeletal overgrowth, with greater femoral cross-sectional tissue area, and increased mineral apposition rate (MAR) at the endosteal and periosteal surfaces *in vivo*. Because MAR is a measure of osteoblast activity, we differentiated primary bone marrow mesenchymal stem cells (BM-MSCs) towards osteoblasts and showed that *Ezh2*$^{R684C/+}$ cells indeed had enhanced osteogenic potential by Alizarin red staining.

The shared features of PRC2 disorders suggest that their etiology lies collectively in epigenetic perturbation, specifically of H3K27 methylation. Thus, a therapeutic strategy may be to counterbalance the loss of PRC2 activity by inhibiting the opposing H3K27 demethylases Kdm6a/Kdm6b. We piloted GSK-J4, which reduced osteogenesis in *Ezh2*$^{R684C/+}$ cells at the phenotypic level. Here, we present novel RNA-seq data to characterize the effects of GSK-J4 at the transcriptomic level.

We treated *Ezh2*$^{R684C/+}$ and *Ezh2*$^{+/+}$ BM-MSCs with either GSK-J4 or DMSO (vehicle) and assessed gene expression after osteoblast differentiation. We looked for genes whose expression was altered in *Ezh2*$^{R684C/+}$ cells, and which were also responsive to GSK-J4. We identified 2,659 differentially expressed genes (FDR < 0.1) between vehicle-treated *Ezh2*$^{R684C/+}$ and *Ezh2*$^{+/+}$ cells, which constituted the baseline effect of the R684C allele. To isolate genes that were responsive to GSK-J4, we compared the transcriptomes of *Ezh2*$^{R684C/+}$ cells treated with GSK-J4 versus vehicle. In total, 1,075 genes met both criteria. Strikingly, 1,045 of these (97.2%) reversed fold-change directionality upon GSK-J4 treatment, including genes involved in osteoblast differentiation and the bone morphogenetic protein (BMP) pathway.

These data show that by targeting the epigenetic etiology of WS, we achieved a transcriptomic response and phenotypic reduction of osteogenesis that substantially reversed the effects of the R684C variant. We are now performing ChIP to assess H3K27me3 occupancy at loci of interest in *Ezh2*$^{R684C/+}$ and *Ezh2*$^{+/+}$ osteoblasts, under GSK-J4 and vehicle conditions. Our findings provide both mechanistic insight and a potential therapeutic strategy for the PRC2 overgrowth disorders.

Title: Mutation-specific pathophysiological mechanisms of *AFF3* differently influence the DNA repair pathway

Authors:

S. Bassani[1], N. Voisin[2], J. Chrast[1], A. Brusco[3], F. Sirchia[4], L. Turban[5], S. Schubert[5], J. Schlump[6], A. Rami[5], D. DeMille[7], P. Bayrak-toydemir[8], G. Nelson[7], W. Kristen Nicole[9], L. Duncan[10], C. Gilissen[11], L. Vissers[12], R. Pfundt[13], R. Kersseboom[14], H. Yttervik[15], G. Hansen[15], C. Jonsrud[15], F. Smeland[16], M. Lyons[17], C. Carvalho Fonseca[18], C. Zhang[19], J. Lupski[20], L. Flores-Gallegos[21], R. Morales-Toquero[21], L. Potocki[22], F. Petit[23], B. Yalcin[24], N. Guex[25], G. Ambrosini[26], C. Iseli[25], M. Scala[27], M. Iacomino[27], F. Zara[28], K. Writzl[29], A. Maver[30], A. Reymond[1]; [1]Univ. of Lausanne, Lausanne, Switzerland, [2]Univ. of Lausanne, Lausanne, Vd, Switzerland, [3]Univ. of Torino, Torino, Italy, [4]Univ. of Pavia, Cuneo, Italy, [5]Univ. of Leipzig Med. Ctr., Leipzig, Germany, [6]Gemeinschaftskrankenhaus Herdecke Gerhard-Kienle-Weg, Herdeck, Germany, [7]ARUP Lab., Salt Lake City, Salt Lake City, UT, [8]ARUP Lab., Salt Lake City, UT, [9]Univ. of Utah Sch. of Med., Salt Lake City, UT, [10]Vanderbilt Univ. Med. Ctr., Nashville, Nashville, TN, [11]Radboud Univ. Med. Ctr., Nijmegen, Netherlands, [12]Radboudumc, Nijmegen, Netherlands, [13]Radboud Univ. Nijmegen Med. Ctr., Nijmegen, Gelderland, Netherlands, [14]Ctr. for genetic developmental disorders southwest, Zuidwester, Middelharnis, Netherlands, [15]Dept. of Med. Genetics, Univ. Hosp. of North Norway, Tromsø, Norway, [16]Dept. of Paediatric Rehabilitation, Univ. Hosp. of North Norway, Tromsø, Norway, [17]Greenwood Genetic Ctr., Torrance, CA, [18]14Pacific Northwest Res. Inst. (PNRI), Broadway, Seattle, WA, [19]Dept. of Molecular and Human Genetics, BCM, Houston, TX, [20]Baylor Coll. of Med., Houston, TX, [21]Hosp. Ángeles Puebla, Puebla, Mexico, [22]Baylor Col Med/TX Child Hosp, Houston, TX, [23]CHRU Lille, Lille, Nord, France, [24]Univ. of Strasbourg, Illkirch, France, [25]Bioinformatics Competence Ctr., Univ. of Lausanne, Lausanne, Switzerland, [26]Bioinformatics Competence Ctr., Lausanne, Switzerland, [27]Univ. of Genoa, Genoa, Italy, [28]Inst G Gaslini, Genova, Italy, [29]Clinical Inst. of Med. Genetics UMCL, Ljubljana, Slovenia, [30]Univ. Med. Ctr. Ljubljana, Ljubljana, Slovenia

Abstract:

We previously described the KINSSHIP syndrome, an autosomal dominant disorder associated with *de novo* missense variants in the degron of AFF3, a crucial sequence involved in its binding to ubiquitin ligase. Affected individuals shared a recognizable pattern of anomalies that included intellectual disability (ID), epileptic encephalopathy, mesomelic dysplasia and horseshoe kidney. Mouse knock-ins and overexpression in zebrafish provided evidence for a dominant-negative (DN) mode of action, wherein an increased level of AFF3 resulted in pathological effects. In line with this hypothesis, we recently identified a patient with a phenotype resembling KINSSHIP syndrome who carried a partial duplication of *AFF3*. Further screening of ID cohorts revealed thirteen individuals with heterozygous loss-of-function (LoF) truncation variants and six probands with biallelic missense mutations in *AFF3*, who displayed a milder syndrome that partially overlaps KINSSHIP. Mouse knockouts displayed matching enlarged lateral ventricles, reduced corpus callosum, neurological and skeletal anomalies such as vertebrae fusion, and abnormal skull shape. Zebrafish knockdowns similarly exhibited broad neurological defects that could be rescued by expressing human *AFF3* mRNA Wt confirming their association with the ablation of *aff3*. Conversely, the missense isoforms identified in patients did not rescue, confirming the deleteriousness of these variants. The different patient groups indicate that *AFF3* mutations can lead to pathogenic effects through recessive, DN and haploinsufficiency mechanisms. To challenge this hypothesis, we profile the transcriptome of engineered isogenic cell models harboring homozygous A258T/A258T, the most common KINSSHIP mutation, and homozygous LoF/LoF *AFF3* variants. One-third of the differentially expressed genes (DEGs) are common to both datasets, indicating that *AFF3* LoF and DN mutations largely modulate transcriptomes differently. The G2M checkpoint, E2F targets and MYC-related pathways are repressed, while TNFA signalling via NFKB, interferon and inflammatory responses are upregulated in both genotypes. Key pathways involved in apical junction and DNA repair displayed opposite modulation, being downregulated in the KINSSHIP and upregulated in the LoF model. Heterozygous LoF/+ cells showed an intermediate effect, sharing only a fraction of the DEGs with LoF/LoF model. Similarly, compound heterozygous LoF/DN cells displayed unique perturbations in DEGs specific to either homozygous LoF/LoF or DN/DN cells, supporting an *AFF3* mutation-specific pathophysiological mechanism.

Title: A Drug Screen Identifies Potential Therapeutics to Improve Secretory Protein Trafficking Defects Seen in Neural Progenitor Cells from Microcephaly, Epilepsy, and permanent neonatal Diabetes Syndrome Patients

Authors:

L. Ahn[1], J. Farr[1], A. Peden[2], A. Schaffer[3]; [1]Case Western Reserve Univ. Sch. of Med., Cleveland, OH, [2]Univ. of Sheffield, Sheffield, United Kingdom, [3]Case Western Reserve Univ., Moreland Hills, OH

Abstract:

Microcephaly, epilepsy, and permanent neonatal diabetes syndrome (MEDS) is a severe autosomal recessive disorder caused by deleterious bi-allelic variants in the *immediate early response-3 interacting protein-1* (IER3IP1) gene. The role of IER3IP1 in the pathogenesis of MEDS remains elusive. In this study we aim to delineate the impact of pathogenic IER3IP1 variant in neuronal system and to identify potential therapeutics to rescue the defects. Prior work in yeast and human cells suggest IER3IP1 is involved in the anterograde transport of cargos from the endoplasmic reticulum (ER) to the Golgi via COPII vesicles. Based on these functional studies, we predict that pathogenic mutations in IER3IP1 impairs secretory protein trafficking during neurogenesis. To test this, we generated isogenic neural progenitor cells (NPCs) from MEDS patient-derived induced pluripotent stem cell (iPSC) lines for phenotypic analysis. MEDS NPCs had reduced expression of the ER structure marker, Calnexin, and swollen ER morphology with abundant multilamellar bodies compared to isogenic controls. The structural abnormalities in MEDS NPC organelles suggested there may be functional defects in protein secretion. Thus, we sought to directly quantify and compare protein secretion capacity in MEDS and control NPCs. We generated a protein trafficking reporter that enabled us to measure the rate of trafficking and secretion in real-time and an inducible manner. We found a defect in protein trafficking and secretion that we confirmed by secretome analysis. To test if the trafficking defect caused by IER3IP1 mutation could be reversed pharmacologically, we performed a drug screen with 3000 FDA approved bioactive small molecules library. We discovered a selective CFTR modulator improved secretion of the trafficking of the reporter protein in MEDS NPCs by >50% compared to control, without a major impact on viability. In summary, we demonstrate that IER3IP1 patient variants likely leads to MEDS by disrupting ER structure and secretory protein trafficking and that this can be potentially rescued by small molecule treatment.

Title: A genome-wide CRISPR screen identifies *TNRC18* as a novel regulator of inflammatory signaling

Authors:

F. Rahimov, S. Ghosh, S. Petiwala, M. Schmidt, E. Nyamugenda, J. Tam, S. Singh, V. Avram, A. Modi, C. A. Espinoza, C. Lu, J. Wang, A. Keller, M. Macoritto, N. A. Mahi, N. Chung, M. J. Flister, K. V. Katlinski, A. Biswas, A. I. Den Hollander, J. F. Waring, J. D. Stender; AbbVie, Inc., North Chicago, IL

Abstract:

Precise regulatory control of the innate and adaptive immune responses is essential for proper tissue homeostasis and prevention of inflammation-driven diseases. Cytokines play a critical role in the maintenance of this homeostasis. To identify genes modulating intracellular cytokine levels in response to activation of innate immune signaling pathways, we performed a genome-wide screen in human monocytic cells (U937) treated with lipopolysaccharide (LPS), a potent activator of the immune system, by sorting cells with the lowest and highest levels of intracellular IL1β following CRISPR-mediated gene perturbations. This screen identified hundreds of statistically significant hits, including members of the TLR4 and JAK/STAT signaling pathways, as essential for production of IL1β. Components of the IL-10 receptor signaling and the Cullin ring-finger ligase complex were identified as candidates that when knocked out lead to enhanced expression of intracellular IL1β. In addition, we identified TNRC18 as a novel modifier of LPS-dependent gene regulatory programs with loss of TNRC18 leading to reduced LPS-dependent induction of pro-inflammatory gene expression and secretion of cytokines. The *TNRC18* gene harbors an intronic single nucleotide variant (rs748670681) enriched in the Finnish population which is genetically linked to several immune-mediated diseases, including inflammatory bowel disease, ankylosing spondylitis, and psoriasis. The mRNA for *TNRC18* and an adjacent gene *WIPI2* are decreased in U937 cells engineered to be homozygous for the risk allele of rs748670681, resulting in reduction in LPS-dependent gene activation and cytokine production, but elevation of interferon responsive gene programs. These findings define global regulation strategies for IL1β production in myeloid cells and identify novel modifiers of LPS-dependent transcription.

Title: Immune dysregulation and *Lama1* upregulation in LAMA2-deficient Congenital Muscular Dystrophy: Insights from a mouse model and implications for gene therapy.

Authors:

Y. Tenorio De Menezes, J. Cheng Zhang, M. Johnson, D. Kemaladewi; Univ. of Pittsburgh, Pittsburgh, PA

Abstract:

Congenital muscular dystrophy (CMD) is a group of genetically inherited neuromuscular disorders characterized by muscle weakness and degeneration. LAMA2-deficient CMD is a subtype caused by mutations in the laminin alpha-2 (LAMA2) gene, for which there is no treatment available. Upregulation of disease compensatory gene *Lama1*, achieved via CRISPR activation and delivered using AAV9, is a promising therapeutic approach for LAMA2-CMD. However, progress towards the clinical application of such strategy is hampered by the lack of comprehensive baseline immunophenotypes in LAMA2-CMD. Such knowledge is important because precise interventions on inflammatory pathways are crucial to minimize the risks and augment the success of gene therapy administrations, irrespective of the disease context. Therefore, we set out to fill such knowledge gap by investigating the lymphocyte and myeloid cell dynamics in skeletal muscles and blood from dyW, a mouse model of LAMA2-CMD**.** We found increased leukocyte infiltration in muscles from the dyW mice as early as 2-week-old, with $CD4^+$ T cells, NK cells, macrophages, and neutrophils being the main infiltrating populations. This suggests the involvement of these immune cell types in the early stage of muscle pathology. We also observed increased neutrophils, monocytes, and inflammatory cytokines in the dyW blood, indicating their role in inducing muscle damage and inflammation. Subsequently, we evaluate how these immune cell profiles change in the mice treated with AAV9 carrying CRISPRa to upregulate *Lama1*. We observed reduced expression of activation markers PD1, KLRG1, and GITR on $CD4^+$ T cells in the treated mice, indicating that inflammation is reduced upon *Lama1* upregulation. Furthermore, it shows that the expression of *Lama1* is not detrimental in the LAMA2-CMD mice. Collectively, our study highlights the role of immune cells and inflammatory mediators in early pathology of LAMA2-CMD, which can be reduced following CRISPRa-mediated *Lama1* upregulation. These findings are important to increase our understanding of the immunological changes associated with LAMA2-CMD and pave the path towards development of safe and effective therapy for this condition.

Session 060: Precision prescription: The therapeutic potential of ASOs and small molecules for genetic disorders

Location: Conv Ctr/Room 202A/Level 2

Session Time: Friday, November 3, 2023, 10:45 am - 12:15 pm

Title: Development of a novel lipid nanoparticle-based adjustable gene therapy platform technology.

Authors:

A. Goraltchouk[1], J. Lourie[2], F. Luppino[1], A. Seregin[1], Z. Kai[2]; [1]Remedium Bio, Inc., Boston, MA, [2]Univ. of Massachusetts, Boston, MA

Abstract:

Despite significant promise, gene therapy remains limited by immunogenicity, cost, and inability to adjust the dose following initial administration. These limitations have thus far restricted gene therapy to the treatment of rare, monogenic diseases. To overcome these challenges, we undertook the development of a novel dose adjustable gene therapy system, which enables safe and low-cost delivery of a broad range of therapeutic genes, while allowing for simple and predictable post-treatment dose adjustment. A novel Lipid Nanoparticle (LNP) DNA delivery system was characterized *in vitro* by evaluating transfection efficiency and cytocompatibility with primary human adipocytes. Expression of furin-cleavable human insulin and Exenatide, a Glucagon-Like Peptide-1 Receptor Agonist, was confirmed by sandwich ELISA. Efficacy, dose-adjustability, and biocompatibility of the delivery system was further confirmed *in vivo* by a series of subcutaneous injections of the LNP formulation in two mouse strains, a diet-induced diabetic and non-diabetic C57BL/6N and a non-diabetic B6N-*Tyr*$^{c\text{-}Brd}$/BrdCrCrl albino. Durability, redosability, and cryolipolysis-induced down-titratability of transgene expression was confirmed using bioluminescence imaging over 5 months. We observed high and dose-dependent transfection efficiency, reaching 75% (±17% SD) GFP+ in preadipocytes and 81% (±8% SD) GFP+ in adipocytes over 72h in culture without dose limiting cytotoxicity. Expression levels of Exenatide and functional, furin-cleavable insulin reached 1.66 fg/adipocyte and 8.4E-9 IU/adipocyte respectively. Transgene expression remained durable in vivo for >5 months, with no decrease in bioluminescence flux over the study period. Dose-dependent up-titratability was confirmed stable at multiple anatomical sites for >4 months, with no decrease in reporter signal over the study period. Finally, effective stepwise down-titration was achieved by reducing reporter transgene expression in 25.7% (±9.7% SD) increments per treatment step, following three 10-minute cryolipolysis treatments with no observable side-effects to the animals. In summary, we report the development of a novel, dose adjustable gene therapy platform technology. The system is capable of safely delivering therapeutic transgenes, without retreatment-limiting immunogenicity, and produces durable gene expression that can be adjusted based on clinical need. This platform, for the first time, enables the delivery of subcutaneously administered protein and peptide treatments as a safe, effective, and low-cost adjustable gene therapy.

Title: Engineering poison exons for allele-specific silencing of the Huntington's disease gene *HTT*.

Authors:

C. Catamura, V. Salinas-Rios, F. Urnov, L. Lareau; Univ. of California, Berkeley, Berkeley, CA

Abstract:

Repeat expansion diseases that follow a dominant negative inheritance pattern, such as Huntington's disease, present a great challenge for gene therapies. The most appealing therapeutic strategy requires allele-specific silencing of the disease allele while leaving the wildtype allele intact. Because some amount of repeat sequence is present in both alleles, selective CRISPR targeting of the mutant allele is profoundly challenging.

Here, we present a general framework for allele-specific suppression of repeat expansion genes and show its efficacy in suppressing the gene responsible for Huntington's disease, *HTT*. Our solution is to engineer a poison exon into only the disease allele: an alternative exon that, when included in an mRNA, will introduce an early stop codon and trigger nonsense-mediated mRNA decay (NMD). mRNA expressed from the edited disease allele would be degraded, while mRNA from the wildtype allele would be unchanged.

To target only the disease allele, we take advantage of benign, heterozygous SNPs within introns of the disease gene. CRISPR guides specific to the SNP allele that is in phase with the pathogenic repeat could recruit an RNA-guided deaminase base editor to introduce nucleotide substitutions within a narrow window on only the disease allele.

Using a deep learning model of splicing based on thousands of bases of sequence context, we screen *in silico* for positions where targeting a base editor could create sequences that form new splice sites. Importantly, the novel exon need not be near the actual disease mutation; any highly heterozygous SNP in an intron could potentially serve this purpose.

We applied this model to screen SNPs within introns of *HTT* and found positions where editing adjacent to the SNP was predicted to give rise to new poison exons. We validated these targets at the endogenous *HTT* locus in human cell culture, showing that the edits created new exons spliced into *HTT* mRNAs with over 75% inclusion. The resulting mRNAs were degraded by NMD, substantially reducing *HTT* expression.

We conclude that editing of just a few carefully chosen nucleotides in a gene can be sufficient to create a new exon, making it a feasible approach for allele-specific mRNA depletion. Our approach combines the power of CRISPR editing with deep learning models of RNA sequence to engineer poison exons with therapeutic potential for a disease that, at present, has no viable treatments.

Title: Preclinical development and *in vivo* delivery of antisense oligonucleotides for targeted *NF1* exon 17 skipping.

Authors:

D. Wallis[1], M. Moore[2], H. Liu[1], X. Zhang[1], G. Long[3], E. Westin[4], R. Kesterson[4], Z. Jiangbing[5], L. Popplewell[6]; [1]UAB, Birmingham, AL, [2]Teesside, Middlesbrough,, United Kingdom, [3]Yale, New haven, CT, [4]Pennington BioMed. Res. Ctr., Baton Rouge, LA, [5]Yale, New Haven, CT, [6]Teesside, Middlesbrough, United Kingdom

Abstract:

Neurofibromatosis type I arises from germline variants in the *NF1* gene, which diminish expression of the tumor suppressor protein neurofibromin. Our research has published *in vitro* evidence highlighting the therapeutic potential of antisense oligonucleotides (ASOs) for numerous NF1 pathogenic variants through targeted exon skipping, including exon 17. To provide pre-clinical validation of ASOs *in vivo* and facilitate translational development. We created a mouse model (hG629R) with the insertion of a human exon 17 carrying the G629R pathogenic variant with partial flanking intronic sequences. Two routes of administration (IV and ICV) and three delivery platforms including i) Naked PMOs, ii) Adeno Associated viral (AAV) vectors with AAV9 serotype carrying U7-SnRNA expression cassettes and iii) conjugation of ASOs with morpholino chemistry to cell penetrating peptides (CPP-PMO) were explored to give proof-of-concept of *in vivo* exon skipping efficacy of our optimized ASOs. *Biodistribution studies:* Viral delivery of AAV9-eGFP was administrated by both ICV and IV to adult and neonatal mice to evaluate biodistribution. eGFP was detected in the liver, heart and brain, with ICV found to give greater expression in the brain relative to IV. Provisional evidence from sub-region analysis indicates eGFP detection in the cortex, cerebellum, and olfactory bulb in adult mice. *Exon Skipping efficacy:* Delivery of naked PMOs via IV results in significant levels of exon skipping by qRT-PCR analysis and restoration of NF1 expression by Western blot primarily in the kidney. AAV9-U7-SnRNA encoding ASO sequences and CPP-PMOs were delivered to adult mice and harvested after one week. Viral delivery of U7-SnRNA constructs showed detectable exon skipping by RT-qPCR analysis and thus a reduction of the G629R variant at the transcript level in the optic nerve, liver and provisional evidence indicates the brain. Both routes of administration of the CPP-PMO produced exon skipping in brain, optic nerve, sciatic nerve and kidney, with exon skipping also evident in the liver following ICV delivery. We provide proof-of-concept evidence that efficient delivery of ASOs *in vivo* resulting in detectable targeted exon skipping is achievable across multiple tissues. Optimization of the dosing regimen for both delivery strategies will be performed with examination of the efficacy and durability of the exon skipping response being used as readouts. Finally, efficacy in an acute or tumor model of NF1 loss would provide strong proof of concept for using exon skipping as a therapeutic.

Title: Potential treatment for CMT2S caused by *IGHMBP2* cryptic splice variant, with ASO based therapeutic

Authors:

S. Smieszek, C. Tyner, B. Przychodzen, A. Kaden, C. Johnson, C. M. Polymeropoulos, G. Birznieks, M. H. Polymeropoulos; Vanda Pharmaceuticals Inc., Washington, DC

Abstract:

Charcot-Marie-Tooth disease Type 2S (CMT2S) is a rare autosomal recessive Charcot-Marie-Tooth disease subtype. Rare variants in immunoglobulin mu-binding protein 2 (IGHMBP2) have been shown to cause CMT2S by resulting in abnormal RNA processing leading to alpha-motor neuron degeneration. A patient was reported with consequential variants within *IGHMBP2*. Whole genome sequencing (WGS) revealed a paternally inherited cryptic splice site variant (non-coding variant (c.1235+894 C>A) deep in intron 8). The resulting transcript undergoes nonsense-mediated decay (NMD), resulting in haploinsufficiency. Our objective was to target this cryptic splice site, rescuing IGHMBP2 protein levels with a novel antisense oligonucleotide (ASO).

We obtained the patient's fibroblast cell line and confirmed the variant with WGS and the existence of NMD. We designed a 19mer ASO targeting deep in intron 8 (c.1235+894 C>A), around sequence CACTTCCAC(A)GGGGGAAGA. Several ASOs were designed with a phosphorothioate methoxyethyl (MOE) backbone and prioritized based on *in silico* binding affinity. Fibroblast cells underwent ASO treatment (1µM) and 48-hour incubation. Flow cytometry and fluorescein labelled ASO (GFP+99.8%) confirmed cellular entry. For additional exploratory analyses on the patient's phenotype, electrophysiology studies on iPSC cells and motor neurons derived from the patient's fibroblast cells were completed.

Upon treatment with ASO, we observed a significant IGHMBP2 protein level increase (~30%) in oligo-treated samples compared to control (untreated) samples (WB antibody Sigma SAB2106426). qPCR results confirmed increased ratio of restored wild-type transcript to cryptic exon-containing transcript (~1.3-fold). Preclinical data support this ASO as potential treatment restoring IGHMBP2 protein levels, with limited off-target effects *in silico*. Electrophysiological studies revealed hyperexcitability and spontaneous firing of motor neurons, resembling an amyotrophic lateral sclerosis (ALS) phenotype.

An increased number of autosomal recessive CMT2S cases caused by *IGHMBP2* consequential variants are being reported. The N-of-1 precision medicine approach may prove instrumental to the design of treatments for this highly diverse genetic disorder. Patient-specific phenotypic analysis of motor neurons further confirms the genetic diversity of this disorder, revealing a phenotypic resemblance to ALS. This case exemplifies the shifting boundary between rapid WGS based clinical diagnoses and research capabilities allowing for the design of personalized ASO based treatments.

Title: Utilizing NMD Inhibition, Readthrough Agents, and Small Molecule Modulator Combination to Achieve Functional and RNA-Level Restoration of CFTR Nonsense Mutations.

Authors:

N. Sharma[1], A. Eastman[1], M. Havens[1], A. Bowling[1], S. Patel[1], E. Kavanagh[1], G. Lin[1], S. Jalloh[1], L. Huang[2], C. Merlo[1], G. Cutting[1]; [1]Johns Hopkins Univ., Baltimore, MD, [2]Ionis Pharmaceuticals, Carlsbad, CA

Abstract:

Trikafta, a CFTR modulator treatment that specifically corrects the dysfunctional protein, has significantly increased the quality of life and median life expectancy for those suffering from Cystic Fibrosis (CF). Unfortunately, those harboring nonsense variants cannot reap the benefits of Trikafta, as the mRNA decay caused by the nonsense variant leads to the production of a minimal, or no protein at all. Thus, it is vital to identify additional combinations that can counteract the consequences of nonsense variants in order to make treatment options accessible to all people with CF. We utililzed a) Human nasal epithelial (HNE) cells, and b) Flp-In stable cells. In total, 12 people with CF, 20 carrier parents, one unrelated carrier, and 8 healthy controls were recruited. Each person with CF had at least one copy of nonsense variant. Cells were treated with combinations of the following - ASOs to NMD-critical transcripts, small molecule inhibitors of NMD, readthrough compounds, and CFTR modulator. CFTR mRNA expression, protein processing and function were evaluated.
CFBE cells bearing W1282X CFTR EMG treated with a readthrough compound or Trikafta alone resulted in a minimal recovery of CFTR function. However, upon treatment with combination of NMD-ASO, readthrough, and Trikafta, we observed a remarkable recovery of CFTR function, e.g., treatment with SMG6-ASO, ELOX-02 and Trikafta, ΔIsc= 32.5±13.2 μA/cm2, corresponding to ~20% of wildtype CFTR function in this system. In NEs harboring W1282X, the same combination yielded a significant increase in CFTR function (2.4±0.1μA/cm2), corresponding to ~18% of control. ASO in combination with readthrough compound and CFTR corrector were able to make full-length glycosylated CFTR protein. Importantly, qRT-PCR results for CFBEs and HNEs support that CFTR RNA was increased 10-fold under these treatment conditions and bulk-RNA sequencing confirmed that the ASOs drastically reduced NMD-causing target genes. ASOs among NMD inhibitors and ELOX-02 among readthrough compounds were least toxic and resulted in higher recovery of CFTR expression and function. Among these variants, W1282X was most responsive and R1162X was least. Altogether, our findings suggest that therapeutics for nonsense variants will benefit from taking into account RNA stability, readthrough, and toxicity. It is clear that recovery of function from some of these variants, including W1282X, benefits greatly from the addition of Trikafta following successful readthrough of stabilized RNA. In conclusion, our methodology can be similarly applied to investigate potential treatments for other genetic disorders that are caused by nonsense mutations.

Title: TYRA-300 Demonstrates Significant Increases in Growth and Bone Length in a Mouse Model of FGFR3-Related Skeletal Dysplasia.

Authors:

J. H. Starrett[1], M. Guillo[2], N. Kaci[2], R. V. Swanson[1], L. Legeai-Mallet[2]; [1]Tyra BioSci.s, Carlsbad, CA, [2]Université de Paris Cité, Imagine Inst., Lab. of Molecular and Physiopathological Bases of Osteochondrodysplasia. INSERM UMR1163, Paris, France

Abstract:

Achondroplasia (ACH) is the most common human skeletal dysplasia affecting ~1 in 25,000 births. Infants with ACH have an increased risk for death related to critical foramen magnum stenosis leading to cervicomedullary compression. A specific mutation in *FGFR3*, G380R, causes over 99% of achondroplasia. FGFR3 is expressed in growth plate chondrocytes where it functions to slow endochondral bone formation. The G380R mutation, as well as other activating mutations, results in increased FGFR3 activity, which suppresses chondrogenesis in the growth plate, disturbing long bone elongation. Vosoritide, a C-naturetic peptide analogue, acting exclusively on the MAP kinase pathway, was recently approved as a daily injection to increase annual growth velocity in children with open growth plates. While an important breakthrough, long-term effects on ACH-associated comorbidities are not yet known. TYRA-300 is an oral, highly selective FGFR3 inhibitor currently undergoing a Phase 1/2 clinical trial, SURF301 (Study in Untreated and Resistant FGFR3+ Advanced Solid Tumors). To explore the effectiveness of TYRA-300 in FGFR3-related skeletal dysplasias, TYRA-300 was evaluated in an FGFR3$^{Y367C/+}$ transgenic mouse model. TYRA-300 administered daily at a 1.2 mg/kg dose for 15 days in the FGFR3$^{Y367C/+}$ mouse model significantly increased body length in mice by 17.6% compared to the vehicle ($p<0.0001$) and increased the length of the femur (+24.4%), tibia (+38.3%) and L4-L6 (+23.9%) in mice ($p<0.0001$). Growth in the skull resulting in elongation (+11.0%) and improvement in the size and shape of the foramen magnum were also observed. Histological staining of the femurs revealed restoration of the hypertrophic zone and secondary ossification center within the epiphysis after TYRA-300 treatment. Collagen X staining also indicated that TYRA-300 improved the structure of the trabecular bone and increased chondrocyte differentiation. These data indicate that inhibiting FGFR3 directly leads to highly increased bone length, as well as foramen magnum diameter, in this preclinical model.

Session 061: Scaling screening: A plethora of perspectives

Location: Conv Ctr/Room 147A/Level 1

Session Time: Friday, November 3, 2023, 10:45 am - 12:15 pm

Title: Comparing uptake of pre-test genetic services and genetic testing via chatbot versus genetic counseling standard of care: findings from the BRIDGE trial

Authors:

K. Kaphingst[1], M. Goodman[2], R. Chambers[2], A. Gammon[1], W. Kohlmann[1], R. Monahan[2], M. Volkmar[1], M. Sigireddi[2], O. Ginsburg[3], G. Del Fiol[1], S. Buys[1], BRIDGE research team; [1]Univ. of Utah, Salt Lake City, UT, [2]New York Univ., New York, NY, [3]Natl. Cancer Inst., Gaithersburg, MD

Abstract:

Introduction. Increasing numbers of patients are being identified who could benefit from genetic evaluation for inherited cancer susceptibility. The development of innovative and sustainable approaches to delivering genetic services is therefore essential. Automated approaches to genetic services delivery utilizing conversational agents (i.e., chatbots) are being developed, but data from clinical trials are needed to compare outcomes between this approach and standard of care. The Broadening the Reach, Impact, and Delivery of Genetic Services (BRIDGE) randomized controlled trial was designed to address this need. **Methods.** In the BRIDGE trial, between 2020-2023 we contacted 3156 primary care patients in the University of Utah Health and New York University Langone Health healthcare systems who were 25-60 years of age and did not have a personal history of cancer. These patients were identified through an automated algorithm as eligible for genetic evaluation based on cancer family history information in their electronic health record and contacted via automated patient portal messages. Patients were randomized 1:1 to an intervention arm in which pre-test genetic services were provided via education through a scripted, rules-based chatbot accessed through a hyperlink embedded in the patient portal message or a control arm in which pre-test genetic services were delivered through a standard of care pre-test genetic counseling appointment. The primary trial outcomes were uptake of pre-test genetic services (either pre-test chat or pre-test appointment, depending on arm) and uptake of genetic testing. **Results.** For the primary trial outcome of uptake of pre-test genetic services, of the 1586 eligible patients randomized to the chatbot intervention arm, 398 (25.1%) completed the pre-test genetics education chat. Of the 1570 eligible patients randomized to the standard of care control arm, 362 (23.1%) completed the pre-test genetic counseling appointment. For the primary trial outcome of uptake of genetic testing, of the patients in the chatbot intervention arm, 236 (14.9%) opted to order genetic testing, while of the control arm patients, 274 (17.5%) opted to order genetic testing. **Discussion.** Our findings show that, among unaffected primary care patients eligible for cancer genetic evaluation, similar proportions in the chatbot intervention and standard of care control arms were reached by pre-test genetic services and opted to have genetic testing. The findings indicate that automated approaches such as chatbots are acceptable and feasible alternative approaches to delivering pre-test genetic services within healthcare systems.

Title: Expanded newborn screening using first-tier genome sequencing for highly penetrant early onset conditions to increase health equity for children

Authors:

W. Chung[1], D. M. Kay[2], C. Koval[1], L. Amendola[3], S. F. Suchy[4], A. Begtrup[4], R. J. Sicko[2], T. Brandt[4], B. Friedman[4], S. Hoffer[4], K. Langley[4], K. Retterer[4], R. Torene[4], A. Coffey[3], K. Monaghan[4], J. Devaney[4], A. Johnson[4], S. Strom[3], K. McWalter[4], Z. Hu[1], R. Hernan[1], J. Ganjun[1], H. Kavus[1], J. Wynn[1], S. Ratner[1], Y. Quevedo[1], N. Pimentel Soler[1], A. Ziegler[1], R. J. Taft[3], K. S. Hruska[4], P. Kruszka[4], M. Caggana[2]; [1]Columbia Univ., New York, NY, [2]Wadsworth Ctr., New York State Dept. of Hlth., Albany, NY, [3]Illumina Inc, San Diego, CA, [4]GeneDx, Gaithersburg, MD

Abstract:

For more than 20 years there has been speculation about a future in which newborns are routinely screened at birth for early-onset disorders using whole genome sequencing, but this vision has yet to be systematically assessed in a diverse population-based study. Genomic Uniform Screen Against Rare Disease in All Newborns (GUARDIAN) is a consented pilot study investigating the use of genome sequencing performed on dried blood spots (DBS) collected as part of routine newborn screening in New York. All enrolled participants receive screening results for 147 genes which are associated with early-onset and treatable conditions. Parents of research participants can also opt in to receive results for 90 additional genes (total 237) associated with seizures and/or early-onset neurodevelopmental disorders. GUARDIAN launched in September 2022, and at the time of abstract submission had enrolled 1,500 research participants. Here we will report on cohort demographics, laboratory findings and clinical outcomes. Additionally, we will discuss genome test development, gene and disease selection, variant reporting criteria and the communication of results to providers and parents.The results to date indicate that genome sequencing can complement traditional NBS, and may reduce healthcare disparities through expanded diagnostic equity and associated reductions in time to diagnosis and delays in care. The GUARDIAN study and other ongoing large-scale genome based NBS studies will provide critical information on the benefits and risks of genomics-based screening and will aid in formulating guidelines for its potential use as a component of routine NBS.

Title: Exploring the implementation of preventive genomic screening in the primary care setting: Scaling suggestions from providers and the implementation team.

Authors:

J. Shea[1,2], C. L. Blout-Zawatsky[1,2,3], K. Bellegarde[2,4], L. Onelio[2,5], I. Miller[2,1], L. Subramanian[2,1], N. R. Encina[2,4], S. Williams[6,7], R. C. Green[1,2,8,3], R. Kennedy[9,10,7], R. Homayouni[7]; [1]Brigham & Women's Hosp., Boston, MA, [2]Ariadne Labs, Boston, MA, [3]Broad Inst., Cambridge, MA, [4]Harvard Sch. of Publ. Hlth., Boston, MA, [5]Harvard Sch. of Publ. Hlth., Boson, MA, [6]Corewell Hlth.East, Troy, MI, [7]Oakland Univ. William Beaumont Sch. of Med., Rochester, MI, [8]Harvard Med. Sch., Boston, MA, [9]Beaumont Res. Inst., Royal Oak, MI, [10]Corewell Hlth.East, Royal Oak, MI

Abstract:

Purpose: Preventive genomic screening (PGS) offers patients the opportunity to identify genetic risk factors allowing for earlier detection, management and/or prevention. Health systems are starting to explore the implementation of clinical PGS in the primary care setting. Corewell Health East (CHE) partnered with the Ariadne Labs Precision Population Health (PPH) program to implement PGS in primary care.
Methods: CHE piloted the Invitae Genetic Health Screen in 6 primary care clinics. Patients who saw a participating primary care provider (PCP) over the prior year were sent a Mychart invitation or were invited directly by their PCP. Genetic counselors (GCs) shared the results with both patients and their PCPs for follow-up and management. We conducted semi-structured qualitative interviews with participating PCPs (n=5) and GCs (n=3) involved in the project to obtain feedback on implementation of PGS, and compiled insights from conversations with project leaders. Themes were identified through review of interview notes and recordings and project team meeting minutes.
Results: From June 2022 to May 2023, over 700 patients completed PGS. Among these, 18% were identified to carry a monogenic disease risk. Major themes about implementing PGS included: PCPs' desire to be more involved in invitations and results disclosure, considerations around equitable access and when in the lifespan is it best to offer PGS, suggestions for PCP education and training through peer learnings and robust partnership between PCPs and GCs, better utilization of IT systems for PGS workflows, and reducing administrative workflow burden for providers/staff when considering scale. All PCPs and GCs indicated that they found PGS valuable, and PCPs were unanimous in recommending that other colleagues offer PGS.
Conclusion: Generating PCP buy-in and identifying feasible clinical workflows is essential in order for health systems to successfully offer PGS at scale. This implementation project highlights that providers are interested in PGS and find it valuable for their patients' care. Designing intuitive workflows, building supportive IT systems, and clarifying PCP and GC roles will be important for scaling the integration of PGS in primary care.

Title: HOP POP: Evaluating the Effectiveness of Population Screening for HBOC and Lynch Syndrome in 25,000 Members of the Healthy Oregon Project

Authors:

P. Spellman[1], S. Jackilen[2], S. Richards[2], A. Shafer[3], O. Timothy[2], A. Potter[2], C. Driscoll[2], G. Goh[2], J. Letaw[2], V. Serrato[2], S. McCabe[2], J. Thanner[2], A. Kulkarni[2], S. Medica[2], J. Buitrago[2], A. Larson[2], K. Johnson-Camacho[2], K. Brown[2], R. Crist[2], P. Farris[2], S. Evans-Dutson[2], R. Lutz[2], A. Mitchell[4], P. Anur[5], L. Marriott[2], K. Hamman[2], A. Mulford[2], W. Wiszniewski[2], J. Sampson[2], A. Adey[2], B. O'Roak[2], C. Harrington[2]; [1]UCLA, Los Angeles, CA, [2]OHSU, Portland, OR, [3]Univ. of Oregon, Eugene, OR, [4]Invitae, San Francisco, CA, [5]Tempus, 15931 Nw Rossetta St, OR

Abstract:

Deploying population screening for genetic diseases to human populations depends on numerous technical and health service challenges. These interventions rely not just on the technical ability to sequence human genomes at scale and identify deleterious variants but also on the effectiveness of the health care system to deliver cost effective interventions. Two critical limitations in the current health care system are the number and availability of genetic counselors and the understanding of genetic reports by those tested. Additionally, in order to be cost effective the price of testing must be low and the utilization of information must remain high. As part of NCT04494945, Identifying and Caring for Individuals With Inherited Cancer Syndromes, we sought to evaluate the effectiveness of general population Lynch Syndrome and HBOC screening. Two critical changes were implemented to standard genetic testing. First, the trial only reports pathogenic or likely pathogenic variants and only provides genetic counseling to these individuals. Overall cost is dramatically lower as genetic counseling both limits utilization and increases costs. Second, costs were engineered from the system so that the test is as inexpensive as possible, with the consequence that the sensitivity is lower than can be achieved in conventional testing. The outcome measures of NCT04494945 are the utilization of interventions that improve health outcomes for those with pathogenic variants.In total, we have returned clinical results to more than 23,000 members of the Healthy Oregon Project (HOP). HOP's recruitment model is built on social media marketing, allowing both broad and targeted recruitment. 137 HBOC and 54 Lynch Syndrome patients were identified. In addition, all members of HOP who wished to participate in genetic health screening were screened for an additional 26 genes. 51 individuals were identified to have familial hypercholesterolemia. In total 835 reportable results were identified, just under 4% of those screened. Within 30 months, we screened and returned clinical results to about 1% of the adult population of Oregon.

Title: Public opinions and attitudes toward non-invasive prenatal testing on reddit: Content and sentiment analysis.

Authors:

B. Xiao[1], J. Yan[1], R. Hayeems[1,2]; [1]The Hosp. for Sick Children, Toronto, ON, Canada, [2]Univ. of Toronto, Toronto, ON, Canada

Abstract:

Background Non-invasive prenatal testing (NIPT) can be used to detect fetal chromosomal abnormalities early in pregnancy. As eligibility criteria broaden and screening targets expand, gauging public acceptability of NIPT becomes increasingly important. In addition to primary data collection using traditional survey and interview techniques, social media platforms such as Twitter and Reddit have grown in importance as a data source for research, clinical, education, and policy stakeholders to track public opinion. Accordingly, using natural language processing (NLP) techniques (i.e., sentiment analysis) as a novel technique, the purpose of this study was to investigate public opinions and attitudes toward NIPT on Reddit. Specifically, we explored topics discussed, perceived benefits and concerns, and emotional responses to NIPT.
Method Textual data were collected from four Reddit communities focusing on the NIPT content posted from September 2012 to September 2022 (367 posts and 7822 comments in total) using the R program. We first conducted a content analysis to explore topics discussed, perceived benefits and concerns related to NIPT. Next, we employed various lexical-based sentiment analysis techniques, including Bing, NRC, Syuzhet, and AFINN, to delve deeper into public acceptability of NIPT. Specifically, the word in the comment that was related to an emotion was assigned a value (0 = the word is not associated with this emotion, 1 = the word is associated with this emotion) based on the different lexicons. The sentiment scores were calculated by subtracting positive emotion from negative emotion.
Results Content analysis findings indicated that social media users consider NIPT to be beneficial and desirable. Concerns related to unwanted anxiety/stress, and the fact that NIPT results would not change anything about their approach to pregnancy were also expressed. The sentiment analysis identified more positive than negative emotions; the mean sentiment scores ranged from 0.48 and 1.22, depending on the specific Lexicon used. Specific emotions (i.e., joy, trust, anticipation, fear) and emotional trade-offs were also identified that have research, clinical and policy implications.
Conclusion Our novel approach to understanding public perception and attitudes toward NIPT yielded results that are consistent with conventional patient-oriented research methods. These findings may not only contribute to ongoing improvements in prenatal patient care, research, and policy but also indicate that sentiment analysis applied to social media data can serve as a suitable means to assess public acceptability of NIPT.

Title: Patient views and understanding of non-invasive prenatal screening in a safety net setting

Authors:

K. Riggan[1], N. Condon[2], M. A. Allyse[3]; [1]Mayo Clinic, Rochester, MN, [2]Univ. of Florida Hlth., Jacksonville, FL, [3]Mayo Clinic, Jacksonville, FL

Abstract:

Thirty-one state Medicaid programs provide coverage of non-invasive prenatal screening (NIPS) as a first-tier screen for fetal aneuploidy, regardless of risk status. The patient experience of receiving NIPS in safety net settings has yet to be fully explored, including any pre- and post- test counseling received from a provider, perceptions of the scope of NIPS, and personal utility of screening results. We interviewed 100 Black and Hispanic patients (n=52 non-Hispanic Black, n=45 Hispanic, n=3 Other) receiving prenatal care at a federally qualified health center in Jacksonville, FL after the state of Florida initiated coverage of NIPS for Medicaid patients. Pre-test counseling was described as limited. Participants generally understood that the test was for certain fetal aneuploidies or understood the test as more broadly pertaining to fetal health. Most remembered Trisomy 21 specifically mentioned by their provider. Some participants believed NIPS detects multifactorial conditions, such as autism, or described a therapeutic misconception about its clinical purpose. In general, participants perceived NIPS as a routinized component of prenatal care, with some believing it was mandatory or required. Participants frequently volunteered that they did not intend to use the results for a pregnancy continuation decision, but felt that early identification of a potential aneuploidy would be useful to acquire information about the condition and identify financial and social resources for parenting a child with special needs. The return of low-probability results and fetal sex was the most recalled aspect of their pre- and post-test experience. These results were reassuring and of high personal value, although some participants wanted more agency in how results were returned. Our findings suggest that adherence to guideline recommendations on informed consent and decision-making for NIPS is challenging to implement in safety net settings. As NIPS continues to expand across diverse clinical practice, and as additional conditions are added to screening panels, it will be essential to evolve clinical paradigms to provide support to patients when they need it while reducing informational burden on a low health literacy, high-need patient population with limited resources for follow-on care.

Session 062: Single-cell analyses across various tissues

Location: Conv Ctr/Room 207A/Level 2

Session Time: Friday, November 3, 2023, 10:45 am - 12:15 pm

Title: Inferring novel disease related cell states within and across-cell types in scRNA and scATAC seq data using condPCA

Authors:

S. Carver; Harvard Med. Sch., Boston, MA

Abstract:

Advancing our understanding of cell type heterogeneity and disease-related states requires effective approaches for inferring cell states from single-cell sequencing data. While many methods exist for cell type annotation and clustering of single cell data, few methods infer states within or across cell types. We introduce a novel method called condPCA, which “conditions” out known cell types from single cell data followed by Principal Component Analysis (PCA) to identify cell states within and across cell populations.

We evaluate the performance of condPCA in simulations and in multiple real datasets. In simulations, condPCA outperforms state-of-the-art techniques like UMAP and NMF for rare cell state identification. For example, condPCA accurately recovers a state present in 100 cells with the same accuracy that requires 1000 cells in NMF. In scRNA-seq data from light-stimulated neurons, where light is a state spanning cell types, condPCA outperforms NMF in capturing stimulus-driven variability (condPCA: R2=0.53 and NMF: R2=0.43) . UMAP fails to capture biologically enriched states, consistent with recent findings highlighting its lack of reproducibility and data distortion.

Finally, we apply our approach to 191,890 nuclei from an Alzheimer's disease (AD) cohort, comprising snRNA-seq and snATAC-seq. condPCA identifies 76 and 48 RNA and ATAC based states, respectively, demonstrating much higher dimensionality to single-cell data than typically considered (based on the BIC cutoff adapted for PCA). Genes in these states show significant enrichment in Gene Ontology (GO), including enrichments for known AD pathways. condPCA outperforms NMF by identifying cell states that are more enriched for GO terms; notably, 5,080 and 1,540 GO enrichments are identified in the top 15 snRNA-seq states inferred by condPCA and NMF, respectively (FDR p&lt0.05). Integrating the inferred cell states with AD GWAS data demonstrated significant enrichment of GWAS heritability, showcasing how cell state inference can be connected to disease risk. snATAC-based states demonstrate substantially greater enrichment for AD heritability compared to snRNA-based states (ATAC: 465, RNA: 1 enrichment, BH p&lt0.05), highlighting potentially “missing” disease mechanisms that can be identified with single-cell epigenomics . In high-dimensional brain data, condPCA identifies established and novel cell states, elucidating states influencing AD-related neuroinflammatory phenotypes, especially in microglia.

Our work demonstrates that novel, disease relevant cell states can be inferred from high dimensions in single-cell data and are often missed by conventional methods.

Title: Contextualizing postmortem bias for single-nuclei transcriptomic studies of human brain.

Authors:

E. Vornholt, B. Kopell, R. Thompson, L. Liharska, E. Cheng, N. Beckmann, A. Charney; Icahn Sch. of Med. at Mount Sinai, New York, NY

Abstract:

Background: The molecular underpinnings of brain disorders remain unknown. This is in part due to the inability to study brain tissue from living people. Instead, the field has relied on postmortem human brain tissue, which may not be an accurate representation of living human brain tissue. The Living Brain Project (LBP) has allowed researchers to ethically biopsy living cortical tissue to study multi-omic differences between living and postmortem brains. **Methods:** Single-nuclei RNAseq was generated from 31 living (LIV) and 21 postmortem (PM) cortical samples and clustered/annotated into neural cell types. Differentially expressed genes (DEGs) were identified across cell types using a linear mixed model. Gene regulatory networks (regulons) were defined using SCENIC and then tested for differentially active regulons (DARs) between groups. The utility of the LIV vs. PM expression signature was explored via elastic net to create a postmortem linear predictor score (PMlink) for postmortem bulk samples. **Results:** After QC, the 52 samples yielded 362,390 quality cells (61% postmortem and 39% living) which were clustered into 10 cell types. Differential expression revealed massive signals across all cell types, with 64% of genes differentially expressed in at least one cell type (n = 16,759 genes). Neuronal markers are disproportionately upregulated in postmortem neurons and oligodendrocyte markers are upregulated in living across all cell types. In examining DARs, living cells show increased regulon activity associated with RNA processing, whereas postmortem cells display increased activity of regulons associated with neuronal signaling. Finally, the calculated PMlink had perfect predictive strength in determining LV vs PM classification across all pseudobulk testing permutations (ROC AUC = 1). When PMlink is incorporated as the dependent variable in a linear mixed model containing only postmortem bulk samples, the LV vs PM pseudobulk DE signal is replicated (ρ=0.65). All key findings using pseudobulk were replicated in an independent bulk dataset. **Conclusion:** The combined DEG and DAR results reveal ubiquitous dysregulation of key biological systems throughout the postmortem brain that are not reflective of the living expression profile. These findings provide necessary context for interpreting postmortem gene expression as a snapshot of biological processes at death rather than a proxy for living brain function. Machine learning methods provide utility for creating correction algorithms that can help address this problem in past and future postmortem transcriptomic studies of the brain.

Title: Single-cell dissection of bipolar disorder

Authors:

X. Han[1,2], W. Ruzicka[3,4,2], S. Mohammadi[1,2], N. Sun[1,2], L. Hou[1,2], L. He[1,5], D. Lee[6], P. NM[6], J. Bendl[6], A. Hong[6], C. Casey[6], Z. Shao[6], M. Alvia[6], S. Argyriou[6], G. E. Hoffman[6], J. F. Fullard[6], PsychENCODE Consortium, P. Roussos[6], M. Kellis[1,2]; [1]Massachusetts Inst. of Technology, Boston, MA, [2]Broad Inst. of MIT and Harvard; Cambridge, 02142, USA., Boston, MA, [3]Lab. for Epigenomics in Human Psychopathology, McLean Hosp.; Belmont, 02478, USA., Boston, MA, [4]Dept. of Psychiatry, Harvard Med. Sch.; Boston, 02115, USA., Boston, MA, [5]The Faculty of Hlth.Sci., Dept. of Publ. Hlth., Univ. of Southern Denmark, 5230 Odense M, Odense, Denmark, [6]Icahn Sch. of Med. at Mount Sinai, New York, NY

Abstract:

Bipolar disorder (BD) is a prevalent psychiatric condition, affecting 1-2% of the global population and resulting in a substantial societal burden. Despite progress in the field of psychiatry, our understanding of the pathophysiology of BD remains incomplete, with diagnosis predominantly dependent on clinical symptoms and behavioral presentations, and few advancements in therapeutic treatments.

To advance our understanding of the neurobiological mechanisms underlying BD and prioritize putative therapeutic targets of this highly heritable and polygenic disorder, we profile approximately 500,000 single-cell nuclear transcriptomes (snRNA-seq) across two independently assayed cohorts of human prefrontal cortex sampled from 88 donors (44 BD cases and 44 controls). We annotate 25 cell types, identify 1,932 differentially expressed genes (bdDEGs), and integrate these cell type-specific bdDEGs with single-cell expression quantitative trait loci (sc-eQTL, N=427) and BD-associated genome-wide association studies loci using type-specific Mendelian randomization (scMR) and multiple statistical genetic approaches.

Our bdDEGs implicate excitatory and inhibitory neurons, and are enriched in glutamatergic and GABAergic synapses, neurotransmitter signaling, and neuron projection development. Many bdDEGs have also been supported by BD genetic evidence based on sc-eQTLs and statistical fine-mapping. We observe substantial overlap of DEGs and pathways between BD and schizophrenia, underscoring their shared genetic architecture.

Our BD single-cell atlas offers annotation and interpretation of BD genetic findings, highlighting both common and rare BD GWAS loci that are dysregulated bdDEGs in one or multiple brain cell types. The large-scale sc-eQTLs enable statistical causal inference and fine-mapping approaches to prioritize multiple putative druggable targets for BD, with shared genes and pathways identified from bdDEGs.

In conclusion, our BD single-cell atlas provides an important resource for exploring the molecular mechanisms underpinnings of BD, annotating and interpreting BD-associated loci, and discovering potential therapeutic targets.

Title: Single-cell transcriptomic profiling of human pancreatic islets reveals genes responsive to glucose exposure over 24 hours

Authors:

C. Grenko[1], L. L. Bonnycastle[1], H. J. Taylor[1], T. Yan[1], A. J. Swift[2], C. C. Robertson[1], N. Narisu[2], M. R. Erdos[2], F. S. Collins[2], D. Taylor[1]; [1]Ctr. for Precision Hlth.Res., Natl. Human Genome Res. Inst., Natl. Inst.s, Bethesda, MD, [2]Ctr. for Precision Hlth.Res., Natl. Human Genome Res. Inst., NIH, Bethesda, MD

Abstract:

Disruption of pancreatic islet function and glucose homeostasis can lead to the development of sustained hyperglycemia, beta cell glucotoxicity, and ultimately type 2 diabetes (T2D). In this study, we sought to explore the effects of hyperglycemia on human pancreatic islet (HPI) gene expression by exposing HPIs from two donors to low (2.8mM) and high (15.0mM) glucose concentrations over 24 hours, assaying the transcriptome at seven time points using single-cell RNA sequencing (scRNA-seq). We modeled time as both a discrete and continuous variable to determine momentary and longitudinal changes in transcription associated with islet time in culture or glucose exposure. Across all cell types, we identified 1,528 genes associated with time, 1,185 genes associated with glucose exposure, and 845 genes associated with interaction effects between time and glucose. We clustered differentially expressed genes across cell types and found 347 modules of genes with similar expression patterns across time and glucose conditions, including two beta cell modules enriched in genes associated with T2D. Finally, by integrating genomic features from this study and genetic summary statistics for T2D and related traits, we nominate 363 candidate effector genes that may underlie genetic associations for T2D and related traits.

Title: Single-nucleus multi-omics profiling across 49 individuals of Hispanic ancestry reveals cell-specific skeletal muscle insulin-responsive gene regulatory signatures.

Authors:

B. Li[1], A. Varshney[1], A. Tovar[1], N. Manickam[1], P. Orchard[1], R. DeFronzo[2], C. Jenkinson[2], L. Norton[2], S. Parker[1]; [1]Univ. of Michigan, Ann Arbor, MI, [2]UTHSCSA, San Antonio, TX

Abstract:

Type 2 diabetes mellitus (T2D) is a prevalent metabolic disease causing severe morbidity and mortality. Hundreds of T2D risk loci have been found, 90% of which map to non-coding regions. This has obfuscated efforts to establish molecular mechanisms, partly because some of these loci may impact gene expression and chromatin accessibility in specific cellular and stimulation contexts. In T2D, insulin is a principal stimulus, and insulin clamp is the gold standard technique to study its signaling effects. However, the method is resource-intensive and there is a pressing need for greater understanding of cell-type-specific changes following insulin stimulation *in vivo*. To this end, we generated and analyzed single-nucleus RNA and ATAC (snRNA-seq, snATAC-seq) libraries from paired skeletal muscle biopsies in individuals of Hispanic ancestry before and after insulin clamp to help illuminate the context-specific mechanisms of risk variants.
49 subjects received a four-hour insulin clamp with vastus lateralis muscle biopsies obtained prior to and immediately following. We processed frozen biopsy samples using genetic multiplexing and a randomized block design to account for age, sex, BMI, and stimulatory condition. We isolated nuclei and performed snRNA- and snATAC-seq using the 10X Chromium platform, performed rigorous quality control of all nuclei, identified and assigned singlets to individuals' genotypes using Demuxlet, and jointly clustered nuclei with LIGER. We performed differential expression analysis using DESeq2 and tested functional enrichment with gprofiler2. After removing outliers, we investigated 94 paired samples constituting 61,140 pass-QC nuclei from 14 cell types ranging in abundance from type 1 fibers (34.4% of nuclei) to Schwann cells (0.04% of nuclei). In type 1 fibers, we identified 3,004 significantly differentially expressed genes following insulin clamp (FDR <5%), adjusting for batch and individual factors. This large response to insulin stimulation resulted in significant enrichment (FDR <2%) of relevant processes such as response to hormone stimulus, KLF transcription factors, and miR-335-5p, which negatively regulates GLUT4. In comparison, 510 genes were differentially expressed in mesenchymal stem cells (3.3% of nuclei), enriched for extracellular matrix processes and genes such as *TCF7L2*, *IGF1R*, and *BCL2*, previously implicated in T2D GWAS meta-analyses. We are currently performing expression and chromatin accessibility QTL mapping and GWAS colocalization to better understand the cell- and context-specific (insulin-responsive) mechanisms of genetic risk for muscle-associated traits and diseases.

Title: Single cell analysis of pancreatic islet cells in type 2 diabetes reveals pathogenic mechanisms enriched in populations of African ancestry.

Authors:

E. Heuston, A. Doumatey, A. Adeyemo, C. Rotimi; Ctr. for Res. on Genomics and Global Hlth., Natl. Human Genome Res. Inst., NIH, Bethesda, MD

Abstract:

Type 2 diabetes (T2D) disproportionally affects individuals of African ancestry (AA) in the United States and there is increasing evidence that a subset of T2D risk variants is ancestry specific. To improve our understanding of T2D pathophysiology in different populations, we conducted single cell analysis of the pancreas, a key organ in T2D pathophysiology. We hypothesize that there are differences in the regulation of molecular mechanisms between pancreatic islet cells from AA individuals compared to individuals of European ancestry (EA). We analyzed single cell RNA-Seq (scRNA-Seq) and single nucleus ATAC-Seq (snATAC-Seq) data for 12 donors with T2D (5 AA and 7 EA) from the Human Pancreas Atlas Program. Following quality control, scRNA-Seq analysis (Seurat v4) was done for 18033 cells (5115 AA and 12918 EA) to define transcriptional profiles while snATAC-Seq analysis (ArchR v1) was done for 62199 cells (23220 AA and 38979 EA) to associate transcription factor binding site (TFBS) enrichment with transcriptional profiles. Gene score profiles in snATAC-Seq data were used to integrate chromatin profiles with RNA expression. Clustering analysis defined 17 transcriptionally distinct pancreatic cell clusters, of which two exocrine populations were highly enriched for cells from AA donors. RNA_cluster0 (19% of all AA cells, 9% of all EA cells) and RNA_cluster3 (15% of all AA cells, 6% of all EA cells) were defined by significantly higher levels of pancreatic lipase (*PNLIP*), protease serine 1 (*PRSS1*), regenerating family member 1 alpha and beta (*REG1A* and *REG1B*, respectively), and serine protease inhibitor gene Kazal type 1 (*SPINK1*), and were enriched for HNF4A, HNF4G, and TRIM28 TFBS. Gene set enrichment analysis (WebGestalt) linked these genes to biological processes including protein and fat digestion and absorption (FDR ≤ 2E-16 and FDR ≤ 8E-3, respectively), exocrine pancreatic insufficiency (FDR ≤ 4E-4), and chronic pancreatitis (FDR ≤ 2E-16). Mutations in *PNLIP*, *REG1A*, and *REG1B* dysregulate islet regeneration programs and oncogenic transformation. We also identified a third cell population (RNA_cluster4) that was depleted for AA cells (3% of all AA cells, 11% of all EA cells). RNA_cluster4 is one of three beta cell clusters identified in the analysis. These cells were enriched for EZH2 and Pol2 TFBS accessibility compared to other populations, suggesting active chromatin remodeling in this population.Our findings demonstrate that pancreatic cells in T2D exhibit molecular mechanisms that may differ by ancestry. Further studies of other T2D-relevant tissues may provide insight into pathophysiological processes in T2D in other populations.

Session 063: Spatial omics: Resources and applications

Location: Conv Ctr/Ballroom C/Level 3

Session Time: Friday, November 3, 2023, 10:45 am - 12:15 pm

Title: The Spatial Atlas of Human Anatomy (SAHA) Project: A High content spatial map of the Human Body

Authors:

J. Park[1], R. De Gregorio[1], B. Robinson[1], E. Hissong[1], S. E. Church[2], E. Metzger[3], L. Pan[3], Y. Liang[3], J. Reeves[3], J. Beechem[3], A. Alonso[1], S. L. Houlihan[1], R. E. Schwartz[1], C. Mason[1]; [1]Weill Cornell Med., New York, NY, [2]NanoString, Seattle, WA, [3]NanoString Technologies, Seattle, WA

Abstract:

The Spatial Atlas of Human Anatomy (SAHA) is a foundational effort to map 250 million cells and transcriptomes and proteomes of 30 non-diseased organs from healthy adults at two spatial scales: whole transcriptome of histological features (50 µm to 2 mm), and 1,000-plex RNA and 64-plex protein panels at spatial subcellular resolution (50 nm across 1 cm2). The project aims to establish and validate best practices in experimental design, sample processing, data analysis, and data standards for high-content spatial analysis across multiple human organs at whole transcriptome and proteome levels. The profiled samples will capture variability across genders and ancestries. All results, including raw and processed data, will be made available to the scientific community through the SAHA data portal and AtoMxTM Spatial Informatics Portal.
Here we present the updated Phase I data collected from human bone marrow and lymph nodes, in addition to our initial three in-depth organs (liver, colon, prostate). The spatial whole transcriptome analysis generated by the GeoMx® Digital Spatial Profiler (DSP) measures the expression of whole transcriptomes matched to the exact shape of functional histological organ features from H&E and immunofluorescence stainings. On serial sections, the 1,000-plex RNA profiles and 64-plex protein profiles collected by the CosMx™ Spatial Molecular Imager (SMI) enables the highest-ever subcellular resolution maps of cell types, lineage states, metabolic capacity, cellular neighborhoods, subcellular movements of organelles, and spatially resolved (and novel) ligand-receptor interactions. So far, we captured ~1.2 million cells collected across ~1,200 fields of view from GeoMx DSP, CosMx SMI RNA and protein slides. Through comparing these data to several colon cancer samples surgically resected, we show how spatial organ atlasing at multiple scales can uncover unique insights into organ development, health, and cancer. We also show how the SAHA data can serve as a benchmark reference for spatial precision medicine.

Title: The Parkinson's Cell Atlas: A spatial map of disease programs in human brains.

Authors:

J. Yuan[1,2,3], N. Haywood[4,3], Z. Liao[1,2,3], I. Tuncali[1,2,3], X. Adiconis[4,3], S. K. Simmons[4,3], Y. Kuras[1,2,3], D. El-Kodsi[1,2,3], Z. Lin[1,2,3], J. Parker[1,2,3], G. E. Serrano[5,3], T. G. Beach[5,3], S-C. Zhang[6,3,7], M. B. Feany[1,3], J. Z. Levin[4,3], C. R. Scherzer[1,2,3], X. Dong[1,2,3], ASAP Parkinson Cell Atlas Consortium; [1]Brigham and Women's Hosp., Boston, MA, [2]Harvard Med. Sch., Boston, MA, [3]Aligning Sci. Across Parkinson's (ASAP) Collaborative Res. Network, Chevy Chase, MD, [4]Broad Inst. of MIT and Harvard, Cambridge, MA, [5]Banner Sun Hlth.Res. Inst., Sun City, AZ, [6]Univ. of Wisconsin-Madison, Madison, WI, [7]Duke-NUS Med. Sch., Singapore, Singapore

Abstract:

Parkinson's Disease (PD) is highly complex and heterogeneous, with many clinical presentations and proposed molecular mechanisms. Leveraging recent advances in single-cell and spatial transcriptomics technology, the Parkinson's Cell Atlas project aims to elucidate the spatial distribution and progression of PD biomarkers in the human brain.

Here we present a preliminary spatial atlas of PD pathology in the middle temporal gyrus (MTG). We perform single cell and spatial transcriptomics assays across 100 brains: thirty-three from healthy controls, thirty-four from patients with pre-symptomatic, prodromal disease, and thirty-three from patients with clinically manifest PD. We assayed 328,349 spatial transcriptomes and 8.2 TB of data - building the largest spatial map of the human temporal cortex to date. Using only unsupervised methods, we accurately annotate six cortical layers and the white matter across all samples. We integrate spatial and single-cell transcriptomics data from the same samples to perform cell type-spatial deconvolution, finding that each cortical layer comprises unique cell populations, with especially high spatial specificity for subtypes of glutamatergic neurons. Layer-specific differential expression and pathway analysis reveal associations with tissue-specific abundance of hallmark Lewy bodies as well as ante-mortem motor and cognitive phenotypes. We find MAPK-signaling and cancer-related pathways are upregulated during disease progression even during disease onset. By contrast, electron transport chain-linked pathways are downregulated during disease progression. Significant PD-related differential expression localizes predominantly to layers 5 and 6, but we also observe activity in other cortical layers, such as a decline in ribosomal gene expression in layer 1 and the white matter.

This atlas may provide a roadmap for developing therapeutics that precisely target molecular disease processes in the right cells, in the right space, and at the right time during the course of the disease.

Title: Statistical analysis of cell type-specific spatial expression patterns for spatial transcriptomics.

Authors:

P. Wu, L. Shang, X. Zhou; Univ. of Michigan, Ann Arbor, MI

Abstract:

Spatially resolved transcriptomics, enabled by a diverse range of technologies, facilitates in-depth exploration of transcriptomic landscapes, extending from individual cellular domains to broader tissue contexts. However, the interpretation of abundant gene expression data derived from techniques quantifying averaged expression per spot is frequently complicated by the heterogeneity in cellular compositions, leading to significant computational and statistical challenges. The spatial heterogeneity of gene expression within specific cell types, influenced by functionality, microenvironments, and intercellular communication, further adds to this complexity. Evident in distinct brain regions, these spatial variations in gene expression play critical roles in cellular differentiation, tissue organization, disease progression, and in identifying potential new therapeutic targets, underscoring the importance of better analytical methods to interpret these spatially resolved transcriptomics data. To tackle these limitations, we introduce SpaTially variable cELL type-specific gene identificAtion (Stella), a statistical method developed to identify genes exhibiting cell type-specific spatial expression patterns. By employing a spatially varying coefficient model, Stella examines one gene at a time and accurately models each gene's spatial expression pattern, in relation to the distribution of cell types across tissue locations. Not only does Stella maintain calibrated type I error control, but it also shows a significant increase in detection power across a spectrum of technical platforms. Applications to four spatial transcriptomics data, including a mouse cerebellum Slide-seq data, Stella identified 5 Purkinje-specific spatial genes and 12 granular-specific spatial genes, thereby disclosing spatial heterogeneity and diverse functional lobules among these cell types in the mouse cerebellum. Thus, Stella offers a significant advance in the reliable interpretation of spatial transcriptomic data, contributing an innovative dimension to our understanding of cellular heterogeneity.

Title: Leveraging spatial transcriptomics data to recover cell locations in single-cell RNA-seq with CeLEry

Authors:

S. Jiang[1], Q. Zhang[2], A. Schroeder[1], J. Hu[3], R. Xiao[1], M. Li[1]; [1]Dept. of Biostatistics, Epidemiology and Informatics, Univ. of Pennsylvania Perelman Sch. of Med., Philadelphia, PA, [2]Dept. of Epidemiology, Biostatistics and Occupational Hlth., Sch. of Population and Global Hlth., McGill Univ., Montreal, QC, Canada, [3]Dept. of Human Genetics, Sch. of Med., Emory Univ., Atlanta, GA

Abstract:

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity in health and disease. However, the lack of physical relationships among dissociated cells has limited its applications. Spatial transcriptomics (ST) technologies can overcome the limitations of scRNA-seq by profiling transcriptome-wide gene expression while retaining the location information of cells within tissues. It is desirable to recover their cell location information by leveraging gene expression-spatial location relationships learned from ST. The location recovered scRNA-seq data can be utilized in downstream analysis, e.g., cell-cell communications, in which knowledge of cell locations is important. Although existing methods such as Tangram, SpaOTsc and novoSpaRc can be utilized to infer the location information for cells in scRNA-seq, cell location recovery is not their primary objective, their performance is suboptimal. To address the limitations, we developed CeLEry, a supervised deep learning algorithm that leverages gene expression and spatial location relationships learned from spatial transcriptomics to recover the spatial origins of cells in scRNA-seq. CeLEry has an optional data augmentation procedure via a variational autoencoder, which improves the method's robustness and allows it to overcome noise in scRNA-seq data. CeLEry can infer the spatial origins of cells in scRNA-seq at multiple levels, including 2D location and spatial domain of a cell, while also providing uncertainty estimates for the recovered locations. To evaluate the performance of CeLEry, we conducted comprehensive benchmarking evaluations on multiple datasets generated from brain and cancer tissues using Visium, MERSCOPE, MERFISH, and Xenium. Our results show that CeLEry can reliably recover the spatial location information for cells in scRNA-seq, maintaining a consistent level of prediction accuracy across all locations. It is able to retain the pairwise spatial distance between every two cells, reconstruct spatial gene expression patterns and detect potential cell-cell communications. By contrast, Tangram, SpaOTsc and novoSpaRc fail to recover cell locations and gene expression patterns in some regions, and have lower accuracy regarding cell pairwise distances. Lastly, we demonstrate that CeLEry is computationally fast and memory efficient, making it a promising tool for spatial transcriptomics studies.

Title: Spatial transcriptomics analysis reveals pathology-specific alveolar niches in pulmonary fibrosis

Authors:

A. Vannan[1], R. Lyu[2], A. L. Williams[1], E. D. Mee[1], N. M. Negretti[3], M-I. Chung[1], S. Hirsh[3], J. Hirsh[3], J. M. Sucre[3], J. A. Kropski[3], D. J. McCarthy[2], N. E. Banovich[1]; [1]Translational Genomics Res. Inst., Phoenix, AZ, [2]St. Vincent's Inst. of Med. Res., Fitzroy, Australia, [3]Vanderbilt Univ. Med. Ctr., Nashville, TN

Abstract:

Pulmonary fibrosis (PF) is a chronic, progressive condition that characterizes the end stage of many interstitial lung diseases (ILDs). Studies using single-cell RNA sequencing (scRNA-seq) have allowed researchers to interrogate the complex cellular heterogeneity of the lung, lending insight into cell-specific molecular mechanisms of PF. However, scRNA-seq is limited by a lack of spatial context and fails to capture the regional heterogeneity that is characteristic of progressive PF. The advent and expansion of commercially available platforms to profile gene expression in situ (collectively termed spatial transcriptomics) are enabling researchers to probe the complex patterns of gene expression and cell composition within local microenvironments and understand how spatial niches vary with disease progression. Using the 10X Xenium platform, an imaging based spatial transcriptomics platform providing subcellular resolution, we characterized the expression of 343 genes in over 1 million nuclei across 28 samples from 13 donors, including 7 unaffected and 21 ILD samples. The ILD samples included donor-matched pairs of less and more fibrotic regions of tissue, allowing us to investigate the dynamics of disease progression. We identified differences in cell type composition and gene expression between unaffected and disease samples. Importantly, these data provide a uniquely accurate representation of cell composition in ILDs to date, with substantial recovery of cell types, such as fibroblasts and endothelial cells, that are typically depleted during scRNA-seq sample preparation. Using both cell-based and cell-agnostic approaches, we observed distinct niches in the lung including several alveolar niches enriched in ILD samples. By combining these computationally-defined niches with clinician-annotated histopathology stains, we identified novel patterns of dysregulation in alveoli informed by spatial context. This includes epithelium-depleted alveolar structures, macrophage accumulation within air spaces, and detachment of aberrant KRT5-/KRT17+ cells from the local basement membrane. To further explore these unique features of PF, we computationally segmented individual air spaces across the spectrum of disease severity to outline processes of dysregulated epithelial plasticity in alveoli. Together, these results advance our understanding of the molecular programs regulating PF in situ.

Title: Dissecting the relationship between autism spectrum disorder risk and cognitive ability: Insights from spatial profiles of gene expression and copy number variants

Authors:

K. Kumar[1], W. Engchuan[2], S. Kazem[3], G. Huguet[3], T. Renne[1], B. Thiruvahindrapuram[2], J. R. MacDonald[2], C. Poulain[1], M. Jean-Louis[3], Z. Saci[3], M. Klein[4], O. Shanta[5], L. Almasy[6], G. Dumas[1], D. Glahn[7], S. W. Scherer[8], J. Sebat[5], S. Jacquemont[3]; [1]Univ. of Montreal, Montreal, QC, Canada, [2]The Hosp. for Sick Children, Toronto, ON, Canada, [3]CHU Sainte-Justine, Montreal, QC, Canada, [4]Radboud Univ Med. Ctr., Nijmegen, Netherlands, [5]UC San Diego, San Diego, CA, [6]Children`s Hosp. of Philadelphia, Philadelphia, PA, [7]Boston Childrens Hosp. & Harvard, Boston, MA, [8]The Hosp. for Sick Children & Univ. of Toronto, Toronto, ON, Canada

Abstract:

To date, no specific genes have been found that solely contribute to the risk of autism spectrum disorder (ASD) when disrupted by Copy Number Variations (CNVs), without affecting cognitive abilities or increasing the risk of Intellectual Disability (ID). While a few studies have attempted to identify genes that are more specific to ASD, it has been challenging to distinguish between ASD risk and its impact on cognitive abilities. In this study, we propose a novel approach that utilizes spatial gene expression profiles to analyze the relationship between ASD risk and cognitive abilities. We utilize rare genomic deletions and duplications (CNVs) that fully encompass one or more genes, as well as, post-mortem derived spatial profiles of gene expression in the cortex, to characterize the impact of genetic variation on ASD risk and cognitive ability.

Towards this, we analyzed i) measures of cognitive ability and CNV calls in over 259,000 individuals from unselected populations, and ii) data from the Psychiatric Genomics Consortium (PGC)-CNV working group, which includes information on 13,000 individuals with ASD and 23,340 unaffected family members from various studies. We i) assigned 16,000 coding genes to 180 cortical brain regions of the Glasser parcellation based on their preferential expression across the human cortex of healthy adults provided by the Allen Human Brain Atlas (AHBA); and ii) performed 180 burden association analyses (linear/logistic regression) to calculate the average effect size on cognitive ability and the risk of ASD of genes altered by CNVs for each of the cortical regions, for deletions and duplications separately.

Preliminary results from our study examining CNVs and cognitive ability indicate a negative correlation between the effect sizes of deletions and duplications across the human cortex. The most pronounced effects are observed in sensorimotor regions for deletions and temporal pole regions for duplications. In contrast, the impact of deletions and duplications on ASD risk exhibits a positive correlation between the effect sizes of deletions and duplications across the human cortex. The ASD risk shows a distinct cortical profile that does not align with the effect sizes observed for cognitive ability (near zero correlations). These findings suggest that spatial profiles of gene expression can provide a novel functional axis for distinguishing the effects of CNVs on cognitive ability from their impact on the risk of ASD.

Session 064: Walking the dogma: Proteomics to inform genomic studies

Location: Conv Ctr/Room 146B/Level 1

Session Time: Friday, November 3, 2023, 10:45 am - 12:15 pm

Title: Targeted mass spectrometry-based proteogenomics approach for large-scale detection of protein isoforms colocalized with disease

Authors:

J. Korchak, E. Jeffrey, M. Murali, N. Perry, M. Civelek, G. Sheynkman; Univ. of Virginia, Charlottesville, VA

Abstract:

Complex diseases including coronary artery disease (CAD) continue to jeopardize the health of individuals worldwide. While there are clear genetic links to changes in alternative splicing of transcripts in CAD, little is understood about the protein isoforms that underpin the genetic variants. Recent advances in technology, including long read RNA sequencing and powerful mass spectrometry approaches, allow detection and characterization of the effectors of disease-associated alternative splicing. Using these advanced technologies, we have developed a proteogenomic approach to enable the characterization and detection of novel protein isoforms.

First, splice quantitative trait loci (sQTLs) colocalized with CAD risk loci were obtained from our previous study (Aherrahrou et al., Circ Res 2023;132). Next, long read RNA sequencing data from primary smooth muscle cells from six heart transplant donors were collected and analyzed. We systematically identified promising protein isoform candidates by linking the splice junction (predicted by Leafcutter) to putative protein isoforms translated from the matched long read RNA sequencing data. Using these protein isoform candidates, we performed in silico tryptic digests to identify protein isoform specific peptides for targeting and development of synthetic trigger peptides. Finally, using targeting mass spectrometry approaches that uses real-time peptide search and dynamic sampling (Tomahto, GoDig), we discovered evidence for isoform-specific peptides that are correlated with CAD.

Using long read RNA sequencing data, we uncovered 11 out of 157 CAD-colocalized sGenes in which the sQTL-associated junction was novel to long read identified isoforms. We found that 105 Leafcutter clusters mapped to both the reference annotation and the long read data, providing orthogonal support for their expression. In characterizing the predicted effects of the sQTLs on the isoforms, we found that 13% contained an alternative donor or acceptor, 26% contained a change in the 5' or 3' untranslated region, 45% corresponded to alternative exons, 7% corresponded to noncoding RNA products, and 8% mapped to NMD candidates. These results provided 109 candidates for protein targeting via mass spectrometry. The advanced targeted approach demonstrated four times more coverage compared to traditional non-multiplexed targeting approaches, providing higher throughput detection for sQTL corresponding proteins.

In conclusion, we present a novel approach that can serve as a template for other complex diseases with GWAS and sQTL datasets, providing the field with protein-level evidence of proposed sQTL drivers of disease.

Title: Integrative analysis of UK Biobank proteomics data with autoimmune disease GWAS reveals complex causal relationships

Authors:

A. Hukku[1], E. Kvikstad[2], S. Vasquez Grinnell[2], J. D. Szustakowski[2], J. Maranville[1], E. Holzinger[1]; [1]Bristol Myers Squibb, Cambridge, MA, [2]Bristol Myers Squibb, Pennington, NJ

Abstract:

Introduction: The UK Biobank Pharma Proteomics Project (UKB-PPP) Consortium has generated data on ~3,000 plasma proteins measurements using the Olink Explore 3072 panel for over 53,000 UK Biobank participants. Using this incredibly large-scale data resource, the consortium has conducted full pQTL mapping, identifying 10,248 genetic associations with 85% being novel discoveries[1]. Many of these proteins are circulating cytokines known to play a crucial and complex role in immune-related disease. For this study, we performed in-depth investigation of complex causal relationships between a set of cytokines and autoimmune disease risks with the overall goal of informing and optimizing drug discovery.
Methods: To do this, we applied Mendelian Randomization (MR), a well-established method that leverages features of germline genetics for inferring a causal relationship between an exposure (e.g. protein level) and an outcome (e.g. disease risk). First, we compared commonly used MR approaches (GSMR[2] and MR-Egger[3]) and a more recently developed approach (MR-Clust[4]) to assess agreement across methods. We also investigated the prevalence of potentially divergent causal relationships between protein levels and disease risk (*i.e.* increased protein levels are associated with both increased and decreased risk of a particular disease).
Results: First, we found inconsistency in results between GSMR and MR-Egger. By design, only *MR-Clust* identified distinct causal components linking plasma protein levels to disease risk, which we hypothesize could explain some of the observed disagreements between MR-Egger and GSMR. Next, using *MR-Clust*, we identified several protein-disease pairs that had causal components with conflicting directionalities. For example, we found that increased IL-10 levels are associated with both increased and decreased risk of inflammatory bowel disease (IBD), in-line with the pro- and anti-inflammatory functions of the cytokine.
Conclusions: This ground-breaking proteomics resource has shed new light on the complexity of the relationships between plasma protein levels and human disease traits. The distinct causal components we identified may highlight different causal mechanisms between proteins and diseases. Insights from these findings could lead to novel therapeutic hypotheses and, potentially, patient subsets defined by distinct disease mechanisms for targeted therapies.
Acknowledgements:UKBB-PPP Consortium, Paradigm4 Team, and Pharmalex Team
Refs:1. PMID: 37794186. 2. PMID: 293354003. 3. PMID: 285270484. 4. PMID: 32915962

Title: Biological and genetic factors associated with protein variation in five human tissues.

Authors:

H. Tang[1], H. Fang[1,2], L. Jiang[1], F. V. Leprevost[3], R. Jian[1], J. Chan[1], D. A. Glinos[4], T. Lappalainen[5], A. I. Nesvizhskii[3], A. P. Reiner[6], GTEx Consortium, M. P. Snyder[1]; [1]Stanford Univ., Stanford, CA, [2]Capital Normal Univ., Beijing, China, [3]Univ. of Michigan, Ann Arbor, MI, [4]New York Genome Ctr., New York, NY, [5]SciLifeLab & NY Genome Ctr., New York, NY, [6]Fred Hutchinson Cancer Ctr., Seattle, WA

Abstract:

Proteins serve as the fundamental workforce driving cellular activities. Perturbations in protein abundance directly impact biological functions, complex traits, and disease risks. Comprehensive characterization of the human proteomes to date has primarily focused on plasma or disease tissues. Factors affecting protein variation in normal human tissues remain a missing piece in the genetic regulation puzzle.

Using a tandem-tag mass-spectrometry based platform, we systematically quantified the abundance of 10,841 unique proteins in more than 700 GTEx samples, encompassing five human tissues: colon, heart, liver, lung and thyroid. These tissue samples included donors of both sexes (~30% female) and span a broad range of ages from 21 to 70. We identified 462 protein-sex associations and 716 protein-age associations. Whereas sex-biased protein abundance patterns were predominantly tissue-specific, we observed a higher level of consistency in age-dependent protein variation across tissues. Over 70% of all genes displayed positive correlations between RNA and protein abundance; however, for most genes, the correlation is modest (median correlation=0.12).

By treating protein abundance as molecular phenotypes, we identified a total of 2,013 cis-protein quantitative trait loci (pQTL) at a false discovery rate (FDR) of 0.1. Over 60% of these pQTLs coincided with transcriptome-level regulatory signatures of gene expression (cis-eQTL) or splicing (cis-sQTL) in the corresponding GTEx tissues. Cross-tissue analyses indicated that 40-80% of the pQTLs in one tissue were likely shared with another tissue. However, tissue-specific pQTLs, including those with opposite effects between tissues, were also identified.

Deciphering the genes underlying genome-wide association study (GWAS) findings remains a major challenge. Lead variants (pSNPs) at over 800 pQTLs discovered in this study had been associated with one or more complex traits or diseases in a GWAS. We propose that the corresponding pGenes can provide insights into functional genes and pathways, complementing existing strategies based on genomic, epigenomic and transcriptomic annotations. We present proof-of-principle examples, in which pQTL information nominates novel gene targets linked to GWAS loci.

In summary, this study illuminates the biological and genetic factors contributing to protein variation between individuals, and provides a valuable resource for understanding the biology of complex traits.

Title: Influences of rare protein-coding genetic variants on the human plasma proteome in 49,736 UK Biobank participants

Authors:

R. Dhindsa[1], O. Burren[2], B. Sun[3], B. Prins[4], D. Matelska[4], E. Wheeler[5], J. Mitchell[6], E. Oerton[4], V. Hristova[4], K. Smith[6], K. Carss[4], S. Wasilewski[4], A. Harper[4], D. Paul[4], M. Fabre[4], H. Runz[7], C. Viollet[4], B. Challis[4], A. Platt[6], AstraZeneca Genomics Initiative, D. Vitsios[8], E. Ashley[9], C. Whelan[10], M. Pangalos[4], Q. Wang[11], S. Petrovski[4]; [1]Baylor Coll. of Med. and AstraZeneca, Houston, TX, [2]Discovery Sci., R&D, AstraZeneca, Royston, United Kingdom, [3]Biogen, CAMBRIDGE, MA, [4]AstraZeneca, Cambridge, United Kingdom, [5]MRC Epidemiology Unit, Cambridge, United Kingdom, [6]AstraZeneca, Melbourn, United Kingdom, [7]Biogen Inc., Cambridge, MA, [8]AstraZeneca, CAMBRIDGE, United Kingdom, [9]Stanford Univ., Stanford, CA, [10]Janssen Res. & Dev., Boston, MA, [11]AstraZeneca, Chapel Hill, NC

Abstract:

Integrating human genomics and proteomics offers a promising avenue for elucidating disease mechanisms, identifying clinical biomarkers, and discovering potential drug targets. Previous population-level proteogenomic studies have primarily focused on common variation via genome-wide association studies. Although rare variants disproportionately impact human traits, their contribution to the plasma proteome remains largely unknown. Here, we investigated the influence of rare protein-coding variants on 2,923 plasma protein abundances measured in 49,736 UK Biobank individuals with matched exome-sequencing data. In a variant-level exome-wide association study (ExWAS), we identified 4,415 rare (MAF<0.1%) protein quantitative trait loci (pQTLs), of which 76% were undetected in a prior GWAS performed on the same cohort. We found an additional 289 significant gene-protein associations in gene-based collapsing analyses. Notably, 99.4% of the 691 cis pQTLs identified through gene-level collapsing analysis of protein-truncating variants (PTVs) were associated with decreased protein levels. *STAB1* and *STAB2*, encoding scavenger receptors involved in plasma protein clearance, emerged as pleiotropic loci, with 77 and 41 protein associations, respectively. We uncover several use cases of this large proteogenomics dataset, including describing an allelic series in NLRC4 relevant to diagnostic sequencing and identifying biomarkers for a fatty liver disease-associated variant in *HSD17B13*. We then introduce a novel collapsing analysis method incorporating missense pQTLs with PTVs to bolster discovery power. In employing this method phenome-wide, we uncover several gene-phenotype associations that did not otherwise achieve statistical significance, including an association between Vitamin B deficiency and *TCN1,* which encodes a vitamin B12 binding protein. Finally, we identified distinct proteomic consequences of clonal hematopoiesis, including the association between somatic variants in TET2 and increased levels of FLT3. FLT3 inhibitors are already FDA-approved for acute myeloid leukemia. If this relationship is causal, it could open up potential drug repositioning opportunities. Overall, our results highlight a considerable role for rare variation in plasma protein abundance and demonstrate the value of proteogenomics in unraveling disease mechanisms and accelerating therapeutic discovery. We make this pQTL atlas publicly available (azphewas.com).

Title: Integration of phenome-wide time-to-event modeling with genetic colocalization results for 2,941 plasma proteins and 310 diseases in 44,896 UK Biobank participants

Authors:

J. Davitte[1], D. Croteau-Chonka[2], S. Gade[3], F. Ziebell[3], P. Surendran[4], Q. Wang[4], N. Bowker[5], M. Ehm[1], A. Torralbo[6], N. Fitzpatrick[6], C. Ytsma[6], S. Denaxas[6], J. Betts[4], R. Scott[4], C. Robins[7]; [1]GlaxoSmithKline, Collegeville, PA, [2]GlaxoSmithKline, Cambridge, MA, [3]GlaxoSmithKline, Heidelberg, Germany, [4]GlaxoSmithKline, Stevenage, United Kingdom, [5]GlaxoSmithKline, Cambridge, United Kingdom, [6]Univ. Coll. London, London, United Kingdom, [7]GSK, Collegeville, PA

Abstract:

Biobank-scale measurements of the plasma proteome provide massive opportunities to assess not only genetic regulation of the human plasma proteome, but also to leverage electronic health data to explore the relationship between the plasma proteome and an individual's risk for disease. While emerging research is deeply interrogating proteogenomics, there is less work focused on integrating results from the observational study of protein measurement and disease risk with genetics.
We conducted time-to-event analyses measuring associations between 2,941 plasma proteins and 310 deeply curated physical and mental health disease phenotypes in 44,896 UK Biobank participants of European ancestry. We used Cox-Proportional Hazards (CPH) models for each protein and each disease with more than 20 events, adjusting for covariates, among individuals that were free of the disease at sample collection. To identify protein-disease pairs with genetic evidence, we checked for overlap with 50,920 significant colocalizations between disease GWAS (for the same n=310 diseases) and pQTL signals: both, cis-pQTLs (colocalization proximal to the protein-coding gene) and trans-pQTLs (colocalization at loci across the genome).
We found 57,615 protein-disease pairs that were significantly associated (Bonferroni threshold, p-value <= 6.15e-8) with time-to-new disease. Of these 57,615 protein-disease pairs, 1,978 proteins were represented (67% of the measured proteome) and 226 diseases (73% of the disease phenome). 93.9% (n=54,092) of significant CPH results indicated that increased protein abundance was associated with increased rate of disease onset.
7,368 (12.8%) protein-disease pairs with significant CPH results had at least 1 significant colocalization (cis-pQTL or trans-pQTL), while only n=126 (0.22%) had a cis-pQTL colocalization. Of 12,756 significant cis- and trans-pQTL colocalizations for protein-disease pairs with significant CPH results, 72.6% (n=9,262) had consistent directionality with CPH results, while only 57.3% (n=90) of the n=157 cis-pQTL colocalizations had consistent directionality.
This work is the largest phenome-wide study of the relationship between plasma protein abundance and time-to-incident disease. While pQTL-disease colocalization provides evidence of protein-disease links that may be important in drug discovery and development, integration of time-to-event evidence provides a new layer of information not easily investigated in genetics: knowledge that the protein is associated not just with susceptibility to the disease, but that the protein is associated with the rate at which individuals develop new disease.

Title: Multi-study pQTL analysis of Somascan proteomics in multi-ancestry TOPMed Cohorts

Authors:

C. Debban[1], U. Tahir[2], K. Pratte[3], J. Brody[4], M. Lee[5], C. Guo[3], A. Hill[3], J. Nicholas[6], D. H. Katz[7], B. Yu[8], J. G. Wilson[7], H. Lin[9], K. Kechris[10], S. A. Gharib[11], S. Rich[12], K. Taylor[13], J. Rotter[14], B. Psaty[11], M. Cho[15], S. London[16], R. Gerszten[7], L. Raffield[17], R. Bowler[3], A. Manichaikul[18]; [1]Univ. of Virginia, Charlottesville, VA, [2]BIDMC, Lexington, MA, [3]Natl. Jewish Hlth., Denver, CO, [4]Univ of Washington, Seattle, WA, [5]NIH/NIEHS, Rancho Palos Verdes, CA, [6]Univ. of North Carolina, Chapel Hill, Chapel Hill, NC, [7]BIDMC, Boston, MA, [8]Univ. of Texas Hlth.Sci. Ctr. at Houston, Houston, TX, [9]UMass Chan Med. Sch., Worcester, MA, [10]CU Anschutz, Aurora, CO, [11]UW, Seattle, WA, [12]Univ. of Virginia, Scottsville, VA, [13]The Lundquist Inst. for BioMed. Innovation, 1124 W. Carson Street, Torrance, CA 90502, Torrance, CA, [14]Lundquist Inst., Harbor-UCLA Med Ctr, Torrance, CA, [15]Brigham and Women s Hosp., Duxbury, MA, [16]NIEHS, Research Triangle Park, NC, [17]UNC - Chapel Hill, CHAPEL HILL, NC, [18]Univ Virginia, Charlottesville, VA

Abstract:

Integration of genome-wide association study (GWAS) with gene expression quantitative trait loci (eQTL) has proven a valuable approach as a first step to identifying molecular mechanisms underlying GWAS signals. However, many GWAS loci do not show evidence of colocalization with eQTLs. Motivated by the hypothesis that high-throughput proteomics can complement eQTLs for enhanced interpretation of GWAS signals, we assembled a pQTL resource by combining SomaScan proteomics versions with 1.3k, 5k and 7k aptamers measured from four community-based cohorts (Cardiovascular Health Study [CHS], Framingham Heart Study [FHS], Jackson Heart Study [JHS], Multi-Ethnic Study of Atherosclerosis [MESA]; total n=8,200), one smoking-enriched cohort (COPDGene; n=5,000), and one asthma-enriched cohort (the Agricultural Lung Health Study [ALHS]; n=1,830). The combined set of proteomics measures reflects multi-ancestry individuals representing European Americans (EUR; n=7,470) and African Americans (AFA; n=7,200) with 1,300-7,000 protein aptamer measures per sample (depending on the SomaScan version). We leveraged whole genome sequence data for the TOPMed cohorts (CHS, FHS, JHS, MESA and COPDGene) and genome-wide imputation from TOPMed for ALHS to peform pQTL mapping. We found that that accounting for unknown sources of variance by including PEER factors or PCs of hidden variance as covariates improves detection of pQTLs, with the PCs achieving similar results at far lower computational burden. Thus far, preliminary analysis of selected proteins with data from all studies, adjusting for age, sex, PCs of ancestry, and PCs of hidden variance recapitulates known variant-expression associations such as the known *SERPINA1* S and Z alleles for alpha-1 antitrypsin levels (AAT) levels. In a subset of 25 proteins on chromosome 21, we detected *cis*-pQTLs for 52% of proteins, and *trans*-pQTLs for 44% of proteins. In analysis stratified by race/ancestry, we observed a greater number of protein-associated signals in AFA compared to EUR, likely reflecting differences in patterns of linkage disequilibrium and deeper variation in the African ancestry populations. We are currently expanding our analysis genome-wide. Our pQTL mapping effort leveraging high-throughput proteomics demonstrates the value of integrating multi-ancestry samples to expand the set of protein-associated variants and identify putative molecular mechanisms underlying GWAS signals.

Session 065: Dysfunction of the power house: Mitochondrial disorders

Location: Conv Ctr/Room 147A/Level 1

Session Time: Friday, November 3, 2023, 1:45 pm - 2:45 pm

Title: Dynamics of mitochondrial DNA heteroplasmy in human normal cells

Authors:

J. An, Y. Ju; Korea Advanced Inst. of Sci. and Technology, Daejeon, Korea, Republic of

Abstract:

Mitochondria are crucial cellular organelles responsible for energy production. Hundreds to thousands of copies of mitochondrial DNA (mtDNA) exist within a single cell. The coexistence of more than one mtDNA genotype within a cell, known as heteroplasmy, has been occasionally observed in diseased tissues, including cancers, Leigh syndrome, and myoclonic epilepsy with ragged red fibers (MERRF). However, the landscape of mtDNA heteroplasmy in healthy normal somatic cells and the underlying mechanism of their acquisition over an individual's lifetime from the fertilized egg to the aging process remain poorly understood. Here, we present a comprehensive investigation of mtDNA heteroplasmy dynamics in normal human cells using single-cell resolution whole-genome sequencing data produced from 2,096 clones established from various tissues of 31 individuals. Our analysis collectively identified a total of 5,685 heteroplasmic variants (an average of 3.1 per cell), exhibiting a heteroplasmic fraction of at least ~0.3%. Among these variants, 217 (3.9%) were determined to be the fertilized-egg origin, as they were shared among multiple clones from the same individual. The variant allele fractions of the earliest heteroplasmy in somatic clones suggested a mitochondrial genetic bottleneck during embryogenesis, which reduces mtDNA diversity by reducing copy numbers to nearly one per embryonic cell. Of the heteroplasmic variants identified, 90.3% were concluded to be clearly somatically acquired in post-zygotic cell divisions. Somatically acquired mutations exhibited diverse heteroplasmic levels among cells, although a predominant strand-asymmetric signature, characterized by G>A and T>C mutations on the light strand, and selective neutrality were observed in most cells. The heteroplasmic levels of the somatic mtDNA mutations increased overall through the age of somatic cells, suggesting their continuous acquisition over time and the presence of an mtDNA's drift and/or selection during the aging process. The levels of drift and/or selection of an mtDNA varied across different cell types, suggesting their heterogeneous dynamics of mtDNA segregation in the mitoses. In conclusion, this study provides novel insights into somatic mosaicism and unravels the origin and dynamics of mtDNA heteroplasmy in normal human cells, leveraging the power of genomics-based approaches.

Title: The impact of cardiolipin acyl chain composition on tissue-specific mitochondrial phenotypes in Barth syndrome: new implications for cellular pathogenesis and therapeutic targeting

Authors:

O. Sniezek[1], N. Senoo[2], G. Butschek[3], K. Lee[4], A. Anzmann[1], S. Claypool[5], H. Vernon[6,1]; [1]Johns Hopkins Sch. of Med., Dept. of Genetic Med., Baltimore, MD, [2]Johns Hopkins Sch. of Med., Dept. of Physiology, Baltimore, MD, [3]Johns Hopkins Sch. of Publ. Hlth., Dept. of Molecular Microbiol. and Immunology, Baltimore, MD, [4]Johns Hopkins Univ., Baltimore, MD, [5]Johns Hopkins Sch. of Med., Dept. of Physiology, Baltimore, MD, MD, [6]Johns Hopkins Hosp., Baltimore, MD

Abstract:

Barth Syndrome (BTHS) is a rare, X-linked disorder caused by pathogenic variants in *TAFAZZIN*, resulting in defective remodeling of cardiolipin (CL), a mitochondrial phospholipid composed of 4 acyl chains bound to a glycerol head group. BTHS is clinically characterized by cardiomyopathy, neutropenia, and myopathy and, in contrast to many mitochondrial disorders, there are limited central nervous system effects. It is not known why defects in TAFAZZIN result in a tissue-specific pattern of disease expressivity, given that it is ubiquitously expressed. We hypothesize that tissue-specific acyl chain composition of CL is a key determinant of tissue pathology. To this end, we developed a TAFAZZIN knock out (TAZ-KO) iPSC model and differentiated these iPSCs into cardiomyocytes (CMs) (an affected tissue in BTHS) and neural progenitor cells (NPCs) (a spared tissue in BTHS).Characterization of TAZ-KO CMs and NPCs via lipidomics revealed cell-type specific CL acyl-chain composition, and RNAseq analysis demonstrated tissue-specific dysregulation of pathways associated with mitochondrial quality control and oxidative phosphorylation. We followed the RNAseq data with functional studies including characterization of respiratory chain composition via blue native (BN)-page and identified reduced abundance of complex I (CI) containing super complexes in TAZ-KO iPSC-CMs. We next evaluated mitochondrial quality control indicators and identified defects in the abundance and processing of the mitophagy protein LC3 and striking mitochondrial morphological abnormalities in TAZ-KO iPSC-CMs, but not NPCs. To investigate the effects of modification of CL acyl chain composition in CMs and NPCs on these cell-type specific pathologies, we supplemented the cells with linoleic acid (LA) to generate more unsaturated acyl chains (similar to the WT heart) or oleic acid to generate more saturated acyl chains (similar to the WT brain).Lipidomic analysis confirmed that the addition of LA or OA successfully shifted the CL acyl chains to a less saturated or more saturated profile, respectively. We are currently investigating how the modification of CL acyl chains affects the TAZ-KO CM and NPC phenotypes by characterizing mitochondrial morphological changes, OXPHOS complex assembly and activity, and mitophagy related phenotypes after lipid treatment, with preliminary studies demonstrating significant differences in mitochondrial phenotype correlating to individual lipid treatments. If modification of the TAZ-KO CM acyl profile with lipid treatments affects the cell-specific abnormalities, this may represent a novel therapeutic avenue in patients with Barth Syndrome.

Title: Impaired iron and sphingolipid metabolism in MEPAN syndrome

Authors:

D. Dutta[1], O. Kanca[1], S. Byeon[2], P. Marcogliese[3], Z. Zuo[1], R. V. Shridharan[1], J. Park[1], Undiagnosed Diseases Network, G. Lin[1], M. Ge[1], G. Heimer[4], J. Kohler[5], M. Wheeler[6], B. A. Kaipparettu[1], A. Pandey[2], H. J. Bellen[1]; [1]Baylor Coll. of Med., Houston, TX, [2]Mayo Clinic, Rochester, MN, [3]Univ. of Manitoba, Winnipeg, MB, Canada, [4]Sheba Med. Ctr., Ramat Gan, Israel, [5]Stanford Med., Palo Alto, CA, [6]Stanford Univ, Palo Alto, CA

Abstract:

Autosomal recessive mutations in *Mitochondrial Enoyl CoA Reductase* (*MECR*) cause a rare pediatric-onset neurodegenerative disease called MEPAN (**M**itochondrial **E**noyl CoA reductase **P**rotein-**A**ssociated **N**eurodegeneration) syndrome. *MECR* encodes an enzyme required for fatty acid synthesis in mitochondria. The affected individuals present with movement difficulties, dystonia, and optic atrophy. However, the underlying cause of this disease has remained elusive. Through the Undiagnosed Diseases Network, two children affected with MEPAN were enrolled. In addition, we identified five new MEPAN cases from Israel. We used patient-derived fibroblasts and fruit flies to investigate the disease mechanism. Importantly, expression of human *MECR* in the fly *mecr* null background rescues the phenotypes including lethality. However, the humanized flies that express patient variants only partially rescue the lethality and display neurodegenerative phenotypes. Similarly, neuronal loss/knockdown of *mecr* leads to impaired synaptic transmission and motor defects in aged flies. Given the role of this protein in fatty acid synthesis in mitochondria, we performed lipidomic analysis of mutants and patient-derived fibroblasts. Unexpectedly, we found that sphingolipid/ceramide levels were elevated upon loss of *MECR*/*mecr*. We also found defects in Iron-Sulfur cluster synthesis as well as elevated iron levels. Further analyses show impaired mitochondrial function and morphology in flies and patient fibroblasts. All patients show low ferritin levels further supporting iron metabolism defects in MEPAN. Finally, reducing either iron levels or ceramide levels alleviates the neurodegenerative phenotypes in flies. Interestingly, iron chelation clearly reduces ceramide levels in fly brains showing a direct effect of iron metabolism on ceramide metabolism. Altogether, we propose that loss of *MECR*/*mecr* affects iron metabolism, which elevates ceramide levels leading to neurodegeneration. In addition, our data indicate a potential therapeutic avenue for tackling MEPAN.

Session 066: Gene discovery from large-scale studies

Location: Conv Ctr/Ballroom C/Level 3

Session Time: Friday, November 3, 2023, 1:45 pm - 2:45 pm

Title: Rare variant imputation with the TOPMed reference panel combined with whole-genome sequence data in 52,658 type 2 diabetes cases and 381,683 controls identifies novel rare variant associations and informs the spectrum of pathogenicity in monogenic diabetes genes.

Authors:

A. Huerta[1,2], P. Schroeder[1,3,4], R. Mandla[1,2,4], A. Alkanaq[5,1,2,4], L. Szczerbinski[1,2,4,6,7], J. Cole[8], B. Porneala[9], K. Westerman[10,1,11], J. Li[2,1,4,10], J. Florez[2,1,4,10,12], A. Leong[2,1,4,10,12], A. Manning[10,1,11,12], M. Udler[1,2,4,10,12], J. Mercader[1,2,4,10]; [1]The Broad Inst., Cambridge, MA, [2]Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [3]Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, United States, Boston, MA, [4]Diabetes Unit, Massachusetts Gen. Hosp., Boston, MA, [5]Boston Childrens Hosp., Boston, MA, [6]Dept. of Endocrinology, Diabetology and Internal Med., Med. Univ. of Bialystok, Bialystok, Poland, [7]Clinical Res. Ctr., Med. Univ. of Bialystok, Bialystok, Poland, [8]Univ. of Colorado Sch. of Med., Denver, CO, [9]Massachusetts Gen. Hosp., Boston, MA, [10]Dept. of Med., Massachusetts Gen. Hosp., Boston, MA, [11]Clinical and Translational Epidemiology Unit, Mongan Inst., Massachusetts Gen. Hosp., Boston, MA, [12]Dept. of Med., Harvard Med. Sch., Boston, MA

Abstract:

Genome-wide association studies (GWAS) have primarily focused on common variants due to limited sample sizes in studies with whole-genome sequence (WGS) data and the inability to impute rare variants with high accuracy, hindering the identification of rare variants with large effect sizes in complex diseases such as type 2 diabetes (T2D). To fill this gap, we imputed the UK Biobank, MGB Biobank, and the GERA cohorts to the TOPMed reference panel and demonstrated over 75% concordance for variants with minor allele frequency (MAF)>0.005%. We then combined these data with WGS data from the All of Us Research Program to perform the largest multi-ancestry GWAS meta-analysis, including rare variants (MAF 0.005%-0.1%). Our study included 52,658 T2D cases and 381,683 controls. We identified nine novel variants at genome-wide significance ($p < 5\times10^{-8}$) that were either rare or ancestry-specific. We identified a rare missense variant (MAF=0.007%) in the *HNF4A* (p.R114W, rs137853336), with an 8-fold increased risk for T2D (OR=8.2, 95% CI [4.6-14.0], $p=1.08\times10^{-13}$). This is a known MODY mutation with intermediate penetrance. To test if a common variant polygenic risk score (PRS) for T2D influences the penetrance of the p.R114W variant, we conducted an analysis stratifying the carriers by the PRS tertiles. Compared to non-carriers in the middle tertile of the PRS, carriers in the highest tertile had an odds ratio (OR=18.3 [7.2-46.9], $p=1.2\times10^{-9}$) that was comparable to the effect of confirmed pathogenic *HNF4A* MODY variants (OR=19.3 [6.1-60.6], $p=4.1\times10^{-7}$). However, carriers within the lower tertile of the PRS showed a much smaller odds ratio (OR=2.62 [0.97,7.09], p=0.06). We leveraged this large-scale rare variant association data to inform the pathogenicity of variants in known monogenic diabetes genes reported in ClinVar. Of all the variants of uncertain significance and conflicting interpretation, we provide evidence supporting the reclassification of 21% of them as benign based on their sufficient statistical power but null association with T2D in our meta-analysis (OR 95% CI upper bound < 2). We also identified five variants with intermediate penetrance (OR>5 and OR 95% lower bound > 2) in *HNF1A*, *KCNJ11*, *GCK*, *POLD1,* and *WFS1*. Carriers of any of these intermediate penetrance variants showed an increased risk of T2D in an independent WGS dataset (OR=4.7 [1.86-11.74], p=0.001). Our findings show that large-scale population-based studies using a combination of high-quality imputation and WGS data can effectively identify rare variants with large effects on T2D and potentially improve the classification of variant pathogenicity in monogenic diabetes genes.

Title: Genome-wide association study identifies 30 loci associated with obsessive-compulsive disorder

Authors:

Z. Gerring[1], N. Strom[2], M. Mattheisen[3], Obsessive-compulsive disorder working group of thePsychiatric Genomics Consortium; [1]QIMR Berghofer Med. Res. Inst., Brisbane, Australia, [2]HU Berlin, Berlin, Germany, [3]Dalhousie Univ., Halifax, NS, Canada

Abstract:

Obsessive-compulsive disorder (OCD) is a chronic psychiatric illness that affects approximately 2-3% of the general population. It is characterized by obsessions and compulsions that vary in type and severity across and between individuals. The SNP-based heritability of OCD is between 28-37%, with heritability estimates for childhood-onset OCD at the higher end of the range. Despite its high heritability, existing Genome-wide Association Studies (GWAS) have provided little information on the genetic aetiology and underlying biological mechanisms of OCD. We conducted a GWAS meta-analysis of OCD by combining data from 28 European-ancestry OCD case-control cohorts, comprising 53,660 cases and 2,044,417 controls (27 cohorts, Neff~ 210k individuals). We identified 30 independent SNPs located in 29 independent genomic regions, with a SNP-based heritability of 7.91% (SE = 0.34%). Additional genome-wide analyses based on the method of cohort ascertainment (clinical, biobank, co-morbid, and self-report) did not reveal evidence that heterogeneity in sample ascertainment methods influenced our genetic findings. Using both positional and functional QTL gene-based methods, we identified 231 unique significant risk genes for OCD, 33 of which were prioritised as “high confidence” genes through their association with multiple methods. Tissue and single-cell enrichment analyses found strong enrichment of OCD GWAS signal in excitatory neurons of hippocampus and cerebral cortex, as well as D1 and D2 medium spiny neurons. OCD was significantly positively correlated with all tested psychiatric phenotypes, the highest correlations being with anxiety followed by depression, anorexia nervosa, Tourette syndrome, and post-traumatic stress disorder. Our findings highlight the need for larger genetic studies to gain a deeper understanding of the complex genetic landscape and to elucidate the biological pathways contributing to OCD susceptibility. Our study represents a significant step forward in advancing our knowledge of OCD genetics and provides a foundation for future research aiming to develop targeted interventions for this debilitating disorder.

Title: Multi-ancestry genome-wide meta-analysis in Parkinson's disease

Authors:

J. Kim[1], D. Vitale[1], D. Véliz Otani[2], M. Lian[3], K. Heilbron[4], the 23andMe Research Team, H. Iwaki[1], J. Lake[1], C. W. Solsberg[5], H. Leonard[1], M. Makarious[1], E-K. Tan[6], A. Singleton[1], S. Bandres-Ciga[1], A. Noyce[7], C. Blauwendraat[1], M. Nalls[1], J. Foo[3], I. Mata[8], Global Parkinson's Genetics Program (GP2); [1]NIH, Bethesda, MD, [2]Inst. Natl. de Ciencias Neurológicas, Lima, Peru, [3]Nanyang Technological Univ. Singapore, Singapore, Singapore, [4]23andMe, Sunnyvale, CA, [5]UCSF, San Francisco, CA, [6]Duke NUS Med. Sch., Singapore, Singapore, [7]Queen Mary Univ. of London, London, United Kingdom, [8]Cleveland Clinic Fndn., Cleveland, OH

Abstract:

Background Genome-wide association studies (GWASes) have identified over 90 independent risk variants for Parkinson's disease (PD), but all previous studies have been performed within a single ancestry at a time. Here we performed the first large-scale multi-ancestry meta-analysis of PD GWASes, incorporating data from 49,049 PD cases, 18,785 proxy cases, and 2,458,063 controls including individuals of European, East Asian, Latin American, and African ancestry. **Method** To account for the linkage disequilibrium pattern differences across ancestries, we use two different meta-analysis models, random-effects model and a meta-regression model MR-MEGA, to identify genetic risk variants shared across ancestries and identify variants with differences in effect estimates due to ancestry-level genetic variations. We leveraged the different LD patterns across different populations to finemap known genetic risk loci. To nominate potential risk genes near the novel loci, we used summary-based mendelian randomization (SMR) with multi-omic data. Multi-marker Analysis of GenoMic Annotation (MAGMA) gene ontology enrichment analysis was used to identify gene sets associated with PD across ancestries. **Results** We identified 78 independent genome-wide significant loci including 12 potential novel loci (*MTF2, PIK3CA, ADD1, SYBU, IRS2, USP8, PIGL, FASN, MYLK2, USP25, EP300, PPP6R2*). 3 of the novel loci (*IRS2, MYLK2, USP25*) showed evidence for significant ancestry heterogeneity ($P_{ANC\text{-}HET} < 0.05$). MR-MEGA fine-mapped 6 putative causal variants at 6 known loci in (*TMEM163, TMEM175, SNCA, CAMK2D, HIP1R, LSM7*). Using summary-based mendelian randomization, we nominated 25 putative risk genes near the novel loci, many of them relevant to known pathways related to PD risk such as vesicular transport pathways. MAGMA gene ontology enrichment analysis found enrichment in immune cell pathways (microglial cell/macrophage proliferation, NK T Cell differentiation), further highlighting the potential role of inflammation and the immune system in PD pathology. **Conclusion** Our work identified PD risk loci shared across ancestrally disparate populations and laid the foundation for future efforts to identify globally relevant genetic risk for PD.

Title: Assessing the value of 490,000 whole-genome sequences in drug target discovery

Authors:

Y. Lo[1], J. Liu[1], L. Howe[2], J. Davitte[1], J. Sandhuria[1], T. Johnson[2], Y. Aulchenko[2], R. Scott[2], A. Cortes[2]; [1]GSK, Upper Providence, PA, [2]GSK, Stevenage, United Kingdom

Abstract:

Genome-wide association studies (GWASs) have successfully uncovered many gene-phenotype associations useful for understanding disease mechanisms and identifying drug targets. Recent advances in whole-genome sequencing (WGS) studies of large cohorts provide almost complete coverage of the human genome, allowing us to explore gene-phenotype associations in genomic regions that are not well captured by targeted sequencing or array genotyping methods. Here we describe the analysis of ~490,000 whole-genome sequences from the UK Biobank. We identified over 800 million high-quality single-nucleotide variants, >95% of which were ultra-rare with minor allele count less than 5 and mostly novel. Among the >140M common variants in WGS, 87% were previously not well-imputed or absent in the UK Biobank array genotypes. Among all exonic variants in the WGS, > 8% were not captured by whole-exome sequencing. We performed genome-wide association tests using these common variants and aggregated coding rare variants by functional annotations to perform region-based association tests. We illustrated the value of WGS by comparing the number and magnitude of associations identified in specific indications. For height, we highlighted known associated loci where the lead variants were not previously captured by GWAS and novel associations. We demonstrate how rare variant associations helped determine the likely causal gene and direction in regions with GWAS signals. Across >1,500 curated binary and quantitative trait, we discovered gene-phenotype associations with both common and rare variant association signals that were not found in array genotyping and exome sequencing of the same cohort. These results show the important added value and potential of biobank-scale WGS for the understanding of variant to gene to function mechanisms and the discovery of potential drug targets.

Session 067: Genetic analysis of neurological disorders

Location: Conv Ctr/Room 207A/Level 2

Session Time: Friday, November 3, 2023, 1:45 pm - 2:45 pm

Title: Multi-ethnic meta-analysis of earlier onset Alzheimer's Disease identifies novel risk loci

Authors:

J. Bradley[1], D. Western[1], C. Wang[1], E. Lucio Da Fonseca[2], J. Thulaseedhara Kurup[3], M. Jean-Francois[4], Alzheimer's Disease Genetics Consortium, G. Beecham[4], C. Reitz[5], V. Fernandez[1], C. Cruchaga[1]; [1]Washington Univ. in St. Louis, St. Louis, MO, [2]Univ. of Miami, Miller Sch. of Med., Miami, FL, [3]Columbia Univ. Taub Inst. for Res. on Alzheimer's Disease and the Aging Brain, New York, NY, [4]Univ. of Miami, Miami, FL, [5]Columbia Univ., New York, NY

Abstract:

Background: Alzheimer’s disease (AD) is a highly polygenic disease that presents with relatively earlier onset (≤70yo; EOAD) in about 5% of cases. 90% of EOAD cases remain unexplained by pathogenic mutations. This project aims to identify novel EOAD-associated genes through a large-scale, multi-ethnic genome-wide association study (GWAS). **Method**: We leveraged data of over 70,000 individuals from the Alzheimer Disease Genetics Consortium and the Knight-ADRC. Individuals with age at onset ≤ 70 were considered EOAD cases (CA), whereas healthy individuals older than 70 were considered controls (CO). Our final dataset, consisting of 6,282 CA and 13,386 CO for non-Hispanic Whites (NHW); 782 CA and 3,663 CO for African American (AA); and 375 CA and 838 CO for Asian, was used to performed single-variant analysis (SVA) and multi-ethnic meta-analysis. Additionally, we used a combination of evidence from annotation, gene-based analyses, and overlap with in-house available protein quantitative trait loci (pQTL) to prioritize functional genes in each locus. **Result**: SVA identified a total of 48 independent signals in 28 loci across the three ethnicities. Two additional loci were identified by trans-ethnic meta-analysis for a total of 30 associated loci, 21 of which were novel compared to the most recent AD risk GWAS by Bellenguez et al (2022). We identify strong genetic overlap between EOAD and LOAD as suggested by results from associating LOAD PRS with EOAD status (with APOE: R2=0.103, P<1×10-300; without APOE: R2=0.028, P=1.200×10-143) in NHW. Lastly, using evidence from annotation, gene-based analysis, and overlap and colocalization with molecular QTLs we nominate a single prioritized gene for thirteen out of 21 novel loci. **Conclusion**: This is the largest trans-ethnic GWAS for EOAD to date and will be instrumental to identifying novel variants and pathways implicated in unexplained EOAD. We confirmed nine previous AD risk loci, identified 21 novel EOAD loci, and were able to nominate prioritized genes for 13 of the 21 novel loci. Some prioritized genes such as PDGFA and NELFB are included in gene sets for immune and reactome response potentially implicating those systems for EOAD onset and progression. Other genes like SLC35E2B and FAM86B3P point to a potential comorbidity with body mass index and schizophrenia, respectively. Cell type enrichment, pathway enrichment, and protein-protein interaction analyses are ongoing to further evaluate the functional significance of the prioritized genes.

Title: Improved late-onset Alzheimer's disease risk prediction and stratification through a combined genetic risk score of common and rare variants

Authors:

E. Suh[1], M. Shivakumar[1], A. Naj[1], A. Saykin[2], M. Ritchie[1], L. Shen[1], D. Kim[1]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Indiana Univ. Sch. of Med., Indianapolis, IN

Abstract:

Late-onset Alzheimer's disease (AD) has shown high heritability of 60 to 80%. Despite GWAS and polygenic risk score (PRS) estimates to quantify risk, more than half of AD heritability remains unexplained. Rare variation is a widely discussed potential source of missing heritability in AD. However, PRS only accounts for common variants and does not incorporate rare variants, which may have greater impact on AD risk. We investigate the effects of rare genetic variations on predicting AD risk and performing risk-based stratification, in addition to the integrative effects of both common and rare variants. We used SAIGE-GENE+ to conduct gene burden rare variant analyses using whole genome sequencing data of European samples from the Genome Center for Alzheimer's Disease (GCAD) release 3 (r3) dataset (N=6197). Different binning strategies were used to define variants, including varying minor allele frequency thresholds and variant annotation masks according to functional importance. Significantly associated genes were used to generate a rare variant score, which was calculated as the weighted sum of the aggregated number of variants carried by each sample within the given genes. This score was tested on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (N=1545) for predictive performance using the area under the receiver operating characteristic curve (AUC) and risk-based stratification analysis. The rare variant score was integrated with PRS to assess the combined effects of common and rare variants for AD prediction. Five genes—*NTM*, *BCAM*, *ADGRL2*, *COMT*, and *FAM81B*—were significantly associated with AD in the GCAD r3 cohort. A literature review revealed that these genes are linked to AD, related diseases, and AD pathology. Interestingly, the significant genes were binned based on the intron variant mask. Compared to PRS alone (AUC=0.602), the rare variant score displayed similar performance (AUC=0.605). When combining common and rare effects, the AUC increased to 0.641. The combined score also demonstrated lower risk for controls and, conversely, higher risk for AD patients. Based on the stratification analysis, the score exhibited improved calibration, in which the number of predicted AD patients increased smoothly as the risk score increased. These findings demonstrate that integrating the effects of common and rare variants enhances AD risk prediction and stratification. This suggests that rare variants may significantly contribute to genetic-based treatment strategies for AD. Additionally, the use of intron masking suggests the potential power of focusing on non-coding rare variants to improve risk prediction.

Title: Neuropathology GWAS identifies novel genes involved in amyloid, vascular brain injury, and cerebrovascular disease from common variants.

Authors:

J. Pasteris[1], D. Godrich[1], T. Phongpreecha[2], E. Martin[3], A. Naj[4], R. Mayeux[5], P. Crane[6], D. Bennet[7], M. Pericak-Vance[3], W. Scott[1], L. Molina[8], M. Cuccaro[9], W. Kukull[6], G. Schellenberg[10], T. Montine[2], G. Beecham[1]; [1]Univ. of Miami, Miami, FL, [2]Stanford Univ., Stanford, CA, [3]Univ. of Miami Miller Sch. of Med., Miami, FL, [4]Univ. of Pennsylvania, Philadelphia, PA, [5]Columbia Univ, New York, NY, [6]Univ. of Washington, Seattle, WA, [7]RUSH Med. Coll., Chicago, IL, [8]Inst. de Investigaciones Biomédicas August Pi i Sunyer, Clínic Barcelona, Barcelona, Spain, [9]John P. Hussman Inst. for Human Genomics, Miami, FL, [10]Univ Pennsylvania Sch Med, Philadelphia, PA

Abstract:

Background. Alzheimer's disease (AD) is a progressive dementia, with pathology characterized by amyloid plaques and neurofibrillary tangles. However, AD pathology rarely presents in isolation; complex co-occurring “related dementia” (RD) pathologies can increase severity of cognitive decline. Genetic studies of these pathologies (amyloid, tau, vascular brain injury (VBI), and cerebrovascular disease (CVD), TDP-43, etc.) have typically had sample sizes, and the traits are typically studied in isolation. To address this, we have performed a genome-wide association study of AD/RD neuropathology.
Methods. We utilized data from as many as 12,509 autopsied and genotyped individuals, primarily through the Alzheimer Disease Genetics Consortium, including samples from NACC, ACT, ROS/MAP, NIA-FBS, Mayo, TGEN, UPitt, the University of Miami, and Instituto de Investigaciones Biomédicas August Pi i Sunyer (IDIBAPS). Participants were age 50+ at death and consistent with European genetic ancestry. Genetic data were available from SNP arrays imputed to TOPMed and analyses were conducted only on common variants (MAF >0.01). For ordinal traits, we used ordinal logistic regressions (R, ordinal package), while binary traits had logistic regressions with Rvtests (v20171009). Models adjusted for sex, age at death, and the first three PCs (PC-AiR; GENESIS; R v4.1), both with and without APOE4 count. We then meta-analyzed across the cohorts (METAL).
Results. We observed novel genome-wide significant (GWS) associations with amyloid, VBI, and CVD severity, including *PSMG1* / *LINC01700* (top SNP: rs2836880, MAF = 0.478, P = 1.24E−8) for amyloid presence, *DOCK4* for infarcts (top SNP: rs6976029, MAF = 0.277, P = 1.25E-8), *SLAIN2* / *SLC10A4* (top SNP: rs4346719, MAF = 0.364, P = 3.88E−8) for white matter rarefaction, *LOC107985854* / *LINC00276* (top SNP: rs41446051, MAF = 0.363, P = 4.75E-9) and *VAPA* / *LINC01254* (top SNP: rs206499, MAF = 0.266, P = 1.56E-8) for atherosclerosis, and *LINC01098* / *LIN00290* (top SNP: rs112992465, MAF = 0.012, P = 2.05E−8) for CVD presence.
Conclusion. This is one of the largest studies to date aimed at identifying genetic factors associated with AD neuropathology. Our findings offer new insights into the genetics of amyloid, VBI, and CVD, which are commonly studied in the AD landscape. We acknowledge the need for a larger dataset and will increase ours as the neuropath phenotype harmonization consortium develops. Continuing to uncover the genetic bases of these phenotypes could possibly help us think about AD treatment in a different frame and ameliorate the harmful side-effects, such as brain bleeds, that have been observed in recent clinical trials.

Title: A large-scale meta-analysis of genome-wide association studies reveals genetics underlying Parkinson's Disease leveraging electronic health records.

Authors:

R. Wang[1], S. Gelfman[1], A. Moscati[1], N. Parikshak[1], V. Rajagopal[1], V. Pounraja[1], A. Ayer[1], J. Mbatchou[1], J. Marchini[1], O. Levy[2], DiscovEHR, Mayo Clinic Project Generation, Penn Medicine Biobank, E. Stahl[1], G. Coppola[1]; [1]Regeneron Genetics Ctr., Tarrytown, NY, [2]Regeneron Pharmaceuticals, Tarrytown, NY

Abstract:

To better understand the genetics of Parkinson's disease (PD), we performed a genome-wide association meta-analysis including 27,450 cases, 19,067 proxy-cases with family history of PD from the UK biobank, and 766,096 controls of European descent across seven cohorts (454,350 individuals from UK biobank, and 339,196 individuals from six EHR cohorts). Single-cohort association analyses and meta-analyses were performed as implemented in REGENIE and METAL. We observed common genome-wide significant loci in *GBA* (OR=1.341, p=5.57e-17), *STK39* (OR=1.087, p=4.65e-10), *SNCA* (OR=1.103, p=1.45e-26), *TMEM175* (OR=1.114, p=1.00e-20), *BAG3* (OR=1.055, p=1.31e-9), and *MAPT* (OR=0.887, p=2.03e-28), consistent with previous studies. In an age stratified analysis including either early-onset PD and younger controls (<65), or late-onset PD and older controls (>=65), we further observed that *SNCA* has higher OR in early-onset PD (EOPD, OR=1.259, p=3.47e-10) than in late onset PD (LOPD, OR=1.158, p=2.19e-8), while common signals near *MAPT* are more strongly associated with LOPD (OR=0.85, p=4e-07) than EOPD (OR=0.82, p=1.8e-05). We also replicated the effects of the genes *PRKN* (OR=47.541, p=2.51e-3) and *PINK1* (OR=102.176, p=8.54e-5) based on the burden of rare coding variants in a recessive inheritance model, supporting evidence for mitophagy in PD risk. Testing for genetic correlation, we identified positive genetic correlation between PD and Alzheimer's disease and cognitive traits, whereas negative genetic correlations are shown with ever-smokers and heavy-smokers in the UK biobank. These results provide a comprehensive survey of underlying genetic risks for PD and validate the use of large EHR cohorts is effective for large-scale PD genetics.

Session 068: More data, same problems: Bias in large-scale studies

Location: Conv Ctr/Ballroom A/Level 3

Session Time: Friday, November 3, 2023, 1:45 pm - 2:45 pm

Title: Correction for Collider Bias in the Genome-wide Association Study of Diabetes-Related Heart Failure due to Bidirectional Relationship between Heart Failure and Type 2 Diabetes

Authors:

Y. Sun[1,2], C. Liu[3], Q. Hui[1], J. J. Zhou[4], J. Gaziano[5,6], P. W. Wilson[1,2], the Million Veteran Program, J. Joseph[7], L. Phillips[1,2]; [1]Emory Univ., Atlanta, GA, [2]Atlanta VA Hlth.care System, Decatur, GA, [3]Emory Univ., Decatur, GA, [4]UCLA, Los Angeles, CA, [5]VA Boston Hlth.care System, Jamaica Plain, MA, [6]Brigham and Women's Hosp., Harvard Med. Sch., Boston, MA, [7]VA Providence Hlth.care System, Providence, RI

Abstract:

Heart failure (HF) is a major complication of diabetes, and contributes to the rising health burden of adults. Type 2 diabetes (T2D) is a known risk factor of HF across demographic groups through worsening numerous cardio-metabolic conditions. On the other hand, metabolic impairment, including elevated T2D incidence is a hallmark of HF pathophysiology. We investigated the bidirectional relationship between T2D and HF, and identified genetic associations with diabetes-related HF after correction for potential collider bias. We performed a genome-wide association study (GWAS) of HF to identify genetic instrumental variables (GIVs) for HF, and to enable bidirectional Mendelian Randomization (MR) analysis between T2D and HF. Since genetic factors and HF can independently influence T2D, collider bias may occur when T2D (i.e., collider) is controlled for by design or analysis. Thus, we conducted GWAS of diabetes-related HF with correction for collider bias. We first identified 61 genomic loci, including 24 novel loci, significantly associated with all-cause HF in 114,275 HF cases and over 1.5 million controls of European ancestry. Combined with the summary statistics of a T2D GWAS, we obtained 59 and 82 GIVs for HF and T2D, respectively. Using a two-sample bidirectional MR approach, we estimated that T2D increased HF risk (OR 1.07, 95% CI 1.04-1.10), while HF also increased T2D risk (OR 1.60, 95% CI 1.36-1.88). Then we performed a GWAS of diabetes-related HF corrected for collider bias due to evidence of HF affecting incidence of T2D. After removing the spurious association of *TCF7L2* locus due to collider bias, we identified two genome-wide significant loci (chromosome 4 and 9) associated with diabetes-related HF in the Million Veteran Program, and replicated the associations in the UK Biobank study. We identified novel HF-associated loci to enable bidirectional MR study of T2D and HF. Our MR findings support T2D as a HF risk factor and provide strong evidence that HF increases T2D risk. As a result, collider bias leads to spurious genetic associations of diabetic HF, which can be effectively corrected to identify true positive loci. Evaluation of collider bias should be a critical component when conducting GWAS of complex disease phenotypes such as diabetes-related cardiovascular complications.

Title: Testing for differences in polygenic scores in the presence of confounding.

Authors:

J. Blanc, J. J. Berg; Univ. of Chicago, Chicago, IL

Abstract:

Polygenic scores have become an important tool in human genetics, enabling the prediction of individual phenotypes from their genotypes. Understanding how the pattern of differences in polygenic score predictions across individuals intersects with variation in ancestry can provide insights into the evolutionary forces acting on the trait in question, and is important for understanding health disparities. However, because most polygenic scores are computed using effect estimates from population samples, they are susceptible to confounding by both genetic and environmental effects that are correlated with ancestry. The extent to which this confounding drives patterns in the distribution of polygenic scores depends on patterns of population structure in both the original estimation panel and in the prediction/test panel. Here, we use theory from population and statistical genetics, together with simulations, to study the procedure of testing for an association between polygenic scores and axes of ancestry variation in the presence of confounding. We use a general model of genetic relatedness to describe how confounding in the estimation panel biases the distribution of polygenic scores in a way that depends on the degree of overlap in population structure between panels. We then show how this confounding can bias tests for associations between polygenic scores and important axes of ancestry variation in the test panel. We then use the understanding gained from this analysis to develop a method that leverages the patterns of genetic similarity between the two panels to guard against these biases, and show that this method can provide better protection against confounding than the standard PCA-based approach.

Title: Height-related assortative mating revealed implications in long-term cohabitation: Tehran Cardiometabolic Genetic Study (TCGS)

Authors:

M. Akbarzadeh[1], **P. Riahi**[1], A. Saeidian[2], A. Tenesa[3], C. Hogan[4], M. March[5], K. Guity[6], M. Amiri Roudbar[7], A. Zahedi[1], M. Zarkesh[1], F. Neshati[1], M. Hedayati[1], F. Azizi[1], H. Hakonarson[8], M. Daneshpour[1]; [1]Res. Inst. for Endocrine Sci., Shahid Beheshti Univ. of Med. Sci., Tehran, Iran, Islamic Republic of, [2]Ctr. for Applied Genomics, The Children's Hosp. of Philadelphia, Philadelphia, PA, [3]The Roslin Inst., Royal (Dick) Sch. of Vet. Studies, The Univ. of Edinburgh, Edinburgh, United Kingdom, [4]Div. of Hepatology, Temple Univ. Hosp., Philadelphia, PA, [5]Children s Hosp. of Philadelphia, Philadelphia, PA, [6]Cellular and Molecular Endocrine Res. Ctr., Res. Inst. for Endocrine Sci., Shahid Beheshti Univ. of Med. Sci., Tehran, Iran, Islamic Republic of, [7]Dept. of Animal Sci., Safiabad-Dezful Agricultural and Natural Resources Res. and Ed. Ctr., Tehran, Iran, Islamic Republic of, [8]Children's Hosp. of Philadelphia, Philadelphia, PA

Abstract:

To provide an applied framework for assessing the genetic contribution to assortative mating (AM) using height as a model trait and disclose the trace of certain pieces of evidence of AM in the form of the shared environmental effects in a long-term cohabitation on the spouses' anthropomorphic and lipid serum levels. 2,315 genotyped couples were extracted from the Tehran Cardiometabolic Genetic Study (TCGS). Pearson correlation analysis assessed the relationship between spouses' height. GWAS was conducted for height and spouses' height of unrelated people (1,655 spouses) with 631,579 autosomal SNPs and correlations between the genetic values of individuals were evaluated using the best linear unbiased predictions (BLUP) effects of all SNPs. We used the recent GWAS meta-analysis on 5.4M individuals of height to calculate the polygenic risk scores (PRS) for spouses' height. A subset of 1,038 spouses out of 2,315 couples were subsequently selected to enter the longitudinal resemblance, to be assessed in terms of their anthropomorphic traits and lipid serum levels in a 15-year follow-up. The correlation coefficient of height between spouses was estimated as r =0.248. We show that a person's genotype determines 8.01 % of the variation in the spouse's height. Results for correlational analysis of the effect sizes driven from GWAS of heights between males and females are estimated as 0.468. Furthermore, the genotype of an individual predicts their partner's height for those spouses with only one genotyped data information (1,982 couples) with ~12.6% (using genetic values) and 20.57% (using PRS) accuracy. Raw estimation of height SNP-based heritability was 0.72, while under random mating was 8% less, 0.64. Long-term spousal resemblance revealed an increasing trend for correlation between husbands and wives in terms of their lipid serum level and obesity-related traits. Bayesian hierarchical meta-analysis validated the increasing trend of the Pearson correlation coefficient for spousal resemblance. Our findings support the AM hypothesis for height with a significant spousal correlation and show that selecting the spouse's height is genetically determined. Besides, we provide data showing that AM is predicted to result in an 8% increase in the heritability of height, which is related to the assortative nature of alleles in the population and not to the segregation of genetic variations. Finally, as one of the evolutionary consequences of AM, long-term spousal resemblance provided an increasing trend for correlation between spouses in terms of their lipid serum level and obesity-related traits.

Title: Estimating relative effective dilution for 34 diseases in the Intervene consortium

Authors:

K. Läll[1], B. Jermy[2], F-D. Pajuste[1], T. Laisk[1], K. Zguro[3], J. Mehtonen[2], B. Wolford[4], T. Hartonen[2], Z. Yang[5], R. Monti[6], S. Kanoni[7], Y. Wang[8], J. Wanner[9,2], O. Youssef[10,11], A. Martin[8], N. Mars[2,12], Estonian Biobank research team, S. Furini[13], D. van heel[14], K. Hveem[15], H. Heyne[9,2,16], A. Palotie[2,12], A. Ganna[2,12], S. Ripatti[2,12,17], R. Mägi[1]; [1]Inst. of Genomics, Univ. of Tartu, Tartu, Estonia, [2]Inst. for Molecular Med. Finland, FIMM, HiLIFE, Univ. of Helsinki, Helsinki, Finland, [3]Univ. of Siena, Siena, Italy, [4]Norwegian Univ. of Sci. and Technology, Trondheim, Norway, [5]Univ. of Helsinki, Helsinki, Finland, [6]Hasso Plattner Inst., Digital Hlth.Ctr., Univ. of Potsdam, Potsdam, Germany, [7]Queen Mary Univ. of London, Cambridge, United Kingdom, [8]Massachusetts Gen. Hosp., Boston, MA, [9]Hasso Plattner Inst., Potsdam, Germany, [10]Helsinki Biobank, Helsinki, Finland, [11]Pathology Dept., Univ. of Helsinki, Helsinki, Finland, [12]Massachusetts Gen. Hosp. and Broad Inst. of MIT and Harvard, Cambrdige, MA, [13]Med Biotech Hub and Competence Ctr., Dept. of Med. Biotechnologies, Univ. of Siena, Siena, Italy, [14]Barts and The London, London, United Kingdom, [15]NTNU, Trondheim, Norway, [16]Hasso Plattner Inst., Mount Sinai Sch. of Med., New York, NY, [17]Dept. of Publ. Hlth., Univ. of Helsinki, Helsinki, Finland

Abstract:

Researcher often wish to define diseases in a harmonized way across cohorts. The aim is studying phenotypic misclassification in three large biobanks and its effect on downstream analyses. We apply Phenotypic Measurement of Effective Dilution (PheMED, doi:10.1101/2023.01.17.23284670), a newly developed tool for detecting hidden biases in genome-wide association studies (GWAS) due to phenotypic misclassification. Using FinnGen's clinical endpoint library, 34 were diseases were defined in Estonian Biobank (EstBB), FinnGen and UK Biobank (UKBB). While defining the diseases, FinnGen filtered for sources of diagnoses and applied exclusion criteria for controls whereas UKBB and EstBB applied no filtering besides UKBB's sole inclusion of secondary care data. GWAS for all diseases were performed in individuals of European ancestry and the summary statistics were used as input for the PheMED analysis. Next, we calculated polygenic risk scores (PGS) using MegaPRS for a subset of 18 diseases. Associations between disease and PGS were estimated by fitting Cox proportional hazard models. Phenotype dilution effect on PGS performance was examined. We also investigated the association between genetic correlation and PheMED analysis. Results presented here are for 24 diseases with at least 1000 cases in each biobank. Using FinnGen data R6 as reference, the average relative effective dilution was p_{MED}= 1.31, (IQR 1.04 - 1.48) in EstBB and p_{MED}= 1.39, (IQR 1.12 - 1.65) in UKBB. After correcting for multiple testing, 13 endpoints in EstBB and 14 endpoints in UKBB displayed statistically significant dilution. When classifying diseases into groups, oncological diseases showed the lowest median effective dilution (median p_{MED}= 1.1, IQR 0.88 - 1.22), and musculoskeletal diseases the highest median effective dilution (median p_{MED}= 1.79, IQR 1.09 - 1.92). The Pearson correlation of -0.53 (95% CI -0.827 - 0.0031) between PGS effects and PheMED values in EstBB indicated that larger misclassification is associated with weaker PGS-disease associations. The Pearson correlation between PheMED values and genetic correlations for diseases with non-zero heritability estimate from LDSR was -0.045 (95% CI -0.49 - 0.41) in EstBB and -0.13 (95% CI -0.54 - 0.33) in UKBB, implying that genetic correlation and effective dilution provide complementary information about phenotypic similarity between cohorts. In conclusion, PheMED analysis sheds light into how well harmonization across biobanks has worked in practice. Our analysis shows that FinnGen's GWAS have systematically higher effect estimates across diseases compared to EstBB and UKBB despite the phenotype harmonization.

Session 069: Not fair and balanced: Learning about regulation from allelic bias

Location: Conv Ctr/Room 146B/Level 1

Session Time: Friday, November 3, 2023, 1:45 pm - 2:45 pm

Title: Large-scale characterization of naturally occurring human regulatory genetic variation.

Authors:

A. Boytsov[1], S. Abramov[1], A. Teodosiadis[1], M. Brannon[1], A. Cote[1], J. Stamatoyannopoulos[2,1], J. Vierstra[1]; [1]Altius Inst. for BioMed. Sci., Seattle, WA, [2]Univ of Washington, Seattle, WA

Abstract:

Genetic variation linked to trait and disease associations is markedly enriched within regulatory regions. Previously, we have shown that functional effects at single nucleotide variants can be ascertained directly from epigenomics assays via quantification of allelic imbalance. However, such analyses had insufficient data to quantify cell-selective penetrance of regulatory variation and to evaluate the extent to which these variants are subject to purifying evolutionary selection. Here, we present a framework for uniform large-scale analysis of regulatory variation directly from chromatin accessibility. We uniformly processed over 3,800 high-quality DNase I and 6,800 publicly available ATAC-seq experiments spanning diverse primary cell types, tissues and biological states. Because chromatin accessibility assays re-sequence regulatory DNA at high depth, we performed direct genotyping from these assays to identify 15 million distinct SNPs within regulatory DNA across all samples. Using these personalized genotypes, we quantified allelic imbalance for over 10 million heterozygous positions. We identified over 110,000 variants altering chromatin accessibility (CAVs), representing the most comprehensive set of regulatory variants functionally characterized within their native genomic context to date. By aligning alleles by population frequency, we find that major alleles at rare variants are systematically more accessible than minor alleles and that sequence context alone does not explain the major allele preference, suggesting that the allelic preferences are driven by negative selection acting on rare variants in regulatory regions. We find that 15% of the identified CAVs impact chromatin in a cell-selective manner and are particularly enriched within variation associated with phenotypic traits and pathologies. Furthermore, the cell-context-dependent CAVs are enriched within the consensus recognition sequences for cell-specific TFs. Surprisingly, we observe that only a small fraction of regulatory variation (~15%) manifests in chromatin imbalance even when occurring in the core nucleotides that mediated protein-DNA interactions, suggesting that higher-order interactions between transcriptional regulators modulate the penetrance of regulatory genetic variation. To investigate this, we developed a novel deep-learning model to predict and interpret the effects of single nucleotide variant effects that considers transcription factor binding interdependencies. Taken together, these data provide an unprecedented mechanistic view on how genetic variation shapes human regulatory DNA and trait variation.

Title: Alternative splicing dynamics and allele-specificity during T cell activation

Authors:

N. Hackert[1,2,3], V. Aguiar[1,2], R. Grieshaber-Bouyer[3,4], M. Gutierrez-Arcelus[1,2]; [1]Boston Children's Hosp., Harvard Med. Sch., Boston, MA, [2]Broad Inst. of MIT and Harvard, Cambridge, MA, [3]Heidelberg Univ. Hosp., Heidelberg, Germany, [4]European Molecular Biology Lab. (EMBL), Heidelberg, Germany

Abstract:

Non-coding risk variants for multiple autoimmune diseases likely affect gene regulation in memory $CD4^+$ T cells. However, co-localization analyses with expression and splicing quantitative trait loci (QTL) have only provided putative functional relevance to up to 38% of tested disease-associated variants, emphasizing the need to ascertain cell state-specific regulatory effects.

Here, we investigate dynamic alternative splicing in activated T cells and develop and apply a computational method to quantify allele-specific junction usage, reflecting the influence of genetic variation on RNA splicing.

To quantify differential splicing based on junction reads, we applied leafcutter to bulk RNA-seq data from anti-CD3/CD28 stimulated memory $CD4^+$ T cells from 24 genotyped individuals at 8 time points between 0 and 72h after activation. We found differential splicing between consecutive time points with a median of 8,546 significant events in 5,416 genes (5% FDR), including genes central to T cell biology, such as *IL7R* and *SRSF2*. We discovered that transcriptome-wide intron retention exhibited a dynamic pattern, displaying a decrease during early and mid activation, while increasing during late activation.

We further developed SNPJunkie, a computational tool to quantify allelic junction read counts from spliced RNA-seq alignments. SNPJunkie is implemented in Rust, allowing fast and memory-efficient quantification of allelic counts from reads aligned to one or more SNP-junction combinations (events).

Using SNPJunkie, we extracted a total of 59,633 distinct non-MHC QC-passing events involving 16,142 heterozygous SNPs across 184 samples (median of 7,116 events per sample). We found evidence of significant allelic imbalance (binomial test, 5% FDR) for 4,971 events (median of 91 events per sample), 1,665 of which did not display significant allelic imbalance at the level of their respective SNP (binomial test, 5% FDR, N = 445) and/or displayed a significantly different allelic count ratio from the SNP-level (Fisher's exact test, 5% FDR, N = 1,542). Those events' variants would thus have been missed or de-prioritized by conventional allele-specific expression analyses. We further found 230 of those events to overlap 111 putative target genes for risk loci of immune-mediated diseases, such as *CD44* for systemic lupus erythematosus or *CD6* for multiple sclerosis and inflammatory bowel disease.

Our work highlights the dynamic nature of alternative splicing during T cell activation and introduces a new computational method allowing fast and unbiased quantification of the effects of genetic variation on exon-exon junction usage on the level of individual samples.

Title: Genotype calling from expression data in Recount3.

Authors:

A. Razi[1], C. C. Lo[1], S. Wang[1], J. T. Leek[2], K. D. Hansen[3]; [1]Johns Hopkins Univ., Baltimore, MD, [2]Fred Hutchinson Cancer Ctr., Seattle, WA, [3]Johns Hopkins Univ, Baltimore, MD

Abstract:

Advances in high throughput sequencing technologies have enabled gene expression studies of human health and disease. Large amounts of transcriptome data have been uploaded to public repositories in the last decade representing a broad range of experimental conditions, sequencing technologies, and hypothesis under studies. Access to existing data promotes reproducibility and reusability, however, processing raw RNA seq data is time consuming and resource demanding. The Recount3 repository has resolved this issue by providing public access to uniformly processed RNA-seq data. The repository includes 316,443 uniformly processed human RNA-seq samples from Genotype Tissue Expression (GTEx), The Cancer Genome Atlas (TCGA), and Sequence Read Archive (SRA). Coupling this expression data to genotypes would enable unprecedented large-scale data analyses such as variant prioritization, eQTL, and allele specific expression (ASE) analysis. However, genotypes are missing for most publicly deposited RNA-seq samples.To address this problem, we have developed a maching learning model to genotype Recount3 human RNA-seq samples. Our model is fast and only uses raw read counts from the reference and alternative alleles whereas previous variant callers require alignment files which are missing from Recount3 repository due to data storage constraints. We have successfully genotyped on average 932,460 SNPs per bulk and 90,863 per single-cell RNA-seq samples and completed the largest RNA-seq repository with matched genotype information to date. Our model had a mean overall prediction accuracy of 99.5% in our independent test sets, including an out-of-study test sets which strongly supports its generalizability across SRA. The model is robust to variations in tissue expression patterns. The accuracy is mainly affected by SNP coverage and allele frequency. Work on sharing this resource is ongoing as it contains private genotype information. To demonstrate that the predicted genotype contains biological signals, we sought to unravel the population composition of publicly deposited RNA-seq. Using 1000 Genome population as a reference, we used principal component analysis coupled with k-means clustering to predict the super-population of a sample. Focusing on samples with an estimated accuracy higher than 70%, the publicly available bulk RNA-seq samples consists of 38% Admixed American, 35.9% European, 12.6% African, 11.7% East Asian, and 1.7% South Asian. This reinforces concerns that despite massive efforts to diversify the genomic data, the European super-population is still overrepresented in publicly deposited data.

Title: Context-specific chromatin accessibility in single human neurons expands the repertoire of regulatory neuropsychiatric risk variants.

Authors:

S. Zhang[1], C. Li[2], H. Zhang[1], X. Su[3], L. Liang[3], A. Kozlova[1], A. Sanders[1], Z. Pang[4], X. He[3], J. Duan[1]; [1]NorthShore Univ. Hlth.System, Evanston, IL, [2]Perelman Sch. of Med., Univ. of Pennsylvania, Philadelphia, PA, [3]The Univ. of Chicago, Chicago, IL, [4]Rutgers Robert Wood Johnson Med. Sch., New Brunswick, NJ

Abstract:

Genome-wide association studies (GWAS) of neuropsychiatric disorders have identified a plethora of risk loci. However, functional interpretation of possible causal variants/genes remains challenging, especially in the context of interaction between polygenic risk effects and environmental perturbation. We have recently shown that many neuropsychiatric GWAS risk variants affect chromatin accessibility (i.e., allele-specific open chromatin or ASoC) and gene expression in human neurons (Zhang *et al.*, *Science* 2020). We hypothesize that many neuropsychiatric risk variants may alter cell type-specific chromatin accessibility preferentially in activated/stimulated neurons. Here, we modelled neural activation by potassium chloride (KCl) stimulation in co-cultured excitatory and inhibitory neurons of ~100 human induced pluripotent stem cell (hiPSC) lines, followed by assaying neural activity-dependent single-cell multiomes (scRNA/ATAC-seq) and by mapping activity-specific ASoC. Currently, we have analyzed scATAC-seq data for 58 out of 100 hiPSC lines at baseline, 1hr, and 6hrs post-stimulation and identified three main neuronal subtypes (GABAergic inhibitory, NEFM+/NEFM- excitatory) within our neural co-culture. Peak calling by MACS2 identified 188-300K peaks per cell type at each time point, with 26-51% of peaks showing altered chromatin accessibility and with more peaks gaining accessibility. Interestingly, we found that compared to the static (unchanged) peaks, the dynamic peaks (i.e., up or downregulated) showed stronger heritability enrichment of schizophrenia (SZ) and other neuropsychiatric disorders. To identify specific neuropsychiatric risk variants that show neural stimulation-specific ASoC, we performed ASoC analysis of scATAC-seq data in each neural subtype, and our preliminary analysis of 18 hiPSC lines identified 2,392-7,271 ASoC SNPs per cell type/time point, with 30-70% more ASoC SNPs upon neural stimulation. More importantly, we found that many neuropsychiatric risk variants only showed ASoC upon stimulation; for instance, of the 34 SZ risk loci with GWAS risk SNPs (n=46) showing ASoC across all cell types/time points, 16 loci showed ASoC only upon stimulation. Moreover, 20~30% of SZ risk ASoC SNPs were brain or hiPSC-neural eQTLs (1.5-2.2-fold enrichment), implying their possible effects on neural gene expression. Our study provides novel mechanistic insights into how neuropsychiatric risk variants act to confer disease risk, expanding the repertoire of regulatory neuropsychiatric risk variants that may affect chromatin accessibility and gene expression only in activated/stimulated neurons.

Session 070: Sequencing, proceed with caution

Location: Conv Ctr/Room 202A/Level 2

Session Time: Friday, November 3, 2023, 1:45 pm - 2:45 pm

Title: Elusive pathogenic copy number variants detected by noncoding and off-target NGS reads in Mendelian conditions and in cancer

Authors:

C. Rivolta, M. Quinodoz; Univ. of Basel, Basel, Switzerland

Abstract:

Copy-number variants (CNVs) play a substantial role in the molecular pathogenesis of many hereditary diseases, in cancer, as well as in normal human interindividual variation. Yet, they are still rather difficult to identify in mainstream sequencing projects, especially involving Whole-Exome Sequencing (WES), because they tend to occur in DNA regions that are not routinely chosen for sequence analysis. We discovered that DNA reads from uncaptured sequences, present as contaminants in WES procedures, could be used to identify elusive pathogenic CNVs, especially when ‘off-target’ reads can be mapped back to the noncoding regions of the genome. We therefore developed OFF-PEAK, a user-friendly tool that harnesses the enormous, although neglected power of such reads and tested it on sequencing data from 96 cancer samples, as well as 130 exomes of individuals with Mendelian diseases. For both sets of data, OFF-PEAK identified an unexpectedly large number of novel and bona-fide pathogenic CNVs, which were then validated experimentally by standard molecular biology techniques, such as for instance MLPA. Specifically, our tool demonstrated a very high performance (97% sensitivity and 87% specificity), when compared to values obtained experimentally in the wet lab. In conclusion, our data show that CNVs represent an underestimated element of pathogenicity in both Mendelian conditions and in cancer, and that such events can be easily identified by the analysis of off-target reads from NGS procedures.

Title: Eight years and 1.8 million variants later: A historical exploration of variant reclassification rates to evaluate the ACMG/AMP variant classification framework and understand the major contributors to resolving Variants of Uncertain Significance

Authors:

Y. Kobayashi[1], E. Chen[1], D. Swartzlander[1], F. Facio[2], B. Johnson[3], S. Aradhya[4]; [1]Invitae, San Francisco, CA, [2]Invitae, Apo, NY, [3]Invitae, Fort Lauderdale, FL, [4]Invitae, Mountain View, CA

Abstract:

Variant classification (VC) is a process by which currently available evidence is evaluated to determine the certainty of pathogenicity assertion for variants observed during clinical genetic testing for hereditary diseases. In 2015, ACMG/AMP published recommendations that established the widely adopted 5-tier classification system (pathogenic [P], likely pathogenic [LP], uncertain [VUS], likely benign [LB], and benign [B]) and standardized the evidence evaluation framework. While this guideline had a stated target of 90% level of certainty for LB and LP classifications, the lack of quantitative methods for evaluating VC accuracy meant that the framework was largely based on heuristics and the resulting classifications were qualitative in nature.

Because VCs are updated as new data becomes available, review of historical reclassification rates offers an opportunity to assess the accuracy of the original classifications. We explored data of ~1.8 million variants observed in individuals referred for testing at Invitae over an 8 year period. In total, 63,327 variants (3.5%) had been reclassified. Most of these reclassifications represented VUS being resolved to one of the other four classifications (n=56,998; 90.0%). In addition, 5,547 (8.8%) were cases where new data corroborated the original “likely” classifications (LP to P or LB to B) while 734 (1.2%) were cases where new data increased uncertainty (P to LP/VUS, LP or LB to VUS, or B to LB/VUS). Just 48 cases saw new data reverse the original classification (13 P/LP to B/LB and 35 B/LB to P/LP). Looking specifically at the "likely" classifications, 0.65% of 24,851 LP have been reclassified to VUS/LB/B, and 0.031% of 581,998 LB have been reclassified to VUS/LP/P. Thus, the level of certainty appears to far out-perform the ACMG/AMP-targeted 90% threshold.

Of the 959,159 variants originally classified as VUS, 5.9% have been resolved over time, including 29.2% of VUS from 8 years ago. This VUS resolution rate varied based on clinical area (5.4% ~ 20.0%), including 7.5% of VUS in hereditary cancer genes and 10.4% for cardiology. The sources of the data that enabled VUS resolution were wide-ranging, including new data resources (e.g. gnomAD), new tools (e.g. SpliceAI), and testing of family members (e.g. segregation or de novo analysis) with differing degrees of impact based on clinical area.

As the medical genetics community progresses toward a quantitative, probabilistic VC framework, this type of review of historical reclassification data gives us a better understanding of the current state of VC and can serve as a guide to developing more effective and efficient strategies to reach definitive classifications.

Title: Simplifying Clinical Genetic Testing: The Impact of HiFi Sequencing on Diagnoses and Variant Discovery

Authors:

E. Farrow[1], I. Thiffault[2], A. Cohen[3], T. Zion[4], A. Walter[4], M. Gibson[5], C. Bi[4], W. Cheung[6], A. Johnson[4], Y. Zhou[7], G. Bourque[8], J. Johnston[4], C. Schwendinger-Schreck[4], T. Pastinen[9]; [1]Children's Mercy Hosp, Kansas City, MO, [2]Children s Mercy, Kansas City, MO, [3]Children s Mercy Res. Inst., Kansas City, MO, [4]Children's Mercy Hosp., Kansas City, MO, [5]Children's Mercy Hosp., Kansas City, MS, [6]Children s Mercy, Kansas City, KS, [7]McGill Univ., Montreal, QC, Canada, [8]Mc Gill Univ, Montreal, QC, Canada, [9]Children s Mercy Hosp. and Clinics, Kansas City, MO

Abstract:

Despite continued advances in sequencing, clinical genetic testing currently relies on multiple testing modalities to access known genomic variation, complicating integrated clinical reporting, and often necessitating the utilization of multiple reference laboratories. Leveraging HiFi long read genome sequencing (lrGS), a single test can evaluate methylation, copy number variation, structural variation, expansion disorders, X-activation studies, telomere length, and single nucleotide variation with haplotype phasing. To evaluate the impact of a single assay we assess the first 1300 HiFi genomes in a rare disease cohort, systematically generated by an institution-wide research program, Genomic Answers for Kids (GA4K). Analyses included copy number, structural variation, single nucleotide variation, repeat expansion, and for a subset of genomes 5-methyl C detection and X-inactivation. Each layer of the lrGS calling pipeline has been demonstrated to reveal new, previously undiagnosed pathogenic variation. LrGS sensitivity and specificity for SNVs were slightly higher than short read genome sequencing (srGS), at 99% with an estimated net gain of 3% detectability for ClinVar pathogenic variants. Exploratory analyses highlight gains in variation not detected with short read genome sequencing: sequencing 900,000 tandem repeat loci (STRs) in 200 trios allowed us to delineate high de novo variation rate (~1*10e-4) among well genotyped STRs. Improved copy number variant detection, limited to an allele frequency <5% that impact a coding region results in ~18 variants/genome (4 in OMIM genes). Direct 5-methyl-C detection (5mC-HiFi-GS) has been completed in 380 genomes and focusing on rare (< 0.5% population frequency) gene proximal (5') hypermethylation suggestive of "promoter silencing," we observed on average 51 such alleles per patient (13 in OMIM genes). Improved and comprehensive algorithms continue to lead to new diagnoses, including deletions encompassing *SNRPN* and *PMP22* in previously undiagnosed patients with atypical disease presentations. Altogether, the implementation of lrGS in an ES/GS negative cohort results in an ~10% increase in diagnostic yield. Importantly, known variants were recapitulated, indicating lrGS could be utilized as a first-tier genome test, simplifying genetic testing algorithms. In addition, novel capabilities of lrGS augment rare disease analyses beyond increasing current diagnostic rate: discovery of non-coding rare variation leading to hypermethylation, personal assemblies augmenting rare structural variant calling and high-resolution analyses of instability of STRs in candidate disease loci.

Title: A high-throughput approach to the MHC assembly challenge in disease genetics

Authors:

K. Wade[1], K. Kizer[1], J. Williams[1], J. A. Boquett[1], R. Suseno[1], N. Pollock[2], T. Wang[3], S. Caillier[1], A. Renschen[1], A. Santaniello[1], D. Sayer[3], P. Norman[2], D. G. Augusto[4], J. R. Oksenberg[1], J. Hollenbach[1]; [1]Univ. of California, San Francisco, San Francisco, CA, [2]Univ. of Colorado, Aurora, CO, [3]Thermo Fisher Scientific, Los Angeles, CA, [4]UNC Charlotte, Charlotte, NC

Abstract:

The major histocompatibility (MHC) locus is a ~5Mb region of the genome that houses the highly polymorphic human leukocyte antigen (HLA) immune response genes, in addition to many other genes. It also contains the greatest number of disease association signals in the entire human genome. However, the region's hallmark characteristics, such as extensive haplotype diversity, variant heterogeneity, structural variation, copy number variation, repetitive content, extreme linkage disequilibrium blocks and balancing selection, have made interpretation and localization of disease association signals extremely challenging. This complexity is further compounded by the unique methodological hurdles of assembling such highly variable genomic sequences. Approaches which rely on alignment to a reference genome can leave sequence data partially, or improperly, aligned, creating a reference-bias, causing novel variation to be missed. Short read sequencing is especially vulnerable to this issue, and while long read sequencing and genome graphs offer much potential, they are not yet widely applicable for large studies. To make the full, 5Mb MHC accessible for future disease studies, as well as the hundreds of thousands of short read disease cohorts already sequenced, we have developed an MHC-specific, short read de novo assembly method, based on an existing pipeline. We validate this method by re-creating six thoroughly characterized, extended MHC haplotypes from haploid cell lines using their original short read data. We find that the error rate for homozygous sequence is <0.25% for single nucleotide variants (SNVs) and small to medium-sized indels and that we can reproduce known, haplotype-associated structural variation in the HLA Class II region. To handle the diploid, heterozygosity observed in human populations, our method contains additional features for identifying heterozygous variation and larger structural rearrangements. For regions with ~60X coverage, we generate MHC assemblies with average contig number, N50 value and assembly length of 979, 10.96kb and 4.3Mb, respectively, at a speed of ~1 ½ hours per MHC haplotype. Finally, we integrate divergent haplotype sequences into a single, concordant coordinate system, for use in disease association. We apply this method to a human multiple sclerosis case-control cohort (n=2,214) sequenced with targeted, 60X Illumina paired-end short read sequencing, as well as the 1000 Genomes Project Phase 3 WGS individuals. By facilitating for the first time accurate high-resolution analysis of MHC sequence, this method has the potential to transform our understanding of the role of the MHC in human disease.

Session 071: The mind's code: Unraveling neuropsychiatric disorders through genetics

Location: Conv Ctr/Ballroom B/Level 3

Session Time: Friday, November 3, 2023, 1:45 pm - 2:45 pm

Title: Genome-wide meta-analyses of cross substance use disorders in European, African, and Latino ancestry populations

Authors:

D. Lai[1], M. Zhang[1], M. Abreu[1], T-H. Schwantes-An[1], C. Parker[2], F. Jin[3], H. J. Edenberg[4], Y. Liu[1], T. Foroud[5]; [1]Indiana Univ. Sch. of Med., Indianapolis, IN, [2]Middlebury Coll., Middlebury, VT, [3]Case Western Reserve Univ., Cleveland, OH, [4]Indiana Univ Sch. of Med., Indianapolis, IN, [5]Indiana Univ Sch Med, Indianapolis, IN

Abstract:

Substance use disorders (SUDs, including alcohol (AUD), cannabis (CUD), opioid (OUD), and nicotine (NUD), etc.) have devasting consequences on individuals, families, and societies, and cause ~5.5% of disability-adjusted life-years globally. Twin and genetic correlation studies have demonstrated the existence of common cross-SUDs genes, but previous cross-SUDs GWAS only identified a few dozen genes. We performed the largest cross-SUDs meta-analyses in European (EA: including AUD, CUD, OUD, NUD, and substance abuse; N=251,334-435,563, total effective same size (Neff)=1,467,929), African (AA: including AUD, CUD, OUD, and NUD; N=9,745-91,026, Neff=159,000), and Latino (LA: including AUD (N=14,175) and OUD (N=34,861); total Neff=45,757) ancestry populations. Cross-ancestry meta-analyses were also performed. Because SUDs are often comorbid and are genetically correlated, we retained SNPs having concordant effects across SUDs. We identified 40, 1, 3, 20, 24, 12 loci in EA, AA, LA, AA+EA, EA+LA, AA+EA+LA, respectively. These loci include 2,196 genome-wide significant (GWS) SNPs and 104 genes having at least one GWS SNP. We found three missense GWS SNPs (rs11601425 in *TMPRSS5*, rs2287922 in *RASIP1*, and rs3736781 in *BTN1A1*) and two missense SNPs (rs6720 in *MDH2*, rs61785974 in *PHACTR4*) that are in high linkage disequilibrium (R^2>0.81) with GWS SNPs, and these missense SNPs were not reported as SUD-related previously. We also used eQTL and chromatin interaction mapping to link GWS SNPs to genes and performed gene-based analyses. In total, we identified 299 genes. Among them, 250 genes were associated with either psychiatric traits, brain measurements, and brain functions, or directly interact with genes in SUD-related pathways derived from model organism studies. The estimated cross-SUDs SNP heritability in EA was 0.10 (SD=0.001). Cross-SUDs SNPs explained 46.59%, 53.01%, 59.07%, and 50.75% of SNP heritability of AUD, OUD, CUD and NUD, respectively. Genetic correlations among SUDs ranged from 0.48 to 0.70 (Ps<5.20E-12). SUDs were positively correlated with psychiatric disorders (genetic correlations>0.33, Ps<3.95E-05). Cross-SUDs polygenic risk scores were significantly associated with SUDs in All of Us samples (EA: P=4.24E-58; AA: P=0.02) and EA (P=1.86E-07) but not AA (P=0.36) samples from Indiana Biobank. We identified 248 drugs targeting 29 genes that can potentially be repurposed to treat SUDs. Our findings can help elucidate the etiologies of SUDs and develop novel prevention and treatment methods.

Title: Deciphering human brain cellular crosstalk in the context of neuropsychiatric disease.

Authors:

C. Porras[1], D. Lee[1], R. Kosoy[1], M. Pjanic[1], J. Bendl[1], P. Fnu[1], K. Therrien[1], D. Mathur[1], S. P. Kleopoulos[1], Z. Shao[1], M. Alvia[1], C. Casey[1], A. Hong[1], S. Argyriou[1], K. G. Beaumont[1], R. Sebra[1], P. Auluck[2], S. Marenco[2], C. Kellner[1], D. Bennett[3], G. Voloudakis[1], V. Haroutunian[1,4], G. Hoffman[1], J. Fullard[1], P. Roussos[1]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Natl. Inst. of Mental Hlth.-Intramural Res. Program, Bethesda, MD, [3]Rush Univ. Med. Ctr., Chicago, IL, [4]James J. Peters VA Med. Ctr., Bronx, NY

Abstract:

Comparative analyses of Alzheimer's disease (**AD**), Schizophrenia (**SCZ**) and Bipolar disorder (**BD**) suggest neuropsychiatric disease manifestations are mediated by immune and nervous system interactions. Pro-inflammatory mediators released from neuroimmune cells have been shown to affect neurotransmitter levels and alter synaptic connectivity. These changes lead to many common symptoms including cognitive deficits, hallucinations, and depressed mood. However, disease-specific interactions between the immune and nervous system are incompletely understood. Further investigation is challenging due to limited availability of human tissues. Here, we establish one of the largest single-nucleus transcriptomic sequencing cohorts for AD, SCZ and BD, named **PsychAD**, consisting of 1,495 donors across three brain banks. Out of ~8 million nuclei we identify 28 unique subclasses and 67 subtypes of neuronal, glial and immune cells. We identify cell-cell interaction (**CCI**) networks important for disease progression and cross-validate our results between samples from different brain banks. We additionally validate with an independent cohort named **FreshMG**, which consists of fresh human brain single-cell transcriptomic sequencing on 137 unique donors and ~1 million neuroimmune cells. Our overall hypothesis is that specific neuroimmune and neuronal subtypes form cell-cell interactions that influence neuropsychiatric disease pathology through the release of immune mediators. In our investigation, we leverage established CCI inference methods coupled with unsupervised tensor factorization to capture CCI patterns. We identify CCI in AD, SCZ and BD with cell-type specificity at an unprecedentedly high resolution. We estimate heritability of these CCI and replicate our findings across both cohorts. Notably, we demonstrate that microglia HMGB1 ligand is associated with AD progression by acting on the brain perivascular macrophage CD163 receptor. Modulation of this interaction has the potential for therapeutic benefit in individuals with AD. Our results decipher CCI patterns in neuropsychiatric disease and identify targets for pharmacological intervention.

Title: Discovering genes linked to both cognition and psychiatric disorders through analysis of 888,052 exomes

Authors:

V. Rajagopal[1], A. Ayer[1], R. Wang[1], N. Banerjee[2], A. J. Averitt[3], A. Moscati[4], Regeneron Genetics Center, REGN-DiscovEHR Collaboration, Mayo Clinic Project Generation, M. Cantor[1], A. Baras[1], E. Stahl[1], G. Coppola[1]; [1]Regeneron Genetics Ctr., Tarrytown, NY, [2]Regeneron, Tarrytown, NY, [3]Regeneron Genetics Ctr., New York, NY, [4]Regeneron Pharmaceuticals, Tarrytown, NY

Abstract:

Background: The phenotypic spectrum of rare deleterious variants in neurodevelopmental genes is a continuum, extending from severe neurodevelopmental disorders to mild or moderate cognitive dysfunction. Hence, large-scale cognitive phenotypes measured in the general population can serve as endophenotypes to discover risk genes for psychiatric and neurodevelopmental disorders.
Methods: We performed an exome-wide association study of cognitive (fluid intelligence, reaction time, and memory) and socioeconomic phenotypes (educational attainment and household income) in the UK Biobank (N=76,360 to 449,878). Using principal component analysis, we derived the *g* factor and used it to improve the power for gene discovery via a conditional FDR approach. Significant genes were evaluated further by phenome-wide rare-variant burden analysis of mental and behavioral disorders defined by ICD10 codes (chapter V: F00-F99; n=80 definitions) in 888,052 individuals from six cohorts. Genetic analyses were performed using REGENIE. We focused on the burden associations of predicted loss of function (pLoF) and/or likely deleterious missense variants either at the gene or gene-set level at different allele frequencies.
Results: We identified 31 significant genes in which rare-variant burden was associated with the (i) g factor and (ii) one or more cognitive and socioeconomic phenotypes. These 31 genes were enriched for associations with autism and neurodevelopmental disorders (11.9 folds; P=4.1e-10) and the strongest associations include known neurodevelopmental genes, such as *GIGYF1* and *KDM5B*. Analyzing ~900k exomes across six cohorts, we found that pLoFs and deleterious missense variants in the 31 genes in aggregate were associated with a variety of psychiatric disorders, with the strongest associations including intellectual disability and autism-related disorders. At the gene level, we identified significant burden associations for seven genes with 10 disorder definitions which include both known and novel associations
Discussion: Rare variant discoveries for psychiatric disorders have been so far mostly restricted to highly penetrant variants associated with diseases at the severe end of the phenotypic spectrum, and many of the risk genes with mild to moderate penetrance are yet to be identified. Our findings demonstrate the strong genetic link between cognition and mental and behavioral disorders and emphasize the importance of leveraging it to make novel gene discoveries.

Title: WGSPD: The largest deep whole genome sequencing cohort of bipolar and schizophrenia patients identifies novel genetic insights

Authors:

R. Ye[1], C. Liao[1], M. Talkowski[2], B. Neale[2], Whole Genome Sequencing for Psychiatric Disorders(WGSPD) consortium; [1]Broad Inst., Cambridge, MA, [2]Massachusetts Gen. Hosp., Boston, MA

Abstract:

Introduction: The etiologies of bipolar disorder (BD) and schizophrenia (SCZ) are well established as having a substantial genetic component. Whole exome sequencing (WES) has shown evidence implicating rare variation, particularly copy number and protein truncating variants (PTVs), in BD and SCZ risk. Whole genome sequencing (WGS) extends rare variation discovery to noncoding regions, including investigating structural variation (SV) not previously captured by WES approaches. Here we report results from the Whole Genome Sequencing for Psychiatric Disorders (WGSPD) consortium - the largest and most ancestrally diverse BP and SCZ WGS cohort to date.
Methods: After sample QC and ancestry inference, our study cohort consisted of 4534 ancestry-matched controls and 9039 cases with a BP or SCZ diagnosis. We employed an enrichment analysis of ultra-rare PTVs, damaging missense, and deletion variants with minor allele counts less than 5. Finally, we characterized the distributions of structural variation, including deletion, duplication, inversion, translocation, and other complex events.
Results: The WGSPD cohort represents a range of diverse global populations - including African (60.46%), non-Finnish European (22.02%), Finnish (14.83%), American admixed (2.67%), and Asian (0.02%) individuals. In our initial analysis, we found a significant enrichment of PTVs in established schizophrenia associated genes ($OR = 1.54; p = 0.0207$), neurodevelopmental disorder associated genes ($OR = 1.36; p < 1e\text{-}03$), and in high loss-of-function constrained genes ($OR = 1.21; p < 1e\text{-}05$). Additionally, we identified SVs across the allelic spectrum spanning both non-coding and coding regions that may not have been identifiable by previous technologies.
Discussion: We recapitulate expected loss-of-function enrichment among BP and SCZ cases in phenotypically-related gene sets. We further demonstrate the potential of large WGS cohorts to identify novel SVs spanning intronic and exonic regions previously obscured by the limited cohort size and coarse resolution of preceding studies. In conjunction with the diversity of populations represented in the WGSPD cohort, these findings provide insights into the complex etiologies of BP and SCZ, as well as the ability for WGS to resolve noncoding and coding effects.

Session 072: Unraveling the genetic determinants of female infertility

Location: Conv Ctr/Room 145A/Level 1

Session Time: Friday, November 3, 2023, 1:45 pm - 2:45 pm

Title: Elucidating the genetic architecture of polycystic ovary syndrome (PCOS)

Authors:

C. Parker[1], R. Bauer[1], C. Welt[2], R. Legro[3], G. Hayes[1], M. Urbanek[1]; [1]Northwestern Univ., Chicago, IL, [2]Univ. of Utah, Salt Lake City, UT, [3]MS Hershey Med. Ctr, Hershey, PA

Abstract:

Background: Polycystic ovary syndrome (PCOS) is a common multisystem endocrine disorder affecting 6-15% of women in reproductive age and is a major cause of anovulatory infertility. PCOS is a heterogeneous disorder with multiple sub-phenotypes modulated by both environmental and genetic factors. PCOS is diagnosed by the occurrence of hyperandrogenemia, oligo-anovulatory, and/or polycystic ovaries. A study conducted in twins concluded that PCOS is highly heritable (h2=0.72); additionally, first-degree relatives of individuals with PCOS have an increased risk of several sub-phenotypes of PCOS.
Objective: Our goal is to determine which genes and pathways directly impact the etiology of PCOS with the intent to better understand PCOS sub-phenotypes and promote more targeted treatment.
Methodology: We are performing whole exome sequencing (WES) on 800 individuals with PCOS and 300 age/ancestry matched controls to generate a comprehensive catalog of PCOS-associated protein-coding variants. We will use prediction tools to identify likely to be deleterious rare (MAF<1%) and/or low frequency (MAF<5%) variants and in silico evidence of epigenetic modification to identify PCOS-associated genes/pathways. We expect the variants identified by WES to map to multiple genes and pathways, including novel genes/pathways as well as genes/pathways that have previously been predicted to play a major role in the etiology of PCOS.
Results: To date, we have obtained high quality sequencing data on 676 PCOS cases and 85 controls; sequencing of the remaining 300 samples is in progress. We have identified 88 rare, protein-coding genetic variants using targeted re-sequencing in three distinct pathways in an independent PCOS cohort. These genes are associated with insulin resistance, the gonadotropin signaling pathway, and the Anti-Müllerian Hormone (AMH) signaling pathway. We identified 37 functionally validated deleterious variants in genes that encode AMH or its cognate receptor AMHR2, 12 rare variants that encode LMNA, 12 rare variants that encode the insulin receptor, and 27 coding variants for the genes encoding the beta subunit of gonadotropins, LH and FSH, and their receptors, FSHR and LHCGR.
Conclusion: We have identified 88 rare, protein-altering genetic variants in pathways that are predicted to be impaired in individuals with PCOS supporting a critical role for genetics in the etiology of PCOS. Our WES screen will allow us to generate a comprehensive catalog of likely to be deleterious PCOS-associated variants. This catalog of variants will identify the genes/pathways that are mutated in PCOS, leading to improved personalized phenotyping and treatment of PCOS.

Title: Mapping loci associated with impaired female reproduction

Authors:

S. Ruotsalainen[1], J. Karjalainen[2], M. Kurki[2], H. Laivuori[3], H. Heyne[4], E. Lahtela[3], FinnGen, M. Daly[2], A. Palotie[5], E. Widen[3]; [1]Inst. for Molecular Med. Finland, FIMM, Helsinki, Finland, [2]Broad Inst., Cambridge, MA, [3]Univ. of Helsinki, Helsinki, Finland, [4]Hasso Plattner Inst., Potsdam, Germany, [5]Inst. for Molecular Med. Finland FIMM, Helsinki, Finland

Abstract:

Female infertility is a complex and common health problem. Approximately 1/7 of couples are affected by it, and approximately 12% of women have received fertility treatment. The underlying etiology is complex and includes disorders of the ovaries, uterine disorders, or congenital reproductive tract malformations. Despite standard fertility investigations, the underlying cause remains unexplained in up to 15-30% of cases. In addition, the implicated pathogenic mechanisms causing the reduction of fertility remain poorly understood, despite a clinically recognized association between infertility and some diseases of the female reproductive tract.
We utilized the large Finnish research project FinnGen to perform the largest GWAS of female infertility to date. Our analyses included 20 409 cases with a medical record of infertility and 180 155 females who have given birth as controls. Applying both additive and recessive analysis models we identified 6 independent genome-wide significant (GWS) loci, out of which only 3 (near genes *WNT4*, *SULT1B1*, and *ESR1*) have been previously reported to associate with diseases of female reproductive organs, such as endometriosis and PCOS, while 2 loci (near genes *EBAG9* and *TBPL2*) are uniquely associated with female infertility, and show a clear recessive mode of inheritance. In addition, we see GWS association in the MHC region, which also seems to have a recessive mode of inheritance.
In both infertility-specific loci (*EBAG9* and *TBPL2*), a Finnish-enriched coding variant was present in a 95% credible set. The lead variant at *TBPL2* is a highly Finnish enriched (43-fold compared to non-Finnish Europeans) stop-gained variant rs144313315 ($p_{rec} = 3.38 \times 10^{-25}$, $p_{add} = 0.1$). This variant was also associated with a reduced number of offspring in females: the average number of offspring of homozygote females was 0.2 compared to 1.75 in heterozygotes or wild-type homozygotes, $p = 3.05 \times 10^{-15}$). Further highlighting the impact of this variant, none of the homozygote females had conceived without years of infertility treatment. Previously, female mice lacking *TBPL2* have been identified to be completely sterile due to anovulation.
In conclusion, we identified 6 novel loci for female infertility, with a broad allelic spectrum. About half of them have previously been associated with syndromes in which infertility is one of the symptoms, while a substantial proportion of the genetic predisposition was unique to female infertility, and not associated with any other diseases. Interestingly, many of the signals seem to have clearly recessive mode of inheritance.

Title: The distinct genetic determinants of infertility and reproductive hormones

Authors:

S. Venkatesh[1], L. Wittemans[1], M. Parker[1], B. Hill[1], N. Baya[1], D. Palmer[1], T. Ferreira[1], B. M. Jacobs[2], J. Figueredo[3], M. Karjalainen[4], A. Pasanen[4], A. Elhakeem[5], P. Rohde[6], G. Thorleifsson[7], FinnGen, Genes & Health Research Team, DBDS Genomics Consortium, V. Steinthorsdottir[7], K. Stefansson[7], M. Nyegaard[6], D. Lawlor[5], N. Timpson[5], H. Laivuori[8], T. Laisk[3], D. van heel[2], C. Lindgren[1]; [1]Univ. of Oxford, Oxford, United Kingdom, [2]Queen Mary Univ. of London, London, United Kingdom, [3]Univ. of Tartu, Tartu, Estonia, [4]Univ. of Oulu, Oulu, Finland, [5]Univ. of Bristol, Bristol, United Kingdom, [6]Aarhus Univ., Aarhus, Denmark, [7]deCODE Genetics, Reykjavik, Iceland, [8]Univ. of Helsinki, Helsinki, Finland

Abstract:

Reproductive hormones regulate physiological processes such as sexual development and the reproductive cycle, and are implicated in a range of disorders, including infertility. However, the mechanisms driving infertility & hormonal imbalances remain poorly understood.

To elucidate the genetic factors driving reproductive hormone levels and infertility, we conducted the first large-scale genome-wide association study (GWAS) meta-analyses, involving up to 690,631 participants from six biobanks across ancestral groups. We assessed 29 million genetic variants for associations with female infertility of all causes (40,070 cases/553,362 controls), anatomical infertility (7,994/315,490), anovulatory infertility (5,233/433,102), & idiopathic female infertility (up to 26,182/664,449), as well as male infertility (7,038/415,114); we additionally performed the largest to date sex-specific GWAS meta-analyses of follicle stimulating hormone ($N_{females}$ = 34,852; N_{males} = 7,133), luteinizing hormone (29,182; 5,982), oestradiol (64,874; 21,682), progesterone (7,754; 1,420), & testosterone (217,359; 215,858) levels. We further analysed exome sequencing data from up to 350,000 individuals to identify rare coding variation associated with infertility and hormone levels.

We discover nine previously unreported loci associated with female infertility (P < 5E-08), including *RYR3*, *PKHD1L1*, & *ESR1*; as well as two novel loci, *NCKAP5* & *SUPT3H*, associated with male infertility. In addition, we report 28 novel loci for reproductive hormones, including two genes in steroid hormone biosynthesis pathways and six loci associated with age at menarche or menopause. By leveraging publicly available ovary single-cell gene expression databases, we prioritised granulosa cells as enriched (FDR < 0.05) for the heritability of idiopathic female infertility, and various ovarian immune & endothelial cell types enriched for the heritability of the different female reproductive hormones.

While 52% of loci with rare variant signals for hormone levels also harbour common variant associations, we identify seven genes with rare variants impacting testosterone levels which have not been implicated in GWASs. These include *HSD17B2* (steroid hormone oxidation) & *GPR137* (regulation of testosterone secretion).

We do not find evidence for significant genome-wide genetic correlation between any reproductive hormone and any definition of infertility, nor do we find any genetically predicted causal effects of hormone levels on infertility (all FDR > 0.05).

Taken together, our work highlights distinct genetic drivers of reproductive hormones and infertility.

Title: Identifying gene regulatory mechanisms that explain PCOS GWAS signals

Authors:

L. Sankaranarayanan[1], K. Brewer[2], R. Sisk[3], A. Barrera[1], C. Li[4], A. Dunaif[2], T. Reddy[1]; [1]Duke Univ., Durham, NC, [2]Icahn Sch. of Med. at Mount Sinai, New York, NY, [3]Northwestern Univ., CHicago, IL, [4]Univ. of Southern California, Los Angeles, CA

Abstract:

Most genetic variants identified in genome-wide association studies (GWAS) are in noncoding regions of the genome. Understanding the molecular mechanisms by which these variants contribute to complex traits remains a major challenge and opportunity in human genetics. We report the discovery of potential gene regulatory mechanisms that explain noncoding variants associated with PCOS in the *GATA4*, *FSHB* and *DENND1A* loci. We developed a novel approach combining genetic association analysis with results from high throughput reporter assays of the genomes of PCOS case and control populations.

PCOS is a leading cause of infertility affecting up to 15% of premenopausal people. PCOS is characterized by increased testosterone levels. Multiethnic GWAS have mapped ~20 PCOS susceptibility loci implicating neuroendocrine, reproductive, and metabolic pathways. Elucidating the functional consequences of noncoding genetic variation in these loci will provide insight into how these pathways are disrupted in PCOS.

We identified ~1100 candidate regulatory elements in two steroidogenic cell lines across 14 PCOS GWAS loci. These elements included cell type-specific and shared regulatory elements. We used high throughput reporter assays (STARR-Seq) and CRISPR-based epigenetic manipulations to confirm effects on gene regulation. Within these 1100 regulatory elements, we identified novel genetic variants with allele specific regulatory activity or PCOS-association. We identified several genetic variants in the *GATA4* and *FSHB* loci significantly associated with PCOS within those regulatory elements. We identified 22 genetic variants with allele-specific regulatory activity in the *DENND1A* locus using STARR-Seq of five multiethnic genomes. We further show changes in endogenous *DENND1A* expression altered androgen biosynthesis in CRIPSR-modified cell models. We specifically perturbed the promoter and three identified regulatory elements using CRISPR activation (dCas9-p300). We observed the increased *DENND1A* expression levels led to increased testosterone produced by the cell line. *FSHB* codes the β-subunit of follicle stimulating hormone, a key hormone involved in granulosa cell steroidogenesis. *GATA4* codes a transcription factor involved in embryogenesis and testicular development. *DENND1A* regulates theca cell androgen synthesis.

Taken together, we highlight a novel approach combining genetic variant analyses with experimental approaches to fine-map noncoding GWAS signals and identify potential gene regulatory mechanisms that contribute to PCOS.

Session 096: Advancements in genome sequencing: Unraveling genetic factors in human health, disease, and phenotypic diversity

Location: Conv Ctr/Room 202A/Level 2

Session Time: Saturday, November 4, 2023, 10:30 am - 12:00 noon

Title: Assessing the contribution of rare untranslated region variants to human diseases using ~500,000 UK Biobank whole-genome sequences

Authors:

K. Kundu[1], F. Hu[1], Q. Wang[2], R. Dhindsa[3], K. R. Smith[1], K. Carss[1], **S. Petrovski**[1]; [1]AstraZeneca, Cambridge, United Kingdom, [2]AstraZeneca, Chapel Hill, NC, [3]Baylor Coll. of Med. and AstraZeneca, Houston, TX

Abstract:

Untranslated regions (UTRs) of protein coding genes play important roles in post-transcriptional gene regulation, including mRNA stability and subcellular localization, and controlling protein levels. Previous studies have indicated that variants within UTRs can be genetic drivers of human diseases. These studies have mostly focussed on common variant associations. The whole-genome sequencing (WGS) data of ~500K UK Biobank participants provides an unprecedented opportunity to explore for the first time at scale the role of rare UTR variants in human diseases, which will uncover disease mechanisms and provide opportunities to develop better therapeutic approaches and precision medicine strategies. Here, we compiled 14.7 million rare variants (minor allele frequency (MAF) < 1%) from both 5' and 3' UTRs of 19,229 protein coding genes from 462,095 participants of European ancestry in the UK Biobank, which constitutes the largest UTR rare variant data set to date. We characterized all the UTR variants based on their MAF and predicted deleteriousness. We then performed collapsing analysis separately for 5' and 3' UTRs with ~15.5K binary phenotypes. For 5' UTRs, we observed a total of 198 significant ($P < 1 \times 10^{-8}$) gene-phenotype associations comprising 67 genes and 124 phenotypes. The 3' UTR variants yielded a total 160 significant associations comprising 77 genes and 90 phenotypes. These included known associations such as *APOA5* and hypercholesterolemia ($P = 1.59 \times 10^{-14}$, OR = 1.29), as well as novel associations that are undergoing further analysis. Overall, our work is by far the largest rare-variant study of UTR variants and will demonstrate how rare variants in UTR regions provide valuable insights into disease biology.

Title: Exome sequencing of 1 million individuals identifies 211 genes associated with adult human height

Authors:

J. Kosmicki, L. Ganel, A. Marcketta, K. Watanabe, D. Sharma, K. Sun, J. Mbatchou, A. Ziyatdinov, T. Joseph, S. Balasubramanian, X. Bai, W. Salerno, M. Jones, J. Reid, Mayo Clinic Project Generation, Penn Medicine Biobank, DiscovEHR, A. Baras, G. Abecasis, J. Marchini, M. Ferreira, A. Locke; Regeneron Genetics Ctr., Tarrytown, NY

Abstract:

Adult human height is the model trait of human genetics and the first for which genome-wide association studies (GWAS) identified nearly all common genetic variants contributing to its heritability. Yet, the contribution of rare (minor allele frequency [MAF]<1%) coding variation in height remains largely unexplored. To address this, we performed the largest exome sequencing study of height in 1,086,942 individuals (joint-called discovery: N=844,029 [24% non-European ancestry]; replication: N=242,913 [29% non-European ancestry]).

Overall, through gene-based tests, we directly implicate rare variation in 211 genes on adult human height (Bonferroni $P<1.75x10^{-9}$), independent of GWAS loci. The rarest and most deleterious variant class, singleton putative loss-of-function variants (pLoFs), significantly impacted height in 17 genes (5 height increasing and 12 decreasing): carriers of height-increasing pLoFs were 9.4cm taller than average, while carriers of height-decreasing pLoFs were 5.6cm shorter than average. The most extreme example was *ACAN*, with a -14.9cm per-singleton pLoF effect. For comparison, individuals in the top 1% of a polygenic score (PGS) for height were 11.8cm taller than average, whereas those in the bottom 1% were 7.6cm shorter than average - indicating singleton pLoFs have roughly the same average effect on height as the 1% tails of the PGS distribution. Additional notable findings include 1) 169 genes not previously implicated in Mendelian forms of height, 2) 47 genes not implicated by GWAS or >1Mb from a GWAS locus, 3) 37 genes identified only by SKAT tests (explained by non-synonymous variants with directionally opposing effects on height) and 4) a positive correlation between stronger gene constraint and larger pLoF effects on height in both directions (height-decreasing genes $P=2x10^{-6}$; height-increasing genes $P=8x10^{-4}$).

Beyond solely identifying genes, we also explored rare variant heritability; in the replication cohorts, rare pLoFs in these 211 genes explained 0.72% of variation in height, compared to ~21% explained by 4017 GWAS loci. When considering all protein-coding genes, we found rare pLoFs and deleterious missense variants explain 1.8% and 1% of the heritability of height, respectively. Overall, our analysis of height in >1 million exomes identified 211 genes and constitutes the most comprehensive description of the rare coding genetic architecture of height to date.

Title: Analysis of >800,000 diverse sequenced humans in gnomAD improves clinical interpretation and provides insight into gene function.

Authors:

K. Chao[1,2], J. K. Goodrich[1,2], K. Laricchia[1,2], M. W. Wilson[1,2], J. M. Fu[1,2], G. Tiao[1,2], Q. He[1,2], D. Marten[1,3], T. Poterba[1,2], C. Vittal[1,2], S. Chen[2,1], W. Lu[1,2], S. Baxter[1], S. Chapman[1,2], C. Cusick[1], P. W. Darnowsky[1,2], L. Gauthier[1], L. Gruenschloss[4], R. Grant[1,2], S. Jahl[1,2], M. Solomonson[1,2], C. Stevens[1,2], gnomAD Project Consortium, D. King[1,2], D. G. MacArthur[4,1], M. E. Talkowski[1,2], B. M. Neale[1,2], A. O'Donnell-Luria[1,2,3], K. E. Samocha[1,2], K. J. Karczewski[1,2], M. J. Daly[1,2,5], H. L. Rehm[1,2]; [1]Broad Inst. of MIT and Harvard, Cambridge, MA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]Boston Children's Hosp., Boston, MA, [4]Garvan Inst. of Med. Res., Sydney, Australia, [5]Inst. for Molecular Med. Finland, Helsinki, Finland

Abstract:

A critical component to the medical and functional interpretation of genetic variants involves the accurate estimation of their frequency across human populations. Here, we present a new release of the Genome Aggregation Database (gnomAD), a publicly available allele frequency resource consisting of 807,069 individuals (730,913 exomes and 76,156 genomes) aligned to GRCh38, and highlight some of the scientific insights gained from a dataset of this scale and diverse human representation.

gnomAD v4 contains 138,536 non-European samples, including 37,485 and 30,013 individuals with African (AFR) and Ameridingenous (AMR) genetic ancestries, respectively. At gnomAD's current scale, we find two variants every three base pairs in the exome and discover 98% of all possible synonymous CpG substitutions. Notably, over 20 million variants (approximately 54%) of the variants seen in v4 were not observed in previous releases, with much of the increase resulting directly from increased representation of diverse genetic ancestry groups.

We expand our methods to improve pathogenic variant classification, calculating a single filtering allele frequency combining exome and genome data. We compute updated metrics for genic constraint (detecting genes under natural selection), with 96% of genes well-powered to detect constraint against predicted loss-of-function (pLoF) variants, and we extend these metrics to estimate mutational intolerance in non-coding regions. We show that diversity is a critical component to functional interpretation, as these scores built from diverse cohorts perform better in discerning genes and genomic regions undergoing natural selection, increasing AUC for identifying non-coding regions with fine-mapped GWAS variants from 0.66 in non-Finnish European samples (NFE) to 0.682 in a multi-ancestry cohort of the same size. Further, over 3,000 genes harbor homozygous pLoF variants (potential human "knock-out" genes), compared to 1,700 genes in v2. We also identified a median of 1 rare coding copy number variant (CNV, <1% site frequency) per person among 84,060 unique rare CNV sites, with rates per gene correlated to constraint metrics. The gnomAD v4 dataset will provide the foundation for better powered aggregate variant co-occurrence and regional missense constraint.

Increasing the scale and diversity of reference datasets results in direct scientific advances in addition to moving the fields towards more equitable science. With better representation, we better understand the landscape of genetic variation, yielding insights into human disease.

Title: Identification of rare *de novo* tandem repeat expansions in congenital heart defects using trio based whole genome sequencing

Authors:

A. Martin Trujillo, B. Jadhav, P. Garg, M. Shadrina, W. Lee, B. Gelb, A. Sharp; Icahn Sch. of Med. at Mount Sinai, New York, NY

Abstract:

The use of exome and genome sequencing in human disease cohorts has identified the genetic basis of numerous Mendelian disorders. However, even after genome sequencing, a substantial fraction of cases often remains unexplained, suggesting the involvement of alternative pathogenic mechanisms.Tandem repeat expansions (TREs) are a class of mutation that, to date, have primarily been associated with late-onset neurological disorders and are typically missed by standard genomic analyses. Given their known pathogenic potential, we hypothesized that TREs might represent the underlying genetic cause for other Mendelian diseases, potentially accounting for a fraction of *unsolved cases*. To investigate this, we analyzed Illumina genome sequencing data from 1,780 trios collected by the Pediatric Cardiac Genomics Consortium (PCGC), consisting of probands with congenital heart defects (CHD) and their unaffected parents. We performed genome-wide screening for expansions of short tandem repeats (STRs) using *STRetch* and *ExpansionHunterDenovo,* followed by targeted genotyping of ~28,000 loci of interest using *ExpansionHunter*. After extensive automated and manual curation, we identified 516 high-confidence rare *de novo* TREs (dnTREs), representing 393 loci where a proband carried an expanded TR allele that was both rare in controls and significantly longer than both parents. Validations were performed for a subset of TREs using long-read sequencing. dnTREs showed a 2.3 fold-enrichment to occur within genes previously implicated in CHD from human and mouse studies ($p<0.05$), providing strong evidence for a causal role in disease. While 77% of the dnTREs were observed uniquely, we also detected recurrent expansions, *i.e.* the same TRE observed in multiple unrelated probands, including several that occurred within genes previously implicated as causal for CHD by exome sequencing, *e.g.* an intronic AG TRE located in *VEGFB* in two unrelated cases.We then performed replication using ~168,000 individuals from the UK Biobank for which WGS data were available. Using ICD codes to identify individuals with likely CHD, we generated genotypes for 393 STRs that showed *de novo* TREs in the PCGC discovery cohort and performed a case-control study for rare outlier TR alleles. 13% of the loci tested showed a significant enrichment ($p<0.05$) for TREs in patients with CHD compared to controls, providing further evidence of their causality.Our observations expand the known pathological effects of TREs beyond neurodegenerative disorders and suggest that TREs represent a mutational mechanism that contribute to a much broader range of human diseases than is currently recognized.

Title: Insights from an exome sequencing study on ethnically diverse cohort of 18,994 patients with suspected rare Mendelian disorders.

Authors:

H. Lee, H. Han, S-I. Hyun, K. Kwon, S. Ryu, R. Khang, E. Lee, J. Kim, Y. Song, W. Jeong, J. Han, D-W. Kim, H. Kim, J. Woo, K. Lee, D. An, S. Yang, S. Lee, S. Jang, S. Kim, D. Jeon, J. Lee, G. Seo; 3billion, Seoul, Korea, Republic of

Abstract:

Over the last decade, exome/genome sequencing (ES/GS) has settled down as a routine clinical test for diagnosing patients with suspected rare Mendelian disorders. There are also multiple large-scale ES/GS studies being carried out. However, these are limited to certain countries and many others are still at earlier stages of exploring how to incorporate ES/GS in their clinical practices and design large population studies. Here, we report on 18,994 patients referred from 50 countries between 2020-2022 to a CAP/CLIA-accredited reference laboratory located in East Asia for ES. ES was performed following the standard protocol.
The diagnostic statistics were comparable to previously reported large-scale exome sequencing studies not selected for specific phenotype or ethnic group. Overall, 30.4% of the patients received a molecular diagnosis with P/LP variants. Additional 12.7% of the patients received an inconclusive result with either VUS in AD genes or at least one heterozygous P/LP variant in AR genes. The diagnostic rate (DR) varied (3%-49%) depending on the disease category and the neurodevelopmental delay group, the largest group with 4,208 patients had a DR of 38%. Early-onset (≦12 yo) disorders were significantly correlated with higher DR of 36.1% than late-onset (>12 yo) disorders (20.1%, $p<0.00001$).
The largest number of patients were referred from South Korea (33%), Malaysia (18%) and Egypt (11%). Ethnicity was predicted for each sample based on gnomAD's variant frequency and ethnicity information. 51.4% of the patients were predicted as East Asian and 25.9% was predicted as 'others', suggesting that the ethnic groups of 'others' patients are under-represented in gnomAD. 18% of the patients with regions of homozygosity (ROH) of over 0.02% of the genome, mostly from consanguinity, had a significantly higher DR of 44.5% than patients with no ROH ($p<0.00001$). As expected, 81% of the diagnosed patients had homozygous P/LP variants and of the 1,018 variants reported as diagnostic, almost half of them (468) were novel and never observed in gnomAD. On the other hand, we also found a subset of variants that were commonly observed in our dataset but not in gnomAD to the level we could classify them as likely benign: we were able to reclassify 558 P/LP/VUS ClinVar variants to likely benign.
This study highlights that countries/ethnic groups at an earlier stage of utilizing ES/GS are under-represented in public databases (DB) and how further efforts should be made to be more inclusive in genomic data generation to allow us to identify novel variants, enrich the DBs and consequently improve the diagnostic performance for patients with suspected rare disorders.

Title: Understanding the causes of phenotypic heterogeneity among carriers of clinical pathogenic variants.

Authors:

A. Wei[1], R. Border[1], B. Fu[1], S. Cullina[2], M. Udler[3], N. Brandes[4], S. Sankararaman[1], E. E. Kenny[2], V. Ntranos[4], N. Zaitlen[1], V. Arboleda[1]; [1]Univ. of California, Los Angeles, Los Angeles, CA, [2]Icahn Sch. of Med. at Mount Sinai, New York, NY, [3]Massachusetts Gen. Hosp., Boston, MA, [4]Univ. of California, San Francisco, San Francisco, CA

Abstract:

Background Approximately 1.8%-3.4% of the population carries a pathogenic or likely pathogenic variant in genes implicated in autosomal dominant, monogenic diseases. However, only a portion of these carriers will develop disease (incomplete penetrance); those who do will range from mild to severe disease (variable expressivity). Here, we apply novel methods to understand the underlying causes of heterogeneity in penetrance and expressivity to improve the accuracy of genomics-informed prognosis. We investigate three underlying mechanisms: variable effect sizes of clinical variants, polygenic background of carriers, and moderation of variants by genetic background (epistasis). **Methods** We analyzed 200K whole exome-sequencing and clinical phenotypes from the UK Biobank (UKB) to identify carriers of rare, monogenic variants known to affect six phenotypes: LDL, HDL, triglycerides, type 2 diabetes, and body mass index (BMI). We first applied a novel protein language model (PLM) variant pathogenicity predictor to all missense variants in the UKB, analyzing the relationship between PLM score and phenotype. Second, we examined the relationship between polygenic risk score (PRS) and carrier phenotype. Third, we applied a novel linear mixed model, FAME (FAst Marginal Epistasis Estimation) that quantifies the fraction of phenotypic variance modulated by epistasis between clinical monogenic variants and polygenic background. **Results** *MC4R* PLM scores were strongly correlated with BMI (r=-0.29, p=7E-4, n=14,008), and, for the first time, delineated gain from loss-of-function variants that protect from or cause monogenic obesity; results replicated in Mt. Sinai's BioMe biobank (r=-0.21, p=0.047, n=1,456). Individuals carrying missense mutations with the lowest 10% of PLM scores had a mean BMI 18.5% higher than carriers of missense mutations with the highest 10% of PLM scores. Similarly strong correlations were found in other gene-phenotype pairs (*LDLR*-LDL r=-0.49, p.=3.22E-22, n=21,333; *GCK*-HbA1c r=-0.28, p=1.388E-2, n=369). We then found that polygenic background contributes to variable expressivity: carrier PRS were significantly associated with phenotype for five of six traits. Finally, the FAME results reveal that epistasis strongly impacts carrier phenotype: for example, the effect of epistasis on carrier phenotype was 48% the size of the effect of carrier status on phenotype for monogenic high LDL carriers (p=2.88E-10). In summary, we identified three significant contributors to phenotypic heterogeneity in monogenic carriers: differing effect size of clinical variants, carrier polygenic background, and genetic epistasis.

Session 097: Cell-type and context-dependent regulation of gene expression

Location: Conv Ctr/Ballroom C/Level 3

Session Time: Saturday, November 4, 2023, 10:30 am - 12:00 noon

Title: Single-cell eQTL mapping across brain cell types reveals context-specific genetic regulation in AD genetics

Authors:

N. Sun[1], Y. Park[2], C. Boix[3], L. Hou[4], X. Xiong[1], Y. Tanigawa[1], X. Han[5], D. Bennett[6], L-H. Tsai[1], M. Kellis[1]; [1]Massachusetts Inst. of Technology, Cambridge, MA, [2]Univ. of British Columbia, Vancouver, BC, Canada, [3]Massachussetts Inst. of Technology, Cambridge, MA, [4]MIT & Broad Inst., Cambridge, MA, [5]Massachusetts Inst. of Technology, Boston, MA, [6]RUSH Univ., Chicago, IL

Abstract:

Genetic variants can impact gene expression patterns, and contribute to diverse brain neurodegenerative and psychiatric disorders, acting through multiple brain cell types. Here, we leverage single-nucleus RNA-seq data from post-mortem human prefrontal cortex samples across 29 cell types, 2.3 million nuclei, and 427 individuals to discover 362k fine-mapped cis-acting single-cell expression quantitative trait loci (sc-eQTLs), linked to 9,608 target genes (eGenes). We find that ~30% of our eQTL links are newly-detected compared to GTEx, which are enriched for cell-type-specific eQTLs.

We investigate underlying mechanism of cell-type-specificity of eQTLs in three ways: 1) accessible chromatin regions from seven brain cell types show cell-type-dependent enrichment of sc-eQTL (FM-eQTL); 2) chromatin loops show cell-type-dependent enrichment of gene-enhancer linking supported by FM-eQTL; 3) we identify 301k TF-FMeQTL-gene trios where effect size of TF-motif-overlapped FMeQTL is correlated with TF (transcription factor) expression level, indicating these TFs contributes to the cell-type-specificity of these FMeQTLs.

We use independent component analysis to discover 161 co-expressed modules (16,409 genes) across combinations of cell types and individuals, and find 56 modules are significantly enriched in eGenes, suggesting that genetic variants have broad effects on gene expression.

We use motif analysis to predict TFsfor each module. We find 17 TFs (e.g. RUNX1, MEF2A, and HIF1A) show significant binding in eSNPs and regulating eGenes in a microglia-specific module, suggesting that these TFs may regulate gene expression by genetic variants. We next identify trans-acting eQTLs for genes in our modules without enrichment in eGenes if the upstream TFs of these genes have cis-eQTLs. For example, GLI2 regulates the expression of ASRGL1 through trans-eQTLs in astrocytes.

We find condition-specific effects in cis-eQTLs for Alzheimer's disease (AD) patients. We discover 104,092 (42.9%) cis-eQTLs specifically in AD individuals and 98,647 (40.7%) cis-eQTLs specifically in controls, 39,845 (16.4%) cis-eQTLs shared in both cohorts, across all cell types. For example, 12 AD risk genes (including SLC24A4 and ARL17B) are specific eGenes in AD individuals, while 8 AD risk genes (including ABCA7 and CD33) are specific in control individuals, in microglia.

Overall, our study provides an unprecedented resource of genetic regulation on gene expression in the brain at cell type resolution, uncovers the potential drivers mediating the regulation systematically, and exemplifies the genetic regulatory circuits that may cause brain diseases.

Title: Initial eQTL Discovery in Individual Pancreatic Cell Types from the Human Pancreas Analysis Program

Authors:

M. Hazuga[1], R. Elgamal[2], K. Lorenz[3], K. H. Kaestner[4], K. Gaulton[2], S. F. A. Grant[4,5,6], B. F. Voight[3,4,7]; [1]Genomics and Computational Biology Graduate Group, Univ. of Pennsylvania - Perelman Sch. of Med., Philadelphia, PA, [2]Dept. of Pediatrics, Univ. of California San Diego, San Diego, CA, [3]Dept.s of Systems Pharmacology and Translational Therapeutics, Univ. of Pennsylvania – Perelman Sch. of Med., Philadelphia, PA, [4]Dept. of Genetics, Univ. of Pennsylvania – Perelman Sch. of Med., Philadelphia, PA, [5]Dept. of Pediatrics, Univ. of Pennsylvania – Perelman Sch. of Med., Philadelphia, PA, [6]Ctr. for Spatial and Functional Genomics, Div. of Human Genetics, Children's Hosp. of Philadelphia, Philadelphia, PA, [7]Inst. of Translational Med. and Therapeutics, Univ. of Pennsylvania – Perelman Sch. of Med., Philadelphia, PA

Abstract:

The pancreas is a central organ in the endocrine system composed of multiple, distinct cell types with etiologically relevant functions for cancer, type 1 and type 2 diabetes, metabolic, and glycemic traits. Pancreatic studies have largely focused on bulk tissue data, limiting the knowledge of how each of its distinct cell types contribute to such traits. Recently, single-cell methods and tissue collection efforts have substantially expanded for the pancreas, facilitating a detailed characterization of genetics, expression, and regulation of these cell types at a population scale and across disease states. One such effort is the Human Pancreas Analysis Program (HPAP), aiming to procure and comprehensively study the human pancreas and pancreatic islet in diabetes and integrate data with the clinical history from de-identified donors. This study creates an opportunity to investigate how each of the pancreas' islet cell types, such as insulin secreting beta cells or glucagon secreting alpha cells, uniquely contribute to metabolic traits. Here, we report an initial expression quantitative trait locus (eQTL) discovery effort using 232,102 cells collected from single-cell RNA Sequencing from 63 pancreatic donors (33% non-European ancestry). Clustering analysis based on single-cell expression data identified pancreatic islets, pancreatic exocrine cells, and immune cell types, and we focused initial efforts on populations of alpha and beta cells where we had sufficient cell numbers across donors for discovery. Using standard analysis pipelines, we observed 559 and 199 eQTLs in alpha and beta cells (permutation $p<0.05$), respectively, with 19 shared across both cell types. Formal colocalization, meta-analysis with other procurement efforts, and integration with genomic regulatory features promises to reveal a deeper understanding of how each cell's function contributes to the pathogenesis of pancreatic disease and nomination of putative effector transcripts underlying disease.

Title: SAIGE-QTL: scalable and accurate expression quantitative trait locus mapping for single-cell studies

Authors:

W. Zhou[1], A. Cuomo[2], G. Chau[3], M. Kanai[3], C. Krishna[3], R. Xavier[3], D. MacArthur[4], J. Powell[5], M. Daly[6], B. Neale[6]; [1]Massachusetts Gen. Hosp., Cambridge, MA, [2]Garvan Inst., Darlinghurst, Sydney, Australia, [3]Broad Inst. of MIT and Harvard, Cambridge, MA, [4]Garvan Inst., Darlinghurst, Australia, [5]Garvan Inst. of Med. Res., Sydney, Australia, [6]Massachusetts Gen. Hosp., Boston, MA

Abstract:

eQTL mapping based on transcriptomes across multiple cell types and tissues provides evidence on functional genetic variants and facilitates understanding the molecular basis for human diseases when integrated with GWASs. More recently, single-cell RNA-seq (scRNA-seq) methods enable the most fine-grained view of cellular diversity and identification of eQTL with a heterogeneity of genetic effects across cellular contexts. However, major limitations remain in current eQTL mapping methods for scRNA-seq data. Here, we propose a scalable and accurate computational method to efficiently map the genetic basis of transcriptomes measured by scRNA-seq for common and rare genetic variations.

Our method, called SAIGE-QTL, is an extension of the previously developed mixed model method SAIGE for GWASs in large-scale biobanks and data sets. It conducts genome-wide eQTL analysis using scRNA-seq data based on the Poisson mixed models to account for within-sample cellular heterogeneity, between-sample relatedness, and population stratifications through multiple random effects. It performs set-based association tests for more powerful analysis of rare variants and identifies cell context-specific eQTLs with interactions of genotype and cellular contexts included in the models. SAIGE-QTL is feasible for identifying cis- and trans-eQTLs in large-scale RNA-seq data of hundreds of thousands or even millions of cells from hundreds to thousands of individuals.

The extensive simulation studies demonstrated that SAIGE-QTL has well controlled type I error rates. We applied it to the OneK1K cohort, which consists of scRNA-seq data on PBMCs from ~1,000 donors and assessed the association power of sc-eQTL in the preliminary analysis of naive B cells (n = 81,284 cells) for 8,314 genes expressed in at least 1% of the cells. Strikingly, the sc-eQTL analysis outperformed the analysis using pseudobulk data with a 21% increase in average χ^2 statistics in 1,072 Bonferroni corrected-significant cis-eQTLs with $p < 6.01 \times 10^{-6}$ in either analysis. We identified ~25% more eQTLs using SAIGE-QTL compared to previous pseudobulk analysis (FDR<5%), and the two analyses have highly correlated effect size estimates of cis-eQTLs (Pearson's R2 = 0.98). We are applying SAIGE-QTL to the full OneK1K data with sixteen cell types to obtain a more comprehensive understanding of the genetic regulation of gene expressions.

In summary, we develop a novel computational method that addresses the unsolved analytical challenges for systematic genetic analysis on single-cell transcriptomics, offering the potential to interpret complex trait associations through genome-wide colocalization events.

Title: Systematic screen of eQTL variants across cell types uncovers novel functional variants

Authors:

E. Padhi[1], E. Greenwald[1], D. Nachun[1], M. Degorter[2], Z. Weng[1], S. B. Montgomery[1], N. S. Abell[1]; [1]Stanford Univ., Stanford, CA, [2]Stanford Univ, Stanford, CA

Abstract:

Expression quantitative loci (eQTLs) mapping has been a powerful approach to identify genetic variation associated with changes in gene expression within and across diverse human tissues. Previous work from our group has aided identifying causal variants at eQTL loci due to linkage disequilibrium by employing massively parallel reporter assays (MPRAs) to assess the functional effects of 40,000 variants at ~700 eQTL loci detected in 1000 Genomes Project population samples. This revealed an intricate genetic architecture underlying these loci, however, whether these variants have functional effects in other cellular contexts is unclear.
To gain deeper insights into the effects of eQTL variants across cell types, we expanded MPRAs of these 40k variants using K562 and HEK293 cells. We discovered 5,722 sequences exhibiting differential regulatory element activity when comparing to our initial results in lymphoblastoid cells (LCLs). These sequences are enriched for binding of transcription factors that are cell types specific and related to function of each cell type, such as such as GATA family members for K562 and IRF family members for LCLs. Quantitatively, these sequences demonstrated complex patterns of regulatory element activity where 62-81% of the directions of effect were shared across cell types. However, when considering the absolute magnitude of the effect, the sharing ranged from 24% to 45%. This demonstrates that cell type largely acts to amplify or dampen regulatory element activity for sequences at multiple eQTL loci.
Next, we identified 3,766 variants that exhibited a significant interaction effect with cell type, with 65% of these variants not being identified in our initial study. These variants also demonstrated complex patterns of activity, with 62-86% sharing the direction of effect and 22-45% sharing the magnitude of the effect. When further summarizing variants to a per locus level, we identified 72% of loci have at least one additional functional variant outside of LCLs, with some loci having up to 33 additional functional variants. Currently, we are conducting analyses to understand how these functional variants intersect with other genomic properties such as conservation and variants from genome-wide association studies. Collectively, this work has identified novel functional variants at eQTL loci across multiple cell types, while also identifying many variants with highly cell type-specific effects despite being within loci that may be considered shared for eQTL effects.

Title: The impact of ethnicity on the single-cell immune response to malaria in West African children.

Authors:

T. Shahin[1], M. Dieng[1], W. Abdrabou[1], O. Bayaraa[1], B. Alamad[1], J. Jurkovic[1], A. Diawara[1], V. Manikandan[1], S. S. Sermé[2], I. Soulama[2], Y. Idaghdour[1]; [1]New York Univ. Abu Dhabi, Abu Dhabi, United Arab Emirates, [2]Ctr. Natl. de Recherche et de Formation sur le Paludisme, Ouagadougou, Burkina Faso

Abstract:

Plasmodium falciparum caused 621,000 deaths in 2021 and mostly in children under the age of 5. Differences in susceptibility to malaria between populations are documented but the underlying mechanisms are poorly understood. The Fulani ethnic group in Africa was reported to be less susceptible to malaria compared to other sympatric groups such as the Mossi, with lower disease rates and parasite load, as well as higher serological protection and immune activation. Unraveling the molecular and immunological basis of this protection remains challenging in part due to a lack of in-depth immunological characterization at the cellular level. Here we study the immune response to *P. falciparum* infection in infected (Inf) and non-infected (NI) Fulani and Mossi individuals in rural Burkina Faso using a multi-omics approach. We generated and analyzed > 95,000 single-cell RNA-Seq profiles from peripheral blood mononuclear cells of 177 male and female children from the Fulani (Inf, n = 52, $\log_2$ parasitemia: 4.8-16.2, and NI, n = 34) and sympatric Mossi (Inf, n = 69, $\log_2$ parasitemia: 2-18.4, and NI, n = 22) ethnic groups. We identified 30 cell subtypes and report the cell types whose proportions differ between the two ethnic groups in the Inf and NI states, including γδ T cells that show a distinct profile in the Inf Fulani samples compared to the Inf Mossi counterpart. Moreover, cell-type specific differential expression analysis revealed ethnic-specific immune signatures in both Inf and NI states and highlight in particular lower pro-inflammatory responses in the $CD14^+$ monocytes of the Inf Fulani, characteristic of asymptomatic disease, driven by downregulation of *TNF*, *IL1B* and *IL6*, among other cytokines and chemokines. Single-cell eQTL analysis across multiple cell types revealed hundreds of significant regulatory variants including cell-type and ethnic-specific effects. Overall, the results demonstrate the power of single-cell transcriptomics in identifying ethnic, cell-type specific and genetic regulatory effects in the host immune response to malaria and provide valuable single-cell eQTL and transcriptomic data from populations underrepresented in genomic studies.

Title: Single nucleus RNA-sequencing of a longitudinal subcutaneous adipose tissue cohort discovers an adipocyte subcellular state with 89 genes for a large weight loss outcome in bariatric surgery.

Authors:

S. Lee[1], B. Balliu[2,3,4], A. Kar[1], S. Rajkumar[1], M. Alvarez[1], S. Heinonen[5], B. van der Kolk[5], U. Säiläkivi[6], T. Saarinen[6], D. Kaminska[7,8], V. Männistö[9], J. Pihlajamäki[7,10], A. Juuti[6], K. Pietiläinen[5,11], P. Pajukanta[1,12]; [1]Dept. of Human Genetics, David Geffen Sch. of Med. at UCLA, Los Angeles, CA, [2]Dept. of Pathology and Lab. Med., David Geffen Sch. of Med. at UCLA, Los Angeles, CA, [3]Dept. of Computational Med., David Geffen Sch. of Med. at UCLA, Los Angeles, CA, [4]Dept. of Biostatistics, UCLA Fielding Sch. of Publ. Hlth., Los Angeles, CA, [5]Obesity Res. Unit, Res. Program for Clinical and Molecular Metabolism, Faculty of Med., Univ. of Helsinki, Helsinki, Finland, [6]Dept. of Abdominal Surgery, Abdominal center, Helsinki Univ. Hosp. and Univ. of Helsinki, Helsinki, Finland, [7]Inst. of Publ. Hlth.and Clinical Nutrition, Univ. of Eastern Finland, Kuopio, Finland, [8]Dept. of Med., Div. of Cardiology, UCLA, Los Angeles, CA, [9]Inst. of Clinical Med., Univ. of Eastern Finland and Kuopio Univ. Hosp., Kuopio, Finland, [10]Dept. of Med., Endocrinology and Clinical Nutrition, Kuopio Univ. Hosp., Kuopio, Finland, [11]Hlth.y Weight Hub, Abdominal Ctr., Helsinki Univ. Hosp. and Univ. of Helsinki, Helsinki, Finland, [12]Inst. for Precision Hlth., David Geffen Sch. of Med. at UCLA, Los Angeles, CA

Abstract:

Bariatric surgery is a common intervention for the management of obesity. However, the degree of weight loss achieved through this procedure exhibits substantial inter-individual variability and the cellular and molecular mechanisms contributing to this variability remain poorly understood. To address this knowledge gap, we performed the largest to date longitudinal single nucleus RNA-sequencing (N=132 samples) on subcutaneous adipose biopsies from 33 Finns at 2 pre- and 2 post-bariatric surgical time points (mean pre-surgery BMI=44 (±6) kg/m^2). We stratified individuals by the median decrease between pre- and post-surgery BMI into high (N=16; BMI decrease=12) and low (N=17; BMI decrease =7) BMI difference (dBMI) groups. In each group, we used Milo to identify cell states within adipocytes and found that 45% and 21% of the identified cell states showed differential abundance by weight loss in the high and low dBMI group, respectively (FDR<5%). Next, we performed differential expression analysis and found 89 and 9 genes up-regulated in the adipocyte states more abundant after weight loss in the high and low dBMI groups, respectively (FDR<5%). The 89 up-regulated genes (e.g., *ABCD2*, *ACACB*, and *PRKAG2*) in the high dBMI group showed significant enrichment (FDR<5%) for 18 fatty acid beta-oxidation and lipid catabolism pathways, indicating increased adipose energy expenditure (i.e., fat burning) in the high dBMI group. In addition, difference in their expression level in adipocytes pre- and post-surgery accounted for 18% of the variance in dBMI, whereas differences in the expression level of all adipocyte-expressed genes explained only 0.9% of the dBMI variance. To validate these findings, we analyzed an independent Finnish bariatric surgery cohort, KOBS, comprising pre- and post-surgery adipose bulk RNA-seq data (N=168). We constructed a longitudinal gene signature (LGS) created using PCA of the delta expression between the pre- and post-surgery time points while only including the adipocyte marker genes among the 89 genes (N=31) into the LGS to minimize cell-type heterogeneity in these bulk RNA-seq data. We observed a significant correlation (P=0.049) between the LGS and dBMI in KOBS, thus providing independent support for the association between the identified adipocyte cell state and increased weight loss. Finally, we mapped adipocyte cis-eQTLs at the baseline (N=69) and found that 25% of the 89 genes had cis-eQTLs (FDR<10%), including several obesity GWAS SNPs. In summary, we discover and replicate an adipocyte cellular state with fat burning centered genes as a marker for a large weight loss outcome during bariatric surgery.

Title: Driver project for advancing long-read de novo genome assembly methods in clinical research

Authors:

E. Delot[1], J. Monlong[2,3], H. Barseghyan[1], S. Berger[1], S. Bhattacharya[1], J. Cool[4], J. Lotempio[5], B. McNulty[2], S. Negi[2], S. O'Rourke[2], K. Miga[2], B. Paten[2], E. Vilain[5]; [1]Childrens Natl. Res. Inst., Washington, DC, [2]UC Santa Cruz Genomics Inst., Santa Cruz, CA, [3]Inst. de Recherche en Santé Digestive (IRSD), Université de Toulouse, Toulouse, France, [4]Chan Zuckerberg Initiative, San Francisco, CA, [5]Univ. of California Irvine, Irvine, CA

Abstract:

To evaluate technology readiness and resulting utility, we launched a pilot project using long-read *de novo* assemblies in a research clinical setting, in complement to our efforts to validate optical genome mapping (OGM) for clinical use for rare disease diagnosis. We focused on regions of the genome of critical importance for disambiguation of Differences of Sex Development (DSD) conditions refractory to diagnosis with short-read sequencing because of the complexity of the genomic architecture. DSD are among the most difficult conditions to diagnose clinically, with a large molecular diagnostic gap, and are considered an underserved population by the Sexual and Gender Minority Research Office of the NIH.
Methods: High-molecular-weight DNA extracted from stabilized blood of participants in the DSD-Translational Research Network (DSD-TRN) was subjected to either OGM (Bionano platform) or Long-Read Sequencing (LRS) both at ~30x coverage. LRS was performed on the ONT Nanopore platform, with Shasta assembly, using the pipeline developed for the NIH CARD project. Results: OGM was able to identify mosaicism in several samples with 45,X/46,XY sex chromosome complement. It also highlighted unexpected complexity in DSD caused by 9pter deletion, revealing translocations between telomeric regions of 9p and other chromosomes, leading to loss of material on both. OGM also detected that the Y-to-X translocations found in the majority of 46,XX males (testicular DSD) actually involve different locations in both X and Y chromosomes. With LRS, alignment to the newly completed Y chromosome T2T reference for the first time allows disambiguation of the X and Y pseudoautosomal regions where those translocations occur. While OGM was not able to resolve the gene/pseudogene conversions/30kb deletions responsible for most pathogenic alleles in *CYP21A2*, the gene causing Congenital Adrenal Hyperplasia (CAH), the region was well covered with ONT reads and we observed a drop of coverage indicative of the presence of the 30 kb deletions in several patients. Furthermore, the density of small variants called from the ONT reads suggested gene conversions events. Conclusions: Nanopore-based LRS was able to identify rearrangements at the pseudoautosomal regions of the sex chromosomes. It also covered the complex *CYP21A2* gene/pseudogene region, bringing hope that a specialized LRS-based test could be designed to replace the current clinical test (which requires 4 long-range PCRs, bidirectional Sanger sequencing, and MLPA) for CAH, the most common 46,XX DSD condition. In parallel, OGM can be used to replace multiple technologies currently required to diagnose several chromosomal DSD.

Title: Low-pass mate-pair genome sequencing versus chromosomal microarray analysis: validation and implementation in prenatal diagnosis

Authors:

Z. Dong[1], J. Qian[1], H. Wang[2], M. Chau[3], H. Liang[4], W. Tse[1], Y. Cao[1], T. Leung[1], Y. Kwok[1], K. Choy[1]; [1]The Chinese Univ. of Hong Kong, Hong Kong, China, [2]Shenzhen Baoan Women's and Children's Hosp., Shenzhen, China, [3]Baylor Coll. of Med., Houston, TX, [4]Shenzhen Res. Inst., The Chinese Univ. of Hong Kong, Shenzhen, China

Abstract:

Low-pass genome sequencing (GS) has been adopted and recommended by the American College of Medical Genetics and Genomics for detecting germline structural variants. Whilst chromosomal microarray analysis (CMA) has been recommended for routine prenatal diagnosis, the clinical diagnostic utility provided by low-pass mate-pair GS (5kb DNA fragments) has yet to be evaluated. Herein, we conducted a prospective back-to-back comparison of low-pass GS and CMA for 222 fetuses with ultrasound anomalies (with single/multiple affected systems). Each sample was subjected to low-pass mate-pair GS and CMA in parallel. CNVs and regions with Absence of Heterozygosity (AOH) were reported by both assays, while low-pass GS also identified structural rearrangements with/without involvement of CNVs. Low-pass mate-pair GS not only identified all 24 aneuploidies, or pathogenic or likely pathogenic (P/LP) CNVs detected by CMA in 22 cases (9.9%, 22/222), but also detected multiple regions with AOH suspected parental consanguinity in two cases consistently with CMA results (0.9%, 2/222). In addition, low-pass GS revealed three cryptic P/LP deletions with insufficient probe coverage in CMA. For instance, a 35kb heterozygous intragenic deletion was detected in *RBFOX2* in a fetus with cardiac anomalies. Further, low-pass GS identified 34 structural rearrangements, including (1) 11 balanced translocations/inversions, and (2) 23 structural rearrangements involving CNVs. All were further verified with optical genomic mapping and Sanger sequencing for the genomic composition. Among them, 15 structural rearrangements (15 cases) showed the potential clinical significance. For example, a 30kb duplication was found to insert into gene *BMPR1B* in a fetus with likely osteocraniostenosis. Lastly, the genomic composition reported by low-pass GS also helped to reclassify the pathogenicity of CNVs defined by CMA. For instance, a de novo 3.5Mb duplication in chr13 was classified as uncertain clinical significance by CMA in fetus with ventriculomegaly; however, it was inserted to chr2 revealed by low-pass GS potentially leading to the dysregulation of *GLI2*. Overall, about 32% of the affected genes showed expressions in amniotic fluid cells, and revealed RNA aberrations. In summary, low-pass mate-pair GS shows its superiority in identifying clinically significant structural variants underlying fetal ultrasound anomalies as compared with CMA platform for prenatal diagnosis.

Title: Next generation cytogenomics: high resolution analysis of chromosomal aberrations using proximity ligation sequencing.

Authors:

Y. Liu[1], H. Fang[1], Y. Wu[1], S. Sullivan[2], I. Liachko[2], S. Eacker[2]; [1]Univ of Washington, Seattle, WA, [2]Phase Genomics, Seattle, WA

Abstract:

Cytogenetic assays are commonly used to identify genomic structural variants (SVs) in clinical and research settings. To cytogenetically characterize a genome, it is common to apply multiple methods including karyotyping, fluorescence *in situ* hybridization, and chromosome microarray. Each of these methods has limitations that necessitates the use of the other methods in parallel. In this study we validate the use of proximity ligation sequencing (PLS, aka Hi-C seq) as a tool to generate cytogenomic profiles for clinical genetics cases. PLS captures ultra-long-range chromatin interactions within the intact nucleus, generating chromosome-scale contiguity information from small numbers of dividing or non-dividing cells (less than 100,000). Using an AI-powered platform, we use the pair-wise sequence interaction data generated by PLS to identify translocations, inversions, insertions, deletions, duplications, and complex rearrangements with high accuracy and resolution from ~7x genome coverage. In this study, we sequenced more than 200 clinical cases with extensive cytogenetic characterization to benchmark performance. We observed an overall sensitivity and specificity of more than 0.95 for clinically reportable variants, including cases of absence of heterozygosity (AOH) and Robertsonian translocations. In addition, unlike other methods such as long read sequencing methods, PLS does not require ultra-long DNA molecule and can be used for various sample types including FFPE samples. Together, these data demonstrate that PLS can provide clinically actionable cytogenomic insights with a single, sequencing-based assay.

Title: A high-throughput and high-resolution diagnostic assay for determination of methylation status of the CGG repeat region in 5'UTR of FMR1 gene

Authors:

Z. Zhang[1], M. Robinson[2], N. T. Leach[1], H. Zhu[1], D. Boles[2], P. Okamoto[1]; [1]Labcorp, Westborough, MA, [2]Labcorp, Research Triangle Park, NC

Abstract:

Fragile X syndrome (FXS) is the second most common cause of inherited intellectual disability (behind Down syndrome), affecting approximately 1 in 4000 males and 1 in 8000 females. FXS is associated with pathogenic variants in the FMR1 gene located on Xq27.3, which cause FMR1 protein (FMRP) deficiency. For >98% of cases, loss of FMRP results from hypermethylation of the FMR1 5'UTR region, which is triggered by CGG repeat expansion. We have developed and validated a fragile X (FRAX) methylation PCR (mPCR) assay that overcomes many of the testing limitations currently encountered in Southern blot (SB) analysis for methylation detection, such as allele-specific resolution of methylation status, processing time and amount of input DNA required. Briefly, the assay uses methylation-specific immunoprecipitation to separate the genomic DNA into methylated and unmethylated fractions, which along with the unfractionated genomic DNA, are processed in parallel. After resolution of the PCR products by capillary electrophoresis, a custom calling tool is used to qualitatively determine methylation status. Various specimen types, including whole blood, direct or cultured fetal samples, saliva, and buccal samples were validated in the FRAX mPCR assay. Based on our validation study, analytical sensitivity and specificity were 100% (95%CI: 95.3%-100%) for 98 alleles, in which all were concordant with SB results after manual review. Intra-assay repeatability and inter-assay reproducibility for methylation status were 100% (95%CI: 92.5%-100%; 60/60 alleles) and 98.3% (95%CI: 89.9%-99.9%; 59/60 alleles), respectively. Our data also demonstrated that this assay can determine the methylation status of mosaic alleles with high sensitivity and resolution. Taken altogether, we have developed a robust FRAX mPCR assay that directly assesses methylation and eliminates the need for SB analysis, thereby improving resolution, specificity, sensitivity and throughput.

Title: Detection and characterization of repeat expansions in patients with neurodevelopmental disorders using short- and long-read sequencing.

Authors:

I. Rajan Babu[1], S. Adam[1], C. Guimond[1], R. Chiu[2], I. Caglayan[1], CAUSES Study, IMAGINe Study, I. Birol[2], J. Friedman[1]; [1]The Univ. of British Columbia, Vancouver, BC, Canada, [2]BC Cancer Agency, Vancouver, BC, Canada

Abstract:

Background: At least 60 short tandem repeat (STR) expansion disorders are known to date. Repeat expansions (REs) are dynamic, often changing in size during parent-to-child transmissions. Repeat length and motif composition (i.e., the presence of non-canonical motifs or interruptions) influence repeat stability, penetrance, age of onset, severity, and/or clinical features of RE disorders. Recent improvements to genomic sequencing technologies and analytical software have greatly enhanced our ability to study REs. Reliable genotyping and characterization of STRs are crucial for diagnosing and counseling patients with RE disorders. **Methods:** We used ExpansionHunter, ExpansionHunter Denovo, STRetch, and STRling to identify both known and novel pathogenic STR expansions in the Illumina short-read exome and whole-genome sequencing (SRS) data of ~600 patient-parent trios (child with a neurodevelopmental disorder and both unaffected biological parents) from two of our research studies. **Results:** We identified 63 individuals with REs at the *AR*, *ATXN1*, *ATXN2*, *ATXN8OS*, *DMD*, *DMPK*, *FXN*, *GIPC1*, *HTT*, or *TCF4* locus, and also apparent REs at 47 novel STR loci, 14 of which are in OMIM disease genes. We used REViewer, a software we developed for STR visualization in SRS data, to verify the ExpansionHunter genotype calls of these candidate REs and also investigate the repeat sequences for evidence of interrupting motifs. Nanopore long-read sequencing (LRS) and analysis with Straglr, a tandem repeat genotyping software we developed for LRS data, were able to characterize the REs identified by SRS more accurately. For example, in a mother and a proband with *ATXN8OS* expansion, both clinical PCR and SRS genotyped the RE as ~80 CTG repeats, while LRS revealed intergenerational instability and expansion of the maternal allele from 86 repeats to 525 repeats in the proband. Interestingly, in the proband, LRS also highlighted the presence of expanded stretches of the non-canonical TTG repeat motif. In another case, we identified a rare paternally-inherited ~35 AAAAT RE in *LITAF* on SRS and analysis with ExpansionHunter. LRS and Straglr genotyping demonstrated ~1200 AAAAT repeats in the locus in the proband and the father. We are currently validating additional SRS-flagged RE candidates using targeted Nanopore and/or PacBio LRS. **Conclusion:** Our research will outline the diagnostic utility of SRS and LRS technologies and complementary bioinformatics methods in resolving and characterizing REs that may be the underlying cause of neurodevelopmental disabilities in some of our undiagnosed patients.

Session 099: Epigenomics in neurodevelopmental disorders

Location: Conv Ctr/Room 146B/Level 1

Session Time: Saturday, November 4, 2023, 10:30 am - 12:00 noon

Title: *Snord116* is a light sensing regulator of gene expression, behavior, and metabolism

Authors:

A. Mendiola[1,2], J. Rutkowsky[3,4,5,6], K. Neier[1], S. Hakam[1], D. Yasui[1,2], J. Ramsey[3,6], J. Lasalle[1,2,7,8,9]; [1]UC Davis Sch. of Med., Davis, CA, [2]UC Davis Dept. of Med. Microbiol. and Immunology, Davis, CA, [3]UC Davis, Davis, CA, [4]UC Davis Mouse Metabolic Phenotyping Ctr., Davis, CA, [5]UC Davis Sch. of Vet. Med., Davis, CA, [6]UC Davis Dept. of Molecular BioSci.s, Davis, CA, [7]UC Davis Genome Ctr., Davis, CA, [8]UC Davis Perinatal Origins of Disparities Ctr., Davis, CA, [9]UC Davis MIND Inst., Sacramento, CA

Abstract:

Prader-Willi syndrome (PWS) is an imprinted disorder characterized by sleep dysregulation, intellectual disabilities, and hyperphagia, caused by deficiencies of paternal 15q11-q13. Common to all PWS cases is loss of paternal *SNORD116,* a non-coding gene whose function remains elusive. We have previously shown that *Snord116* interacts with light by establishing diurnal rhythms of DNA methylation and gene expression in cortex; affecting gene pathways of circadian entrainment, metabolism, and neurodevelopment. In addition, an overexpressing transgenic *Snord116* mouse *(Snord116TG)* was spliced and processed into snoRNAs and host gene components in wildtype but not PWS *(Snord116del)* neurons. To test the hypothesis that *Snord116* orchestrates cortical gene expression changes in response to dynamic light cycles, we compared wildtype *(Snord116+/+,CTG-/-)*, deletion *(Snord116+/-,CTG-/-)*, overexpression *(Snord116+/+,CTG+/-)*, and compensation *(Snord116+/-,CTG+/-)* mice in a circadian entrainment treatment. Mice born into a 24 h light:dark (LD) cycle entrained at ~2 mos for 21 d to either 22 h (11:11) LD or 24 h (12:12) LD. Following entrainment, mice were transferred to an indirect respiration calorimeter (CLAMS) for 3 days in entrained LD cycles, followed by 5 days in constant darkness (DD). Running wheel (RW) behavior was measured for 1 week in entrained LD cycles followed by 1 week of 12:12. Tissues were harvested at Zeitgeber +6 (middle of sleep) following measurements. Overall, significant entrainment, genotype, light cycle, and sex effects were observed. Both males and females showed the previously reported decrease in respiratory exchange rate (RER) in the light cycle (sleep stage for mice) in *Snord116del* in LD, but not DD. Females show greater rhythmicity as well as adaptability to reduced amplitude of RER and activity in response to 11:11 entrainment. 11:11 entrainment resulted in reduced RER rhythmicity that corrected the *Snord116del* effects in DD. All genotypes were able to adapt to changes from 11:11 to 12:12 post-entrainment, however *Snord116del* males had reduced activity which was recovered by the presence of *Snord116*TG. To understand gene pathways involved in these adaptive responses. We perform RNA-seq analyses on brain cortex of *Snord116* genotypes utilizing a systems approach to defining rhythmic gene modules in wildtype cortex collected every 3 h to test the hypothesis that rhythmic metabolic genes are altered by light and *Snord116* genotype interactions. Together these results suggest that the *Snord116* locus is part of a learned adaptive response of mammals to change metabolism and activity in response to environmental change.

Title: The role of TCF20 complex in neurodevelopment and *MECP2*-related ASD pathogenesis

Authors:

J. Zhou[1], H. Hamdan[1], H. Yalamanchili[1], S. Bajikar[1], Z. Liu[1], M. Rasband[1], H. Zoghbi[2]; [1]Baylor Coll. of Med., Houston, TX, [2]Baylor Coll. of Med./HHMI/NRI, Houston, TX

Abstract:

Loss of function mutations in *MECP2* cause Rett syndrome (RTT) while duplications spanning the gene cause *MECP2* duplication syndrome (MDS). While the phenotypes of both disorders overlap with those of other autism spectrum disorders (ASDs), the precise molecular mechanism driving pathogenesis remains unclear. MeCP2 binds methylated DNA and recruits chromatin modifying proteins but the relationship between these proteins and gene expression changes is not clear. Therefore, identifying and characterizing MeCP2 interactors is crucial to fully understand the molecular function of MeCP2 and the pathogenesis of *MECP2*-associated disorders. To this end, we performed proximity-dependent biotin identification (BioID) in cultured rat primary neurons using a biotin ligase fused to MeCP2. In addition, we used two mutant alleles, $MECP2^{R111G}$ and $MECP2^{\Delta NLS}$, which disrupt DNA binding and nuclear localization of MeCP2, respectively, to filter out non-chromatin associated MeCP2 interactors. Our unbiased approach identified a novel MeCP2-interacting complex which includes TCF20, encoded by a known ASD-causing gene, and three other transcriptional regulators: RAI1, PHF14, and HMG20A. We found MeCP2 interacts with the TCF20 complex via PHF14 and that several RTT-causing MECP2 mutations reduce the binding between MeCP2 and the TCF20 complex. Next, we found that *Tcf20* modulates *MECP2*-mediated synaptogenesis in cultured primary neurons by co-regulating the key neuronal gene *Bdnf*. Further, reducing *Tcf20* partially rescued the behavioral deficits caused by *MECP2* overexpression in mice, underscoring a functional relationship between MeCP2 and TCF20 in MDS pathogenesis. We next assessed global gene expression changes in mouse models and found a significant proportion of differentially expressed genes in $Tcf20^{+/-}$ mice were also altered in $Mecp2^{-/y}$ mice; a majority of these genes changed in the same direction and with similar magnitude. Through CUT&RUN experiments, we found a significant reduction of TCF20/PHF14 binding to the downstream genes shared between $Tcf20^{+/-}$ and $Mecp2^{-/y}$ brains upon loss of MeCP2, suggesting that MeCP2 recruits TCF20 complex to chromatin to co-regulate gene expression. Notably, we identified one de novo *PHF14* missense mutation in a patient that displays clumsy gait, speech delay, and mild regression in gross motor skills and found that this mutation disrupts MeCP2-PHF14-TCF20 interaction. Our data demonstrate the critical role of a novel MeCP2-TCF20 complex for brain function and revealed a converging molecular mechanism whereby mutations of genes encoding several subunits in the same complex contribute to shared ASD symptoms.

Title: *ASXL3* in cortical development: Molecular insights from rare genetic variants

Authors:

S. Regan, B. McGrath, C. Ryan, S. Bielas; Univ. of Michigan, Ann Arbor, MI

Abstract:

De novo dominant *ASXL3* frameshift variants are the genetic basis of Bainbridge-Ropers Syndrome (BRS) and syndromic autism spectrum disorder (ASD). ASXL3 is a component of the Polycomb repressive deubiquitinase (PR-DUB) complex that is critical for Polycomb-mediated transcriptional repression. PR-DUB catalyzes the removal of ubiquitin from Histone 2A (H2AUb1), a repressive histone modification catalyzed by Polycomb repressive complex 1 (PRC1). The molecular mechanism of H2AUb1 and the cellular processes it regulates during normal development and disease remain largely unexplored. To investigate the role of ASXL3 in development we generated a mouse model that carries a clinically relevant *Asxl3* frameshift variant (*Asxl3fs)*. Genetic inactivation of *Asxl3* leads to perinatal lethality, multi-organ developmental defects, and increased levels of H2AUb1. Within the developing cerebral cortex, loss of ASXL3 perturbs the composition of excitatory neurons and fidelity of cortical layer deposition. We carried out single-cell RNA sequencing at three developmental stages during neurogenesis to characterize the cellular composition and transcriptomic changes. The emerging pathogenic model based on analysis of multiple cell types, at sequential developmental timepoints, implicates overactivation of Notch signaling that alters NPCs proliferation and timing of differentiation. These early developmental defects lead to altered composition of excitatory neurons with aberrant expression of proneural genes responsible for layer specificity at later timepoints. Across cortical development, dysregulated genes were enriched for high confidence ASD risk genes, implicating a convergent pathological mechanism. Together our findings underscore the importance of ASXL3 in Polycomb transcriptional repression during development and provide insight into developmental mechanisms altered by ASD risk genes.

Title: *De novo* variants in *RYBP* are associated with a severe neurodevelopmental disorder and congenital anomalies.

Authors:

M. Weisz-Hubshman[1,2], L. Burrage[3,2], J. Rosenfeld[1], S. Von Hardenberg[4], A. Bergmann[4], M. Rydzanicz[5], R. Ploski[5], R. Smigiel[6], A. Stembalska[7], W. Chung[8], R. Hernan[8], F. Lim[8], T. Brunet[9,10], S. Syrbe[11], B. Keren[12], S. Heide[12], D. Murdock[1], H. Dai[13], F. Xia[3], S. Ketkar[1], S. Chen[1], B. Dawson[1], S. Jangam[1,14], M. Wangler[1,14], Undiagnosed Disease Network, C. Bacino[1,2], B. Lee[1,2]; [1]Baylor Coll. of Med., Houston, TX, [2]Texas Children's Hosp., Houston, TX, [3]Baylor Coll. Med., Houston, TX, [4]Dept. of Human Genetics, Hannover Med. Univ., Hannover, Germany, [5]Med. Univ. of Warsaw, Warsaw, Poland, [6]Dept. of Pediatrics, Endocrinology, Diabetology and Metabolic Diseases, Med. Univ. of Wroclaw, Wroclaw, Poland, [7]Dept. of Genetics, Med. Univ. of Wroclaw, Wroclaw, Poland, [8]Columbia Univ., New York, NY, [9]Technical Univ. of Munich, Sch. of Med., Inst. of Human Genetics, Munich, Germany, [10]Dr. von Hauner Children's Hosp., LMU Hosp., Ludwig-Maximilians-Univ., Munich, Germany, [11]Div. of Paediatric Epileptology, Ctr. for Paediatrics and Adolescent Med., Univ. Hosp. Heidelberg, Heidelberg, Germany, [12]Dept. of Genetics, Assistance Publique - Hôpitaux de Paris, Univ. Hôpital Pitié-Salpêtrière, Paris, France, [13]Baylor Coll. of Med. / Baylor Genetics, Houston, TX, [14]Jan and Dan Duncan Neurological Res. Inst., Texas Children’s Hosp., Houston, TX

Abstract:

Polycomb group (PcG) proteins are key epigenetic regulators of gene transcription in humans and other organisms. PcG proteins are modular complexes, and their function is based on the interacting components. Multiple human neurodevelopmental disorders have been associated with pathogenic variants in genes encoding PcG protein subunits or proteins that interact with PcGs. *RYBP* encodes a zinc finger protein that was initially identified as a binding protein of RING1 and YY1 proteins and was found to be an important component of a subgroup of PcGs called non-canonical Polycomb Repressor Complex 1 (ncPRC1). RYBP-PRC1-complex has been shown to play an important role in Histone 2A monoubiqutination, however its putative role in pathogenic processes is unclear. In this study we describe six families with heterozygous *de novo* variants in *RYBP* and show that heterozygous missense variants and copy number variants in *RYBP* are associated with a syndromic intellectual disability disorder with multiple congenital anomalies. We show that all the missense variants in *RYBP* localize to exon 1 of the gene which encodes the zinc finger domain and ubiquitin binding domain of the protein. *In vitro* studies demonstrate that the RYBP p.C34W variant protein does not affect binding of YY1 or binding of ubiquitin. *In vivo*, initial overexpression studies in the Drosophila model showed that the human reference protein causes significant change in the wing morphology including changes in wing margin, and wing shape but both RYBP p.C34W and p.C34S variants failed to recapitulate the phenotype seen with the reference protein. These findings suggest that the assessed variants can be deleterious and validate the use of Drosophila model in the exploration of *RYBP* variants.

Title: Epigenomic and phenotypic characterization of DEGCAGS syndrome.

Authors:

S. Douzgou Houge[1,2], M. Horáčková[3], I. Aukrust[1], J. Paulsen[4], J. Kerkhof[5], R. Mendoza-Londono[6,7], L. Dupuis[6], M. Dickson[6], A. Aldeeri[8], T. Yu[8], R. Maroofian[9], T. Lay[9], F. Laccone[10], A. Soltysova[11], J. Rzasa[5], K. Karimi[5], D. Weis[3], B. Sadikovic[5,12]; [1]Dept. of Med. Genetics, Haukeland Univ. Hosp., Bergen, Norway, [2]Div. of Evolution and Genomic Sci., Sch. of Biological Sci., Univ. of Manchester, Manchester, United Kingdom, [3]Dept. of Med. Genetics, Kepler Univ. Hosp. Med Campus IV, Johannes Kepler Univ., Linz, Austria, [4]Dept. of Med. Genetics, St. Olavs Hosp., Trondheim Univ. Hosp., Trondheim, Norway, [5]Molecular Diagnostics Program, and Verspeeten Clinical Genome Ctr., London Hlth.Sci. Ctr., London, ON, Canada, [6]Div. of Clinical and Metabolic Genetics, The Hosp. for Sick Children, Toronto, ON, Canada, [7]Univ. of Toronto, Toronto, ON, Canada, [8]Div. of Genetics and Genomics, Boston Children's Hosp., Harvard Med. Sch., Boston, MA, [9]Dept. of Neuromuscular Diseases, Queen Square Inst. of Neurology, Univ. Coll. London, London, United Kingdom, [10]Dept. of Med. Genetics, Med. Univ. of Vienna, Vienna, Austria, [11]Dept. of Molecular Biology, Faculty of Natural Sci., Comenius Univ., Bratislava, Slovakia, [12]Dept. of Pathology and Lab. Med., Western Univ., London, ON, Canada

Abstract:

DEGCAGS syndrome (developmental delay with gastrointestinal, cardiovascular, genitourinary, and skeletal abnormalities, MIM #619488) is caused by biallelic variants in the *ZNF699* gene. *ZNF699* encodes a large, nuclear, zinc finger protein, suggesting a genomic regulatory role. The corresponding gene in D. melanogaster (*hang* from 'hangover') is required for development of ethanol tolerance.[1] The clinical characteristics of the affected individuals partially overlap with conditions with distinguishable episignatures, such as chromatin remodeling disorders (microcephaly, feeding difficulties, hypertrichosis).[2] We aimed to delineate the phenotypic features and investigate the DNA methylation profiles in affected individuals with variants in *ZNF699*. Through international collaboration, we collected data on 24 individuals with DEGCAGS, including 10 previously unpublished cases.[3,4] To determine whether *ZNF699* pathogenic variants are associated with a distinct epigenetic signature, we applied methylation arrays on peripheral blood DNA samples in 10 affected individuals. The methylation levels were fitted in a multivariate linear regression model to identify the differentially methylated probes. We then constructed a binary SVM classification model to differentiate DEGCAGS cases from controls. We report 7 novel *ZNF699* variants (NM_198535.3): c.421_424delGAGT p.(Glu141Profs*15), c.339del p.(Cys113Trpfs11), c.918_1001del p.(Cys318_Ser345del), c.1014_1097del p.(Ser346_Ser373del), c.175+1G>A (n/a), c.1039del p.(Ser347ProfsTer5) and c.1019G>A p.(Cys340Tyr). All 24 affected individuals present with global developmental delay and intellectual disability and a distinctive pattern of facial dysmorphisms, generalized hypertrichosis and early white hair (83%), immunodeficiency and/or anemia (50%), genitourinary (83%), skeletal (67%), and gastrointestinal (33%) anomalies. DNA methylation analysis indicated a clear and robust separation between individuals with variants in *ZNF699* and controls. We discovered a sensitive and specific DNA methylation episignature as a robust diagnostic biomarker for DEGCAGS, which can be applied as a tool in screening and genetic diagnosis. In conclusion, we expand the genetic heterogeneity of DEGCAGS, further delineate the related morbidity including a pattern of immunodeficiency, and reveal a specific DNA methylation episignature.
References: 1. Scholz, H. et al. (Letter) Nature 436: 845-847, 2005. 2. Kerkhof J et al. Genet Med. 2022;24:51-60. 3. Bertoli-Avella, A. M. et al. Genet. Med. 23: 1551-1568, 2021. 4. Biela M et al. Genes (Basel). 2022;13:168. 2022.

Title: Characterizing Metabolomic and Transcriptomic Signatures in Kabuki Syndrome

Authors:

Y. Jung, C. Hung, J. Choi, E. Lee, O. Bodamer; Boston Children's Hosp., Boston, MA

Abstract:

Kabuki Syndrome (KS) is a rare multisystem disorder with a variable clinical phenotype. The majority of KS cases are caused by dominant loss-of-function mutations in two genes, KMT2D (lysine methyltransferase 2D, KS1) and KDM6A (lysine demethylase 6A, KS2). Both KMT2D and KDM6A play a critical role in chromatin accessibility, which is essential for cell differentiation. In a previous study, we reported unique enhancer signatures in KS1 compared to control samples, providing insights into the underlying pathogenicity of KS. Early detection is crucial for timely therapeutic intervention. To identify potential biomarkers for early detection and to inform clinical trial readiness, we conducted a study in which we collected and analyzed plasma and urine metabolites from 40 KS patients and 12 healthy controls. We employed an untargeted approach using LC-MS/MS. Additionally, we profiled gene expression in the majority of KS patients. Our analysis revealed a total of 151 significantly altered metabolites between KS1 patients and controls, with these metabolites being clustered based on genotypes. Importantly, we identified a metabolite that showed the most significant change in both KS1 and KS2 patients. To classify the KS and control groups, we utilized a multivariate regression model called partial least squares discriminant analysis. Using this trained model, we achieved a high level of discrimination between the KS data and controls. Furthermore, pathway analysis revealed several disrupted pathways associated with the significantly altered metabolites and expression data. These findings provide valuable insights into the metabolic dysregulation underlying KS and highlight potential targets for further investigation and therapeutic interventions.

Session 100: Go beyond GWAS in type 2 diabetes, obesity and related metabolic disorders

Location: Conv Ctr/Room 147A/Level 1

Session Time: Saturday, November 4, 2023, 10:30 am - 12:00 noon

Title: Discovering stimulatory state specific type 2 diabetes GWAS mechanisms with single cell multi-omics on iPSC-derived fibro-adipogenic progenitor cell villages

Authors:

C. Ventresca[1], A. Varshney[1], P. Orchard[1], Y-C. Tsan[1], A. Monteiro da Rocha[1], M. Laakso[2,3], J. Tuomilehto[4], T. Lakka[2,3,5], K. Mohlke[6], M. Boehnke[1], L. Scott[1], H. Koistinen[4], F. Collins[7], T. Herron[1], S. Bielas[1], S. Parker[1]; [1]Univ. of Michigan, Ann Arbor, MI, [2]Univ. of Eastern Finland, Kuopio, Finland, [3]Kuopio Univ. Hosp., Kuopio, Finland, [4]Finnish Inst. for Hlth.and Welfare, Helsinki, Finland, [5]Kuopio Res. Inst. of Exercise Med., Kuopio, Finland, [6]Univ. of North Carolina, Chapel Hill, NC, [7]NIH, Bethesda, MD

Abstract:

Variants and genes at type 2 diabetes (T2D) and related trait genome wide association study (GWAS) signals may impact T2D risk via skeletal muscle, a primary insulin responsive tissue. We previously generated *in vivo* single nucleus (sn-)multi-omics (RNA+ATAC) profiling of 286 skeletal muscle biopsies from the FUSION Tissue Biopsy Study and found that a subset of GWAS signals colocalize with cell-type specific e/caQTL active in fibro-adipogenic progenitors (FAPs). In a subset of 50 of these individuals with muscle sn-multi-omics data, we derived induced pluripotent stem cell (iPSC) lines from fibroblasts. To investigate mechanisms for the GWAS signals, we differentiated iPSC lines to FAPs, allowing us greater flexibility to dissect the underlying pathways using multiple techniques. These approaches include sn-multi-omics and CUT&Tag, a sensitive method to analyze genome-wide histone marks. We previously optimized both techniques on multiple cell types. We hypothesize that FAPs can be differentiated from iPSC lines to investigate the impact of FAP-specific molecular mechanisms on T2D risk and T2D-related traits. Here, we demonstrate that FAPs can be derived from iPSC lines for deep molecular phenotyping. These iPSC-derived FAPs display FAP morphology, expression of FAP marker genes, and loss of expression of pluripotency and hematopoietic markers. Over the course of the differentiation, *PODXL*, a pluripotency marker, drops from 99.1% of cells expressing this gene down to 0.19%, based on flow cytometry analysis. Meanwhile, expression of *NT5E*, a FAP marker, increases from 0.095% of cells up to 98.9%. We performed a time course analysis with this experimental system with 10 independent iPSC lines, all multiplexed into a single cell village, to explore the trajectory from iPSC progenitors to fully-differentiated FAPs. To compare with the *in vivo* results, we took cell-type specific chromatin accessibility peaks for the differentiated FAPs and the *in vivo* skeletal muscle cell types and performed a logistic regression between the two. This comparison reveals the chromatin accessibility profile of the iPSC-FAPs to be very similar to the *in vivo* FAPs, indicating that the differentiated FAPs recapitulate what we observe *in vivo* and that these cell villages can be used as a model system to characterize FAPs. We are currently upscaling the experiment to 50 iPSC-derived FAP samples to investigate stimulatory state specific genetic regulatory effects (e/caQTL) within T2D pathways. We will validate our results by performing CRISPRi knockdown and CRISPRa activation on a subset of nominated loci.

Title: Integrative analysis of a multi-ancestry GWAS of 2.5 million individuals with multiple omics and quantitative trait datasets identifies effector transcripts, proteins, and metabolites for 528 type 2 diabetes loci.

Authors:

R. Mandla[1], K. Lorenz[2], X. Yin[3], O. Bocher[4], L. Southam[5], K. Suzuki[6], K. Hatzikotoulas[7], H. Taylor[8], A. Huerta[1], A. Arruda[9], N. Rayner[10], DIAMANTE Consortium, VA Million Veterans' Program, J. Meigs[11], M. McCarthy[12], A. Mahajan[12], C. Spracklen[13], M. Boehnke[14], M. Vujkovic[2], J. I. Rotter[15], B. Voight[2], E. Zeggini[16], A. Morris[17], J. Mercader[1], Type 2 Diabetes Global Genomics Initiative; [1]Broad Inst., Cambridge, MA, [2]Univ. of Pennsylvania, Philadelphia, PA, [3]Nanjing Med. Univ., Nanjing, China, [4]Helmholtz Zentrum Muenchen, Neuherberg, Germany, [5]German Res. Ctr. for Environmental Hlth., Neuherberg, Germany, [6]The Univ. of Manchester, Manchester, United Kingdom, [7]Helmholtz Zentrum München (GmbH), Neuherberg, Germany, [8]Natl. Human Genome Res. Inst., Bethesda, MD, [9]Helmholtz Zentrum München, Neuherberg, Germany, [10]Helmholtz Zentrum München, Munich, Germany, [11]Massachusetts Gen Hosp, Boston, MA, [12]Univ. of Oxford, Oxford, United Kingdom, [13]Univ. of Massachusetts, Amherst, MA, [14]Univ. of Michigan, Ann Arbor, MI, [15]Lundquist Inst., Harbor-UCLA Med Ctr, Torrance, CA, [16]Helmholtz Zentrum Munich, Neuherberg / Munich, Germany, [17]Univ. of Manchester, Manchester, United Kingdom

Abstract:

While hundreds of genetic loci associated with type 2 diabetes (T2D) have been identified, it remains a challenge to discern the functional relevance of these association signals. To understand the mechanisms of T2D associated loci, we conducted the largest, most ancestrally diverse genome-wide association study (GWAS) meta-analysis (428,452 cases and 2,107,149 controls). We then performed colocalization analyses with expression quantitative trait loci (eQTL) datasets from 6 tissues relevant for T2D, circulating plasma protein quantitative trait loci (pQTL), metabolite quantitative trait loci (metabQTL), and 43 cardiometabolic trait GWAS.

We identified 1,289 genome-wide significant (P<5x10-8) index variants comprising 859 regions. 528 of the 859 regions demonstrated evidence of colocalization (posterior probability >0.8) with at least one molecular QTL dataset. Among the eQTL colocalizations, the majority were tissue-specific, with pancreatic islets having the highest percentage of tissue-specific colocalizations (50%). For example, an index variant associated with higher T2D risk colocalized with higher expression of SCTR within pancreatic islets only. SCTR, a secretin receptor, is expressed primarily in pancreatic islets and is in the same protein family as GLP1R, a known T2D drug target. SCTR has been previously associated with T2D, including a protective, missense variant prominent in East Asian populations and we observed lower SCTR expression in individuals with T2D.

Notably, 32 index variants showed evidence of simultaneous colocalization with an eQTL, pQTL, quantitative trait association, and metabQTL, allowing for in-depth functional characterization. For instance, one index variant colocalized with expression levels of WIPI1 specifically in pancreatic islets and ATF6B protein levels in plasma, two genes known to function in autophagy and endoplasmic reticulum stress pathways and that have been linked with diabetes risk. This variant additionally colocalized with high-density lipoprotein levels in plasma and alkaline phosphatase (ALP) GWAS signals, which may be driven by the role ATF6 plays in lipid metabolism and ALP activity regulation.

Overall, this work identified putative molecular pathways underlying T2D, allowing functional follow up of and gene prioritization of 528 (61%) of the most extensive list of T2D loci to date and may inform future therapeutic strategies. Further causal inference analyses are needed to confirm causal pathways of T2D loci, and additional ancestrally diverse molecular QTL datasets from relevant tissues will be needed to dissect mechanistic pathways of population-specific T2D signals.

Title: Hypothalamic variant-to-function analysis of the key childhood obesity chr12q13 GWAS locus implicates rs7132908 as a causal variant embedded within the 3'UTR of *FAIM2.*

Authors:

S. Littleton[1,2], K. Trang[2], C. Volpe[1,2], K. Cook[2], N. DeBruyne[1,2], J. Maguire[2], M. Hazuga[1,2], K. Boehm[2], J. Pippin[2], M. Pahl[2], S. Grant[1,2]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Children's Hosp. of Philadelphia, Philadelphia, PA

Abstract:

Our childhood obesity GWAS with the Early Growth Genetics consortium found one of the most significant loci, namely on chr12q13 and tagged by rs7138803, was more pronounced in children than adults. As the top adult obesity/BMI loci have received the most attention thus far, this locus has been largely overlooked. Our chromatin accessibility and conformation capture data generated in various cell types, including human embryonic stem cell (ESC)-derived hypothalamic arcuate neurons and primary astrocytes, and genetic fine-mapping independently implicated proxy rs7132908 as a candidate causal SNP. This non-coding variant is embedded within the 3' untranslated region of *FAIM2*. We hypothesized it resides in a *cis*-regulatory element influencing expression of effector genes within its topologically associating domain (TAD). Due to evidence that BMI GWAS loci collectively confer an effect via neuronal processes related to feeding behavior, we sought to characterize the effects of the rs7132908 risk allele in neural cell types. Given that primary astrocytes are a relatively tractable model, we performed luciferase reporter assays in this setting and revealed that the non-risk allele region increased luciferase expression driven by the *FAIM2* promoter 1.75 fold, while the risk allele decreased expression 0.53 fold. In contrast, there were no significant expression changes with additional promoters implicated by chromatin capture: *LIMA1* and *RACGAP1*. We also engineered human ESC lines homozygous for each allele and then differentiated them to a heterogeneous model of hypothalamic arcuate neurons, which are responsible for regulating appetite. Next, we performed bulk or single nucleus RNA-seq at 3 timepoints: ESCs (day 0), neural progenitors (day 14) and hypothalamic arcuate neurons (day 40); when considering all 21 genes within the TAD, 0 genes in ESCs, 5 genes in neural progenitors and 7 genes in hypothalamic arcuate neurons were differentially expressed. Consistent with the luciferase results, *FAIM2* was the only TAD gene down-regulated with the risk allele in both the neural progenitors (log2 fold change -0.97) and hypothalamic arcuate neurons (log2 fold change -1.58). As FAIM2 functions in neural development, plasticity and protection from apoptosis, we also assessed cellular differences due to genotype via single nucleus RNA-seq on differentiation day 40. We observed a marked difference, with the non-risk allele cells comprising 62% neurons and the risk allele cells comprising only 3% neurons. We conclude that rs7132908 is an obesity causal variant regulating *FAIM2*, and potentially other genes, that function in neuron development and/or survival.

Title: A GWAS-in-a-Dish using isolated human islets identifies diabetes risk variants associated with insulin secretion and content

Authors:

S. Sharp[1], C. E. Ellis[2], H. Sun[1], J. Lyon[2], V. Rajesh[1], N. Smith[2], S. Thaman[1], A. Spigelman[2], A. Bautista[2], J. E. Manning Fox[2], P. E. MacDonald[2], A. L. Gloyn[1]; [1]Stanford Sch. of Med., Palo Alto, CA, [2]Alberta Diabetes Inst., Edmonton, AB, Canada

Abstract:

Introduction: Translation of genetic association signals to biological insights remains a significant bottleneck for complex traits. Coupling human genetics with cellular phenotyping, a so called "GWAS-in-a- Dish" could inform on cellular mechanisms of disease risk. Pancreatic islets, a critical tissue for type 1 and type 2 diabetes are responsible for making and secreting insulin which regulates blood glucose. Genetic evidence supports a major role for this tissue in mediating disease risk. We used human islets from deceased donors and coupled cellular phenotyping with multi-omics to determine the impact of genetic variation on insulin secretion and content. We integrated human islet expression (e) and chromatin (ca) QTLs to identify effector transcripts driving pancreatic islet-cell dysfunction.
Methods: Human islets were isolated from deceased donors (n=289). Donor acinar tissue was array genotyped (Omni2.5Exome) and imputed (TOPMED). Hand-picked islets were processed for RNA-seq and ATAC-seq and exposed to low (1mM) and high glucose conditions (16.7mM) and insulin secretion and content measured by ELISA. We log-normalized Secretion Index (high/low secretion) and percent secretion (secretion/content). We examined relationship with sex, age, BMI, ancestry, diabetes genetic risk scores and partitioned scores. We performed GWAS using REGENIE adjusting for sex, age, BMI, population structure and excluding rare variants (MAF<1%). We tested variants with existing evidence of association with diabetes, glycemic traits and eQTL/caQTLs.
Results: Insulin secretion at high glucose was reciprocally associated with BMI ($p=0.026$) whilst secretion decreased with increased age ($p=0.014$). Insulin secretion was lower in female donors ($p=0.029$) whilst those of African ancestry had decreased insulin content ($p=0.041$). Type 2 diabetes partitioned risk for beta cell function associated with both lower insulin secretion and content ($p<0.001$, $p=0.005$). Islet eQTL/caQTL co-localized variants associated with diabetes risk adjacent to FAM234A altered both insulin content and secretion ($p<0.001$). GWAS revealed novel loci associated with insulin secretion (n=5), secretion index (n=8) and insulin content (n=19) for follow up at suggestive significance ($p<1x10-6$).
Conclusion: We have coupled cellular phenotyping of human pancreatic islets with genetic and epigenomic characterization to link genetic variation to regulatory elements, gene expression and insulin secretion. We demonstrate the feasibility of this approach in human islets and show that genetic determinants of in vivo measures of beta-cell function are associated with in vivo function.

Title: eQTL mapping in fetal-like pancreas reveals novel insights into diabetes etiology

Authors:

J. Nguyen[1], T. Arthur[2], K. Fujita[1], B. Salgado[1], M. Donovan[2], iPSCORE Consortium, H. Matsui[2], A. D'Antonio-Chronowska[2], M. D'Antonio[2], K. Frazer[2]; [1]Univ. of California, San Diego, San Diego, CA, [2]Univ. of California, San Diego, La Jolla, CA

Abstract:

The impact of early life events on the susceptibility to disease is an emerging area of research. In the pancreas, adverse events during fetal development can result in insulin resistance, impaired glucose metabolism, and the loss of beta cell function. These factors contribute to an increased risk of developing type 2 diabetes in adulthood. While current expression quantitative trait loci (eQTL) datasets have been instrumental in characterizing genetic variants associated with disease, they often overlook the crucial genetic contributions that occur during fetal development since they primarily focus on adult tissues. In this study, we introduce a comprehensive eQTL resource for the pancreas at the fetal stage for gene expression and alternative splicing. We utilized RNA-seq data from 106 iPSC-derived pancreatic progenitor cells (iPSC-PPCs) obtained from individuals with whole genome sequences. Through colocalization analysis with eQTLs in adult islets and whole pancreases, we discovered 2,426 eQTL associations that were unique to the fetal-like iPSC-PPCs and 4,636 that were unique to the adult. Notably, we found strong evidence linking 228 of these eQTLs to genome-wide association studies (GWAS) loci associated with type 2 diabetes, type 1 diabetes, and obesity-related traits. Among the fetal-associated type 1 diabetes loci were *CDC37L1-DT* and *MEG3*, neither of which have been previously explored in the fetal context. Additionally, we investigated the functional differences of genetic variants between the fetal-like and adult pancreas and identified 1,139 eQTLs that were active in both stages but associated with different eGenes. For instance, cholesterol-associated rs138349 impacted *ADSL* expression during pancreas development but affected *ST13* expression in adulthood. Overall, we uncovered new insights into the roles of disease risk variants which may operate specifically during early development or in adult stages. We also discovered pancreatic eQTLs that have altered regulatory functions between early fetal and adult stages. Further, we provide a valuable resource for interpretation of regulatory variants in early pancreatic progenitor cells. Currently, we are mapping and integrating genome-wide chromatin accessibility QTLs (caQTLs) in iPSC-PPCs for further understanding of the epigenetic mechanisms underlying diabetes risk and obesity traits. By characterizing causal variants across multiple contexts, we can begin to delineate the etiology of disease risk with a finer resolution.

Title: Identification of causal genes for nonalcoholic fatty liver disease using single-cell eQTL analysis

Authors:

S. Hong, K-S. Suh, W. Kim, M. Choi; Seoul Natl. Univ., Seoul, Korea, Republic of

Abstract:

Background and aims Nonalcoholic fatty liver disease (NAFLD) is a liver disease associated with metabolic syndrome with increasing medical and socioeconomic burdens. Lack of effective treatment drugs urges the discovery of novel therapeutic targets. This study utilizes single-cell expression quantitative trait loci (sc-eQTL) based analysis to discover biomarkers and therapeutic targets of NAFLD. **Methods** Liver biopsy samples obtained from 23 control individuals and 25 NAFLD patients were subjected to single nucleus RNA-sequencing (snRNA-seq). DNA samples obtained from the same participants were genotyped by low coverage whole genome sequencing. snRNA-seq profiles of the NAFLD livers were analyzed using various bioinformatics tools. sc-eQTL were mapped via poisson mixed effects model. sc-eQTLs were tested for interaction with various cell level phenotypes. Colocalization with genome-wide association studies (GWAS) were conducted. **Results** A total of 250K cells were detected, including hepatocytes and various non-parenchymal cells. Differential gene expression analysis and intercellular interaction analysis revealed cell type-specific changes in NAFLD. Gene modules discovered by network analysis were associated with distinct pathophysiology of liver cells. Multiple sc-eQTL signals were detected and replicated. Disease-interacting sc-eQTLs were identified. Numerous loci colocalizing with NAFLD GWAS were characterized. **Conclusions** We present transcriptomic profile of NAFLD in a single-cell resolution. sc-eQTL analysis identified NAFLD-associated genes and their regulatory variants in relevant cell types. The role of putative regulatory genes and variants will be subjected to functional validation.

Session 101: Going beyond DNA sequencing to identify disease causes

Location: Conv Ctr/Ballroom A/Level 3

Session Time: Saturday, November 4, 2023, 10:30 am - 12:00 noon

Title: Expanding the clinical reach of RNA sequencing: Evaluating testing outcomes of concurrent germline DNA and RNA genetic testing in a cohort of 43,000 individuals undergoing hereditary cancer testing

Authors:

R. Karam, C. Horton, L. Hoang, H. Zimmermann, C. Young, J. Grzybowski, K. Durda, D. Burks, M. Richardson, S. Harrison, E. Chao; Ambry Genetics, Aliso Viejo, CA

Abstract:

Importance: Precision medicine has become mainstreamed in health care, and its utility is especially evident in the field of oncology. Personalized surveillance, prophylaxis, and cancer treatment options for individuals with hereditary cancer predisposition are informed by results of germline genetic testing. Improvements to genomic technology, such as the availability of RNA sequencing, may increase identification of individuals eligible for personalized interventions by improving the accuracy and yield of germline testing.
Objective: Assess the cumulative impact of paired DNA and RNA testing on detection of disease-causing germline genetic variants and resolution of variants of uncertain significance (VUS) and the resulting clinical implications.
Design, setting, and participants: Paired DNA and RNA sequencing was performed on 43,524 individuals undergoing germline testing for hereditary cancer indication at a single diagnostic laboratory from March 2019 through April 2020. Demographics, clinical data, and test results were curated as samples were received and changes to variant classification were assessed over time.
Main Outcomes and Measures: We assessed the overall results by variant type, the effect of RNA evidence on variant classification, and the corresponding impact on cancer risk management. Increase in diagnostic yield, decrease in VUS rate, and difference in positive and negative predictive values were also evaluated.
Results: Variant classification was impacted in 549 individuals. Medically significant upgrades were made in 97 individuals, including 70 individuals who had a variant reclassified from VUS to Pathogenic/Likely Pathogenic (P/LP) and 27 individuals who had a novel deep intronic P/LP variant that would not have been detected using DNA sequencing alone. This corresponded to eligibility for increased surveillance in 14.2% (n=78) and for surgical options in 5.8% (n=32) of RNA impacted individuals. We found that 17.1% (93 of 545) of P/LP splicing variants were dependent on RNA evidence for classification and 71.1% (312 of 439) of existing splicing VUS were resolved by RNA evidence. The evidence generated during this one-year study period was then applied to individuals with DNA-only testing and have led to reclassifications in 7,602 individuals.
Conclusions and Relevance: The ability to perform RNA sequencing concurrently with DNA sequencing represents an important advancement in germline genetic testing by improving detection of novel variants and classification of existing variants.

Title: Diagnostic Utility of Genome-wide DNA Methylation Screening to Identify Molecular Causes of Genetically Unsolved Developmental and Epileptic Encephalopathies

Authors:

C. Laflamme[1], C. Rastin[2], E. Almanza-Fuerte[1], E. Bonkowski[1], H. Pennington[1], S. Sengupta[1], S. Russ-Hall[3], A. Schneider[3], U. of Washington Center for Rare Disease Research[4], M. Galey[4], D. M. Nyaga[5], N. Lieffering[5], H. McConkey[2], J. Kerkhof[2], M. Levy[2], R. Relator[2], T. Sagie[6], Z. Wang[1], S. Berkovic[3], L. G. Sadlier[5], D. E. Miller[4], I. E. Scheffer[3], B. Sadikovic[2], H. C. Mefford[1]; [1]St. Jude Children's Res. Hosp., Memphis, TN, [2]London Hlth.Sci. Ctr., London, ON, Canada, [3]Univ. of Melbourne, Melbourne, Australia, [4]Univ. of Washington, Seattle, WA, [5]Univ. of Otago, Wellington, New Zealand, [6]Wolfson Med. Ctr., Holon, Israel

Abstract:

Sequence-based genetic testing identifies causative genetic variants in only about half of individuals with developmental and epileptic encephalopathies (DEEs). Aberrant changes in DNA methylation have been implicated in various neurodevelopmental disorders but have been largely unstudied in the DEEs. Rare epigenetic variation ("epivariants"), often coupled with rare and difficult-to-detect genetic variants (i.e. CG-rich repeat expansion), can drive disease by modulating gene expression at a single locus. Additionally, the DNA methylation statuses of multiple epigenetic loci can collectively serve as distinct "episignatures" for over 100 monogenic disorders as part of an EpiSign™ clinical diagnostic test. Here, we sought to interrogate the diagnostic utility of genome-wide DNA methylation array analysis on peripheral blood-derived DNA samples from a cohort of 534 individuals with DEEs who have previously undergone extensive screening (e.g. gene panel, microarray, exome sequencing) but remain etiologically unsolved. We identified rare differentially methylated regions (DMRs) and explanatory DNA methylation signatures to collectively identify causative genetic etiologies in 8 individuals. We then used long-read sequencing on the Oxford nanopore platform to identify candidate DNA defects underlying rare DMRs, including three CG-rich repeat expansions affecting gene expression in patient-derived fibroblasts (n=3), two copy number variants, and one balanced translocation. We also identified pathogenic sequence variants associated with episignatures, most of which were missed when the individual had undergone exome sequencing. Importantly, the genes for which EpiSign testing led to a diagnosis are broadly considered neurodevelopmental disorder (NDD) genes rather than epilepsy genes, emphasizing the contribution of NDD genes to DEE. We also redefine and examine the robust episignature for the DEE gene *CHD2* on the 850K metharray (n=29) and whole-genome bisulfite sequencing (n=11) to investigate potential insights into *CHD2* pathophysiology. Our study demonstrates the utility of genome-wide DNA methylation analysis in identifying the etiology of the DEEs, both for the discovery of genetic variants to diagnose unsolved cases and to increase our understanding of the basic mechanisms underlying disease.

Title: Analysis of RNAseq from over 5,000 individuals in the 100,000 Genomes Project identifies new potential diagnoses for patients with rare disease

Authors:

J. Lord[1], C. Jaramillo Oquendo[1], N. McGinness[2], A. Ho[2], C. Odhams[3], T. James[4], M. Ross[4], L. Hoa[2], G. Elgar[2], D. Baralle[5], Genomics England RNAseq Steering Committee, Genomics England Research Consortium; [1]Univ. of Southampton, Southampton, United Kingdom, [2]Genomics England Ltd., London, United Kingdom, [3]Genomics England, London, United Kingdom, [4]Illumina Cambridge Ltd., Cambridge, United Kingdom, [5]Univ. of Southampton, Faculty of Med., Southampton, United Kingdom

Abstract:

Background: Diagnosis of rare disorders has been revolutionised by whole exome and genome sequencing (WES/WGS), but even with WGS, around half of patients' disorders remain unsolved. Interpretation of non-coding variation, which can affect splicing and gene regulation, has lagged behind coding variation, and improvement in this area will improve diagnostic yields. Here we present analyses of whole blood based transcriptome sequencing data from over 5,000 probands with rare disorders that underwent WGS in the 100,000 Genomes Project, but did not receive a molecular diagnosis.
Methods: We used multiple tools to identify candidate expression (OUTRIDER via DROP) and splicing (LeafCutterMD, FRASER2 via DROP) outliers. PanelApp gene panels relevant to each proband's phenotypes were applied and cross-referencing with WGS data was utilised to flag candidate diagnostic events.
Results: We will present the full 100,000 Genomes Project RNAseq cohort, including information on disease classes and demographics. OUTRIDER identified on average 3.1 expression outliers per proband genome wide, with 8% of the cohort possessing an outlier event in a disease gene in a relevant gene panel. LeafCutterMD initially identified over 150 splicing outliers per proband, but stringent filtering reduced this to 4.0, consistent with FRASER2's number of identified candidates, although the overlap of events found by the two tools was low. Analysis and review of candidates is ongoing, but we estimate around 21% of the cohort has at least one significant outlier in a disease relevant gene. Interesting examples of diagnostic discoveries will be identified and highlighted.
Discussion: Although work is ongoing, we estimate at least 20% of the sequenced cohort will have a diagnostic candidate, which is likely to bring a significant uplift to diagnostic rates.

Title: Enhanced genetic diagnosis of neurological disorders through fibroblast-to-neuron transdifferentiation and RNA Sequencing: A comprehensive workflow for clinical applications

Authors:

S. Li[1], S. Zhao[1], J. C. Sinson[1], A. Bajic[1,2], J. A. Rosenfeld[1], M. B. Neeley[3], M. N. Pena[1], K. C. Worley[1], L. Burrage[1], M. W. Hubshman[1], S. Ketkar[1], W. J. Craigen[1,4], G. D. Clark[5,4], S. Lalani[1,4], P. Moretti[1], K. Machol[1,4], J. Sheppard[2], M. T. Nguyen[2], A. Khoramnia[1], P. P. Hernandez[1], S. C. Nagamani[1], Z. Liu[2,3,5], Undiagnosed Diseases Network, B. Lee[1,4], C. M. Eng[1,6], P. Liu[1,6]; [1]Dept. of Molecular and Human Genetics, Baylor Coll. of Med., Houston, TX, [2]Jan and Dan Duncan Neurological Res. Inst. at Texas Children's Hosp., Houston, TX, [3]Graduate Program in Quantitative and Computational BioSci.s, Baylor Coll. of Med., Houston, TX, [4]Texas Children's Hosp., Houston, TX, [5]Dept. of Pediatrics, Section of Neurology, Baylor Coll. of Med., Houston, TX, [6]Baylor Genetics, Houston, TX

Abstract:

Despite the increasing rates of molecular diagnosis with clinical exome and whole genome sequencing, about half of the patients remain undiagnosed. As a complement to DNA sequencing, RNA sequencing (RNA-Seq) has recently been successfully used to detect aberrant transcripts, increasing molecular diagnosis yield. However, a major obstacle to the application of diagnostic RNA-Seq is the limited expression of tissue-specific transcripts in clinically accessible tissues (CATs). For example, the successful detection of neuron-specific genes is important for the diagnosis of Mendelian disorders that manifest with neurological phenotypes. These genes, unfortunately, may not be adequately expressed in CATs. In this study, we performed fibroblast-to-neuron transdifferentiation for patients with primary neurological disorders from the Undiagnosed Diseases Network (UDN) project, resulting in enhanced detection of neuron-specific genes, exons, and junctions. The application of the analysis workflow resulted in the molecular diagnosis of 19 patients (27% of the total cohort) with various neurological disorders. The molecular diagnostic findings included 5 with aberrant expression, 8 with aberrant splicing, and 6 with combined aberrant expression and splicing. Notably, neuron induction assisted in the molecular diagnosis of 6 patients (31.6%). The diagnosis was facilitated by the increased expression levels of specific neuronal genes in five patients (*ITPR1*, *DCX*, *CACNA1A*, *PIZEO2*, and *POLR3A*), while in one patient (*MBD5*), the detection of neuron-specific isoforms proved instrumental in the diagnosis. The diagnostic yield was particularly notable in patients with brain malformations (40%, 12/30), intellectual disabilities (33%, 15/46), and epilepsy (32%, 10/31), emphasizing the potential of the proposed approach to enhance diagnosis in these specific patient cohorts. Based on this experience, we developed an innovative and simplified workflow tailored to the requirements of clinical diagnostic laboratories. Emphasizing simplicity, cost-effectiveness, and timely execution while ensuring reliable reproducibility, the workflow incorporated two crucial quality control checkpoints. These checkpoints involved an early-stage qPCR assay to exclude samples with insufficient lentivirus infection and a final-stage evaluation of RNA-Seq data to ensure adequate transdifferentiation and identify potential chromosomal abnormalities. We believe that the innovative methodologies outlined in this work can be generalized as a diagnostic tool to improve the transcriptome evaluation of individuals with genetic disorders.

Title: DNA Methylation Signatures for sub-phenotype characterization in Kabuki Syndrome.

Authors:

M. Miron Toruno[1], M. Vladoiu[2], N. Bekheirnia[2], P. Liu[2], M. C. Braun[3], A. W. Reynolds[1], M. R. Bekheirnia[2]; [1]Baylor Univ., Waco, TX, [2]Baylor Coll. of Med., Houston, TX, [3]Texas Children's Hosp., Houston, TX

Abstract:

Kabuki Syndrome (KS) (OMIM 147920) is a Mendelian disorder of the epigenetic machinery characterized by multiple inherited anomalies and developmental delays. Several studies have reported the incidence (12%-43%) of congenital anomalies of the kidneys and urinary tract (CAKUT) in KS diagnosis, with significant clinical variability in the manifestation and severity of CAKUT phenotypes. However, the molecular mechanisms contributing to this feature remain largely unexplored. In this study, we aimed to characterize the epigenomic signature of KS patients with reported CAKUT sub-phenotypes. Using the Infinium methylation EPIC array, we performed genome-wide profiling of >850,000 methylation sites in peripheral blood samples from 7 patients with CAKUT sub-phenotypes, as well as 71 healthy samples serving as controls. We identified a distinct epigenomic profile between KS patients with CAKUT sub-phenotypes and healthy patients, consisting of 597 differentially methylated CpG sites and 74 differentially methylated regions. Genetic and functional annotation revealed the presence of significant CpG sites in *ILF2*, *TMEM181*, *DSCAML1*, and *PLEKHG4B*, a series of genes involved in renal function and development. Moreover, CpGs with methylation signatures were found in genes related to kidney failure, nephrotic syndrome, kidney size, and CHARGE syndrome (*HDAC4*, *SNYNPO2*, *BMPR1A*, and *CHD7*). To the best of our knowledge, this study provides the first insight into the epigenetic landscape underlying CAKUT sub-phenotypes in KS patients. By elucidating the epigenomic alterations associated with renal and urinary diseases we aim to increase the knowledge of disease pathogenesis and provide possible therapeutic targets. Furthermore, the identified epigenomic signatures may be used as markers for diagnosis and prognosis, enabling early identification of sub-phenotypes in KS patients.

Title: Transcriptome profiling of pediatric extracranial solid tumors enables rapid, low cost diagnostic classification.

Authors:

K. Opoku[1], T. Santiago[2], N. Bhakta[2], T. Alexander[1], J. Wang[1]; [1]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [2]St Jude Children's Res. Hosp., Memphis, TN

Abstract:

Approximately 85% of pediatric solid tumors occur in Low- and Middle-Income Countries (LMIC) where diagnostic tools routinely used in High Income Countries (HIC) including cytogenetics, immunohistochemistry, Fluorescence In Situ Hybridization (FISH) and traditional short read sequencing platforms for cancer diagnosis are often unavailable or incomplete. Pediatric tumors require very specific diagnosis of types and subtypes to inform treatment decisions and prognostication. A reliable, high accuracy, cost effective diagnostic tool can help bridge the diagnostic gap and improve cancer outcomes in resource limited settings. We investigate the feasibility of using whole transcriptome sequencing on Oxford Nanopore Technologies' sequencing platforms and a composite machine learning model as a more affordable alternative approach for solid tumor diagnosis in resource limited settings. Whole transcriptome cDNA sequencing was performed on a heterogenous mix of 137 FFPE pediatric solid tumor samples on Oxford Nanopore Technologies' MinION sequencing platform. A composite machine learning model was then used to classify reads into clinically actionable tumor types and subtypes by Gene Expression Profile (GEP) mapping.Overall, 95.6% of tumor types and 92.3% of tumor subtypes were correctly classified. 71.5% of tumor specimens had prediction probabilities >0.8 and were classified with 100% accuracy. 88.3% of tumor specimens had prediction probabilities > 0.6 and were classified with 99.2% accuracy. Additionally, *FOXO1* fusion status, an important determinant of treatment and prognosis for rhabdomyosarcoma (RMS) was predicted accurately for 97.4% (37/38) RMS samples. Tumor types for which we had more samples were classified more accurately with higher prediction probabilities. Whole transcriptome sequencing of FFPE pediatric solid tumor specimens has the potential to provide clinical classification of both tissue lineage and core genomic classification for certain pediatric tumors. Expansion, refinement, and validation of this approach is warranted to explore whether this technology could be part of the solution to address the diagnostic limitations in LMICs.

Session 102: Novel tools and data resources for genetic data analysis

Location: Conv Ctr/Ballroom B/Level 3

Session Time: Saturday, November 4, 2023, 10:30 am - 12:00 noon

Title: gnomAD v4: building scalable frameworks to process and quality control 730,913 exomes and 76,156 genomes.

Authors:

J. Goodrich[1,2], K. Laricchia[1,2], M. W. Wilson[1,2], K. R. Chao[1,2], J. M. Fu[1,2], G. Tiao[1,2], Q. He[1,2], D. Marten[1,3], T. Poterba[1,2], C. Vittal[1,2], W. Lu[1,2], S. Baxter[1], S. Chapman[1,2], C. Cusick[1], P. W. Darnowsky[1,2], L. Gauthier[1], L. Gruenschloss[4], R. Grant[1,2], S. Jahl[1,2], M. Solomonson[1,2], C. Stevens[1,2], gnomAD Project Consortium, D. King[1,2], D. G. MacArthur[4,1], M. E. Talkowski[2,1], B. M. Neale[1,2], A. O'Donnell-Luria[1,2,3], M. J. Daly[1,2,5], H. L. Rehm[1,2], K. E. Samocha[1,2], K. J. Karczewski[1,2]; [1]Broad Inst. of MIT and Harvard, Cambridge, MA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]Boston Children's Hosp., Boston, MA, [4]Garvan Inst. of Med. Res., Sydney, Australia, [5]Inst. for Molecular Med. Finland, Helsinki, Finland

Abstract:

The Genome Aggregation Database (gnomAD) is a publicly available large-scale aggregate dataset of human genetic variation that aids the research and clinical communities with variant interpretation. We present a substantial update (v4) that combines 730,913 exomes with 76,156 genomes from v3, resulting in a total of 807,069 humans with sequence data mapped to GRCh38. This dataset includes 138,536 non-European samples and 416,553 exomes from the UK Biobank, representing a nearly six-fold size increase compared to either of the combined gnomAD v2 and v3 datasets. At these scales, data processing and quality control that retain all quality information require novel analytical paradigms, which we present here, as well as a demonstration of scaling these approaches to a jointly called dataset of nearly one million exomes.

We implemented various updates to our data processing and quality control pipeline to generate this aggregate dataset from sequence data generated on a variety of exome captures, as well as sequencing technologies and sites. To optimize storage and analysis, we combine gVCFs into a Hail VariantDataset (VDS), a format implementing the Scalable Variant Call Representation. We developed a novel contamination metric (CHARR) that calculates the mean reference allele balance of homozygous alternate genotypes from the VDS, speeding up computation of contamination metrics by three orders of magnitude. We built a robust pipeline for sex karyotype imputation in the presence of multiple exome capture platforms, using per-platform distributions of sex chromosome ploidy estimates derived from the depth of reference and/or variant data coupled with Gaussian mixture modeling and extensive manual curation. To enable kinship coefficient estimation on this scale of sample pairs, we developed cuKING, a CUDA-based KING relatedness estimator, speeding up a quadratic step of the pipeline by several orders of magnitude. We improved sample outlier detection by creating an ensemble method that combines our previous ancestry PC regression-based approach with the comparison of samples against their 50 nearest neighbors in PC space, ensuring the inclusion of maximally diverse samples.

The public release of gnomAD v4, which is designed to be of high utility to research and clinical groups, provides downloads of allele frequencies and quality metrics for over 780 million variants, accompanied by a gnomAD browser update. The release of gnomAD v4 with its expanded dataset and improved analysis pipeline provides the research and clinical communities with a comprehensive and valuable resource to advance our understanding and interpretation of human genetic variation.

Title: Tabular Encoding of Rare-Variant Genotype Data for Enabling Efficient Random Access in Memory and Analysis of Rare Allele Sharing

Authors:

B. Guo[1], R. Laboulaye[1], V. Borda Pua[1], D. Veliz-Otani[1], S. Zollner[2], R. Hernandez[3], T. D. O'Connor[1]; [1]Univ. of Maryland Sch. of Med., Baltimore, MD, [2]Univ. of Michigan Sch. of Publ. Hlth., Ann Arbor, MI, [3]Univ. of California, San Francisco, San Francisco, CA

Abstract:

Rare variants (with minor allele frequency < 0.5%) comprise the majority of human genomic variation and represent genetic alterations from recently shared ancestry. The increasing availability of high-density genotype data such as high-depth whole-genome sequencing data makes possible disease association or mapping analyses via rare variants. However, utilizing rare variants in these analyses faces many challenges, including inefficient access to genotypes of rare variants for large datasets, such as TOPMed. Currently, genotype data is mainly stored as the variant call format (VCF, its binary equivalent BCF, or compact equivalents XSI/SAV) and PLINK format (BED/PGEN), which, if decompressed, are essentially site-oriented matrices with each element encoding a genotype call per genome per site. As the majority of segregating sites (97% in TOPMED) are rare variants, explicit matrix encoding of genotype data of rare variants for large sample sizes is computationally inefficient for data input/output and has a high demand for memory for random access, especially for sample-oriented calculations, restricting fast, large-scale rare variant analysis. To address this issue, we propose a simple, in-memory, tabular encoding of the rare variant genotype data (phased and imputed) with each row representing the genotype call of a rare allele with a 3-tuple (position, genome index, allele index). This encoding avoids the explicit representation of common alleles which account for >99 percent of the full matrix encoding. Tabular encoding of all rare SNPs (73 million variants with allele frequency <0.001, including multiallelic sites) of 184K WGS samples from TOPMED freeze 10b only needs 4.7 Gigabytes of memory which is highly compact compared with matrix encoding (>3 Terabytes decompressed). The compact representation allows for fast data loading from disk (seconds for whole chromosome 1 compared to hours from BCF), reordering by samples/genomes, and, more importantly, in-memory random access to efficiently calculate (genome) pairwise rare variants sharing (*Jaccard* index calculation in our new tool *gtencode* is over a hundred times faster than *bcftools roh* and *plink2 sample-diff* calculation). Our initial implementation of this encoding has used the Rust language for memory safety and easy parallelization and the arrow/parquet format for cross-language accessibility. Ongoing work will focus on functionality associated with rare variants, such as sample pairwise rare variant sharing statistics and the correlation of rare variants with identity-by-descent segments and local ancestry, and provide an interface for other popular languages such as Python and R.

Title: At scale enhancements to NCBI RefSeq Functional Elements, a growing resource for functional discovery beyond genes.

Authors:

C. Farrell, O. D. Ermolaeva, C. L. Wallin, T. D. Murphy; NCBI/NLM/NIH, Bethesda, MD

Abstract:

NCBI provides RefSeq Functional Elements (RefSeqFEs; www.ncbi.nlm.nih.gov/refseq/functionalelements/) for diverse and functionally important non-genic elements in human and mouse, including gene regulatory regions, recombination regions and other regions that have been experimentally validated in the literature. The dataset includes richly annotated sequence records, descriptive records in the Gene database, genomic feature annotation, and interactions between regulatory regions, target genes and each other.

Since our initial publication describing the resource (PMID:34876495), we have tremendously improved content by transitioning from high dependence on manual curation to automated record loading and RefSeq creation processes. The human dataset has grown from <10K functional features to >150K features (June 2023), with continued growth expected over the coming year. We added thousands of new records to the dataset, including regulatory elements validated by massively parallel reporter assays (MPRAs), e.g. ChIP-STARR-seq, or by CRISPR/Cas-based genomic perturbations. Recent qualitative improvements include extractable cell type activity data for annotated features, additional fields for data mining in download files, and new CRISPRi-validated target gene linkages. We increased accessibility by providing annotation on the T2T-CHM13v2.0 genome assembly, with periodic updates of GRCh38 and T2T-CHM13v2.0 annotations and the RefSeqFE track hub. Moreover, we improved clinically relevant content by adding features for functionally validated regulatory variants, e.g. MPRA-validated melanoma-associated variants from PMID:36423637, and we curated regulatory elements for genes associated with coronavirus biology and marked up biological regions that overlap known COVID-19-associated variants.

This increased content and genome coverage makes our freely available resource more useful for discovery of non-coding function beyond genes, and directly links it to the experimental literature. In addition to its utility for basic genomics, this resource can be used for genetic variant mapping, e.g. disease-associated variants can be directly mapped to RefSeqFE biological regions, regulatory interactions can be used to determine target genes, and features representing MPRA- or CRISPRi-validated SNPs can be directly used to assign variant function. Further details on RefSeqFE data access, improvements and uses will be presented, as well as overlap analyses with variant datasets. We welcome feedback from the genomics research community, with an aim to optimize RefSeqFEs as a reference resource for non-genic regions.

Title: A unified computing environment for genomics data storage, management, and analysis: NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL).

Authors:

S. Mosher[1], M. C. Schatz[1], A. Philippakis[2], AnVIL Team; [1]Johns Hopkins Univ., Baltimore, MD, [2]Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

Recent years have seen astronomical growth in human genomics. Together with single-cell and functional genomics, electronic medical records and other biomedical data, the field is well-positioned to make great advances in human health. However, the complexity of genomic data sharing, where data is downloaded from centralized datastores for local analysis, is unsustainable and cost prohibitive. Furthermore, housing genomic data across redundant institutional compute infrastructures makes assuring data security and compliant usage of protected data a massive challenge. The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space, or AnVIL (https://anvilproject.org/) was developed to address these and other concerns by providing a unified cloud-based computing environment for genomics data storage, management and analysis. The AnVIL platform inverts the genomics data sharing model by eliminating the need for data movement, which in turn allows for active threat detection and monitoring and provides scalable, shared computing resources for researchers as needed. AnVIL currently provides harmonized access to more than 600,000 genomes from several key NHGRI projects, such as as the CCDG (Centers for Common Disease Genomics), CMG (Centers for Mendelian Genomics), eMERGE (Electronic Medical Records and Genomics), GTEx (Genotype-Tissue Expression Project), T2T (Telomere-to-Telomere), and many more.The platform is built on a set of established components that have been used in a number of flagship scientific projects. The Terra platform provides a compute environment with secure data and analysis sharing capabilities. Dockstore provides standards based sharing of containerized tools and workflows. Jupyter, R/Bioconductor and Galaxy provide analysis environments for users at all skill levels to interactively explore and understand data with thousands of tools available. The Gen3 data commons framework provides data and metadata ingest, querying, and organization. Together, AnVIL provides a collaborative environment for creating, analyzing, and sharing data and analysis workflows for even the largest projects.Long-term, the AnVIL will provide a unified platform for ingestion and organization for a multitude of current and future genomic and genome-related datasets. Importantly, it will ease the process of acquiring access to protected datasets for investigators and drastically reduce the burden of performing large- scale integrated analyses across many datasets to fully realize the potential of ongoing data production efforts.

Title: MSSNG 2023 release: an expanded cloud-based whole-genome sequencing resource for Autism Spectrum Disorder research.

Authors:

S. Scherer[1], B. Thiruvahindrapuram[1], J. L. Howe[1], J. Whitney[1], R. V. Patel[1], M. Bookman[2], O. Hamdan[1], T. Nalpathamkalam[1], G. Pellecchia[1], P. S. Danthi[1], W. Engchuan[1], J. Fuerth[3], H. Ward[3], V. Seifer[4], M. Quirbach[4], M. Mendes de Aquino[1], N. X. Bautista[1], N. Hoang[1], T. Selvanayagam[1], R. K. C. Yuen[1], C. R. Marshall[1], M. Chahrour[5], D. Amaral[6], S. Lewis[7], K. A. Fakhro[8], E. Anagnostou[9], D. Glazer[2], M. Fiume[3], D. M. Hartley[4], B. Trost[1]; [1]The Hosp. for Sick Children, Toronto, ON, Canada, [2]Verily Life Sci., South San Francisco, CA, [3]DNAstack, Toronto, ON, Canada, [4]Autism Speaks, Princeton, NJ, [5]Univ. of Texas Southwestern Med. Ctr., Dallas, TX, [6]Univ. of California, Davis, Davis, CA, [7]The Univ. of British Columbia, Vancouver, BC, Canada, [8]Sidra Med., Doha, Qatar, [9]Holland Bloorview Kids Rehabilitation Hosp., Toronto, ON, Canada

Abstract:

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition whose core characteristics are social-communication deficits, restricted interests, and repetitive behaviors. The genetic architecture of ASD is complex, with ~100 high-impact ASD-linked genes and copy number variants (CNVs) having diagnostic value discovered to date. Both coding and non-coding regions of the genome contribute to ASD susceptibility, meaning that whole-genome sequencing (WGS) data are required to fully understand ASD genetics. The Autism Speaks MSSNG resource is a massive dataset consisting of WGS and phenotype data from families with ASD. Recently, we performed a comprehensive analysis of MSSNG, identifying 134 ASD-associated genes, along with rare ASD-associated variants in ~15% of autistic individuals. We also defined the genomic architecture of ASD in terms of the contributions of many types of rare genetic variants—dominant and recessive sequence-level variants, mitochondrial variants, and various types of structural variants, including chromosomal abnormalities, genomic disorder-causing CNVs, uniparental disomies, and tandem repeat expansions. We also showed that polygenic scores estimated from common variants, while explaining some variance in ASD susceptibility, are not yet particularly informative at the level of individuals or families. Here, we present the 2023 release of MSSNG, which includes WGS data from 2,500 additional individuals, including 1,000 having an ASD diagnosis. MSSNG now contains a total of 14,000 individuals, 6,000 with ASD. A significant portion of the new data is from populations historically underrepresented in ASD genomics studies, in particular from Middle Eastern populations. These new data will allow researchers to better explore ASD genetics in the context of population-specific genetic backgrounds. MSSNG also includes detailed phenotype information, including a wide variety of psychometric assessments (>5,000 measures from >70 tests). To address challenges in interpreting large numbers of phenotypic measures, we have also developed several consensus phenotype measures, each combining several data points into a single easy-to-understand measure. Any qualified researcher with an ASD-related research question can apply for access to MSSNG, with 370 users from 72 institutions in 20 countries having been approved for access to date. MSSNG data can be analyzed via either a cloud-based environment or a web-based portal. In summary, MSSNG combines high-quality WGS data with extensive phenotype information to allow researchers to further our understanding of ASD genetics.

Title: New Insights into Genomic Variation from Long-read Sequencing of >1000 African-American Participants in the All of Us Research Program

Authors:

K. Garimella[1], S. Huang[1], Y. Mostovoy[2], K. Patterson[3], M. Danzi[4], F. Cunial[1], J. Smith[5], B. Shifaw[1], T. Dutka[6], C. Berngruber[7], W. Harvey[3], S. Schwartz[1], E. Laplante[1], M. Mahmoud[8], L. Lichtenstein[1], Y. Wang[9], D. Muzny[10], K. Doheny[11], M. Schatz[11], S. Zuchner[12], S. Levy[13], F. Sedlazeck[10], M. Talkowski[14], E. Eichler[5], A. Ramirez[15], A. Musick[16]; [1]Broad Inst., Cambridge, MA, [2]Mass. Gen. Hosp. / Broad Inst., Winchester, MA, [3]Univ. of Washington, Seattle, WA, [4]Univ. of Miami, Dakota Dunes, SD, [5]Univ of Washington, Seattle, WA, [6]Natl. Inst. Hlth., Bethesda, MD, [7]HudsonAlpha, Huntsville, AL, [8]Baylor Coll. of Med., Houston, TX, [9]Vanderbilt Univ., Nashville, TN, [10]Baylor Coll. Med., Houston, TX, [11]Johns Hopkins Univ., Baltimore, MD, [12]Univ. of Miami Miller Sch. of Med., Miami, FL, [13]Element BioSci.s, San Diego, CA, [14]Massachusetts Gen. Hosp., Boston, MA, [15]NIH, Mclean, VA, [16]All of Us Res. Program, NIH, Bethesda, MD

Abstract:

The All of Us (AoU) research program is a national biobank initiative that aims to collect health and genomic data from at least one million participants for the purposes of advancing precision medicine. While there has been tremendous progress to date (short-read whole genome sequencing [srWGS] of >245,000 participants), recent studies have highlighted large swaths of the human genome and hundreds of biomedically relevant genes are uniquely accessible to long-read sequencing (lrWGS). Here, we present the launch of the AoU lrWGS initiative as a multi-institutional and multi-phase project to capture and openly release genomic variation from ~15,000 diverse individuals with matched srWGS and extensive phenotypic data.

The AoU long-read working group has now completed Phase I of the program and released lrWGS data from 1,027 participants who self-identify as African-American or Black. Here, we describe the novel genomic variants and human disease association discoveries that are only tractable from an integrated genomics dataset of this scale, which includes ~8x Pacific Biosciences lrWGS and ~30x Illumina srWGS. We have completed fully cloud-native processing of all samples. Our analyses to date have generated alignments on both GRCh38 and the telomere-to-telomere CHM13 reference, genome assemblies, short-variant (SNV/indel) callsets using DeepVariant, and structural variant (SV) callsets using a combination of PBSV, PAV, and Sniffles2 on all individuals. These analyses have identified an average of 4-5M SNVs/small indels, ~22.4k SVs per individual. Collectively, we provide an unparalleled resource of 2.2M SVs derived from this diverse population cohort, and through downsampling analyses, estimate current sensitivity of ~80%.

In our ongoing analyses, we resolve variation across repeat expansions, mobile elements, and highly complex genomic regions such as the major histocompatibility complex, LPA, BRCA1/BRCA2, etc. We are further benchmarking complete genome-wide characterization of the value added for SVs captured between these lrWGS and matched srWGS datasets and optimal strategies for future variant interpretation.

Complete data and results are accessible in the All of Us Researcher Workbench, as well as reproducible cloud-native workflows that can be reused in other population-scale long-read sequencing projects. Finally, we will preview Phase 2 work currently underway, which will expand long-read data generation to ~15,000 diverse participants on PacBio, Oxford Nanopore, and Illumina platforms.

Session 103: Understanding kidney traits through genetics

Location: Conv Ctr/Room 145A/Level 1

Session Time: Saturday, November 4, 2023, 10:30 am - 12:00 noon

Title: Proteogenomic Mendelian Randomization identifies putative novel drug targets for chronic kidney disease

Authors:

J. Fadista[1], S. Hu[2], V. Das[3], A. Karihaloo[4], J. Howson[2], D. Shungin[1]; [1]Human Genetics Ctr. of Excellence, Novo Nordisk, Måløv, Denmark, [2]Human Genetics Ctr. of Excellence, Novo Nordisk, Oxford, United Kingdom, [3]Bioinformatics and Data Mining, Digital Sci. and Innovation, Novo Nordisk, Måløv, Denmark, [4]T1D & Kidney Disease, Novo Nordisk Res. Ctr., Seattle, WA

Abstract:

Background/Objectives: Although more than 800 million people worldwide are affected by chronic kidney disease (CKD), kidney pathophysiology remains poorly understood. We integrated proteomics with genetics in a Mendelian randomization (MR) and Bayesian genetic colocalization framework to inform novel potential therapeutic targets for CKD using estimated glomerular filtration rate (eGFR), which is used to diagnose CKD and is a key clinical indicator of renal function.
Methods: We performed GWAS of 2,932 plasma proteins in 48,645 participants from the UK Biobank Pharma Proteomics Project of which 2,026 proteins were instrumented by cis- protein quantitative trait loci (pQTL; P<5x10-8). These cis-pQTLs were compared with published cis-pQTLs from deCODE, ARIC, INTERVAL, and SCALLOP, where we performed a systematic two sample MR study to test the effect of 2,956 individual proteins (with cis-pQTL present in ≥1 study) on eGFR measured in 1,508,659 participants (Liu et al 2022). To assess tissue effects and specificity, we combined results with published cis-expression quantitative trait loci (eQTLs) from kidney (tubular (N=667), glomerular (N=543)). We investigated potential interactions between proteins, by mapping trans-pQTLs (P<5x10-8).
Results: We identified genetically predicted levels of 91 proteins that were associated with eGFR (MR P< 2.5x10-6) with posterior probability of genetic colocalization > 0.7 and consistent effect direction across all protein cohorts. Integration with kidney eQTLs highlighted 26 proteins with same effect direction for protein and gene expression, representing a set of proteins of potential therapeutic interest. Genetically predicted kidney gene expression of additional 388 genes were also causally associated with eGFR (MR P< 2.5x10-6). Among prioritized proteins, UMOD, MANBA, and DPEP1 have been reported as relevant for CKD, while others are novel, including A4GALT which is involved in Fabry disease nephropathy. Trans-pQTL analyses highlighted UMOD as a central hub for kidney function and pointed to CKD-relevant REN pathway.
Conclusion: This largest proteogenomic study to date robustly identified new causal protein-disease links, elucidating the value of integrating genetics with protein levels and gene expression to identify novel putative therapeutic targets.

Title: Use of polygenic risk scores to improve GFR estimating equations in CRIC and MESA.

Authors:

L. Zhou[1], Q. Sun[1], J. Mychaleckyj[2], H. Kramer[3], S. Rich[2], J. Rotter[4], M. Shuey[5], N. Cox[5], NHBLI Trans-Omics for Precision Medicine (TOPMed)Kidney Function Working Group, Chronic Renal Insufficiency Cohort (CRIC), L. Inker[6], N. Franceschini[1], Y. Li[1]; [1]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [2]Univ. of Virginia, Charlottesville, VA, [3]Loyola Univ., Chicago, IL, [4]Lundquist Inst., Harbor-UCLA Med Ctr, Torrance, CA, [5]Vanderbilt Univ. Med. Ctr., Nashville, TN, [6]Tufts Med., Burlington, MA

Abstract:

Glomerular filtration rate (GFR), an estimate of kidney function, is usually not directly measured in clinical practice. Instead, predictive equations of GFR were developed to estimate GFR (eGFR). The 2021 race-free Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equations relate measured GFR (mGFR) to age, sex, and (1) serum creatinine or (2) serum creatinine and cystatin-C. Serum creatinine and cystatin-C have spline forms to account for the changing linear trend (creatinine: knot at 0.7 mg/dl for women and 0.9 mg/dL for men, cystatin-C: knot at 0.8 mg/l). We hypothesized that accounting for the genetic variation in eGFR, through a polygenic risk score (PRS), would improve the accuracy of eGFR compared to mGFR. Using the 2021 CKD-EPI forms for serum creatinine, cystatin-C, age, and sex, we added eGFR PRS as a covariate to the model. We fit a linear model including PRS for (1) creatinine only and (2) creatinine and cystatin-C equations, where eGFR PRS was computed with PRS-CS using European summary statistics from the Chronic Kidney Disease Genetics Consortium (SNPs = 8834748; minor allele count > 10). Using cross validations, we compared our PRS estimating equations with the (1) CKD-EPI 2021 creatinine-only and (2) CKD-EPI 2021 creatinine and cystatin-C equations in the Chronic Renal Insufficiency Cohort (CRIC) study of 1327 African American and White individuals. Performance measures included bias (median of difference (mGFR-eGFR)), precision (IQR of difference), accuracy (median of absolute difference, RMSE relative to mGFR, percent of estimates within 30% of mGFR), and ROC. All performance measures were improved or performed comparably to CKD-EPI 2021 for predicting mGFR. For creatinine-only equations, our PRS eGFR creatinine equation improved the bias by 86% compared to the 2021 CKD-EPI creatinine only equation (0.50 vs 3.66). Our creatinine and cystatin-C equation with eGFR PRS improved bias by 92% compared to CKD-EPI 2021 creatine and cystatin-C equation (0.10 vs -2.02). Chronic kidney disease, defined as GFR < 60, is a highly prevalent disease that affects many clinical decisions in the US. Thus, we are also interested in the performance of the eGFR equations in the cohort with mGFR< 60 versus the cohort with mGFR > 60. For individuals with mGFR< 60, our PRS eGFR equation improved the accuracy by 73% (median absolute difference 1.51 versus 5.63 for CKD-EPI 2021) in the creatinine only equation and 7% (11.4 vs. 14.1) in the creatinine and cystatin-C equation. Further exploration on the form of PRS from diverse populations should further increase the performance. Additionally, we will replicate and validate our results in the MESA cohort.

Title: Clinical exome sequencing efficacy and phenotypic expansions in Congenital Anomalies of Kidney and Urinary Tract (CAKUT)

Authors:

E. Rivera-Munoz[1], X. E. Zhao[1], J. Rosenfeld[1], P. Luna[1], C. Shaw[2], J. E. Posey[1], D. A. Scott[3]; [1]Baylor Coll. of Med., Houston, TX, [2]BAYLOR Coll. OF MEDICINE, Houston, TX, [3]Baylor Coll. Med, Houston, TX

Abstract:

Congenital Anomalies of Kidney and Urinary Tract (CAKUT) are estimated to make up 40-50% of pediatric malformations and are the predominant cause of kidney failure in children and young adults. CAKUT results from aberrant embryonic development and can be triggered by both genetic and environmental factors. Over 100 genes have been implicated in CAKUT, many of which cause other genetic syndromes with multiple congenital anomalies. Given the high genetic and phenotypic heterogeneity of CAKUT, a high-confidence molecular diagnosis would improve clinical management, counseling, and recurrence risk estimates for affected individuals.

We sought to determine the molecular diagnostic efficacy of clinical exome sequencing (cES) in patients with CAKUT and to identify under-recognized CAKUT genes using machine learning. By combining phenotype- and gene-first approaches, we leveraged our database of ~17,000 exomes and publicly available genomic knowledge to identify new genes that cause CAKUT.

Retrospectively, we analyzed phenotypic and genetic findings of 269 individuals with CAKUT as a component of their referral to a diagnostic laboratory for cES. After variant pathogenicity reanalysis, 22.0% (59/269) of cases had a definitive/probable molecular finding in a gene that explained most/all their presenting phenotypes, but only 12.6% (34/269) were findings in an established CAKUT gene. We found no statistically significant difference (p-value >0.05) in cES efficacy among age, sex, specific CAKUT phenotypes, and presenting organ system abnormalities. Based on these results, we recommend that cES be considered on all individuals with CAKUT without a molecular diagnosis.

Next, we hypothesized that genes involved in the same phenotypes will share biological pathways, tissue expression, and molecular interactions. Therefore, they will be similarly annotated across genomic knowledge sources. First, we trained a published machine learning algorithm using a curated list of known CAKUT genes to compare the annotations between our training set and all RefSeq genes. Scores were cross-validated genome-wide using a leave-one-out approach and represent the ranked similarity for our phenotype-specific training set. Using these scores, in conjunction with available mouse models, published case reports, and cES cases from our cohort with matching diagnoses and CAKUT phenotypes, there is sufficient evidence to suggest that deleterious variants in *PCNT*, *WNT7A*, *CBL*, *MED13L*, and *PHIP* predispose to CAKUT.

Our findings demonstrate the diagnostic utility of cES in patients with CAKUT and broaden the phenotypic features associated with multiple disease-causing genes.

Title: *ARID3A*: a novel gene associated with congenital anomalies of the kidneys and urinary tract.

Authors:

H. Milo Rasouly[1], S. Krishna Murthy[1], S. Bheda[1], M. Marasa[2], S. Shril[3], F. Hildebrandt[4], E. Fiaccadori[5], A. Materna-Kiryluk[6], G. Masnata[7], M. Saraga[8], V. Tasic[9], G. Ghiggeri[10], K. Kiryluk[11], S. Sanna-Cherchi[12], A. G. Gharavi[12]; [1]Columbia Univ., New York, NY, [2]Columbia Univ., New York City, NY, [3]BCH, Harvard Med. Sch., Boston, MA, [4]Boston Children's Hosp, Boston, MA, [5]Univ. di Parma, Parma, Italy, [6]Uniwersytet Medyczny w Poznaniu, Poznan, Poland, [7]Univ. of Cagliari, Cagliari, Italy, [8]Univ. Hosp. of Split, Split, Croatia, [9]Univ. Children's Hosp., Skopje, Macedonia, The Former Yugoslav Republic of, [10]Giannina Gaslini Inst., Genova, Italy, [11]COLUMBIA Univ. HEALTH Sci.S, New York, NY, [12]Columbia Univ, New York, NY

Abstract:

Background: Congenital anomalies of the kidney and urinary tract (CAKUT) are identified in about 0.5% of live birth and are associated with increased risk for chronic kidney disease and kidney failure. While many genes are known to cause CAKUT, they are identified in only 10-20% of individuals with the most severe anomalies like kidney agenesis or hypodysplasia. This study aimed at identifying novel genes associated with severe kidney anomalies (KA). **Methods**: Exome sequencing was performed on 2,016 cases with KA. We matched 1,723 KA cases to 22,252 controls from the database of the institute of genomic medicine at Columbia University based on genetic ancestry to perform genome-wide enrichment analysis of qualifying variants in protein-coding genes. Cases and controls FASTQ data were all processed, aligned, and called using the same pipeline. To control population stratification, cases and controls were first divided into 10 clusters based on genetic ancestry, the analysis was performed in each cluster separately and then combined P-values and odds ratios were calculated using the Cochran-Mantel-Haenszel test. **Results**: Under the loss-of-function (LoF) model, only two known genes reached genome-wide significance (p-value$<10^{-6}$, *HNF1b* and *PAX2*), but all other known genes did not. When restricting the analysis to constrained genes (pli>0.9 and LOEUF<0.35) with at least 3 cases and no controls with LoFs, we identified 2 novel genes: *ARID3A* (AT-rich interaction domain 3A) and *NR6A1* (nuclear receptor subfamily 6 group A member 1). An additional *ARID3A* LoF variant not included in the case-control analysis was identified in a proband with highly functioning autism, and emotional hypersensitivity in addition to KA. Through collaboration with Dr. Hildebrandt group, 2 additional individuals with de-novo LoFs in *ARID3A* were identified, one in a patient with heart malformation and hypoplasia of facial muscles in addition to KA. No additional variant was identified in *NR6A1*. Interestingly, ARID3A was reported to promote nephric regeneration, and the *Arid3a*$^{-/-}$ kidney mouse cell line was reported to spontaneously develop into multicellular nephron-like structures in vitro. Finally, renal anomalies have been reported in individuals with a 19p13.3 microdeletion which encompasses *ARID3A*. **Conclusions**: We identified 6 individuals with LoF in *ARID3A* and no LoF variant in 22,252 matched controls. Model organisms demonstrated a role for arid3a in nephron development, suggesting that *ARID3A* is causing KA in patients.

Title: First GWAS of cystatin-C kidney function in continental Africans identifies novel loci & refines known associations

Authors:

R. Mayanja[1,2,3], T. Machipisa[4,5], O. Soremekun[2], A. B. Kamiza[6,7], C. Kintu[6,1], A. Kalungi[6], R. Kalyesubula[8], O. J. Sande[1], D. Jjingo[9], J. Fabian[10,11], C. Robinson-Cohen[12], N. Franceschini[13], D. Nitsch[14], M. Nyirenda[15,14], E. Zeggini[16,17], A. P. Morris[18], T. Chikowore[19,20], S. Fatumo[6,7,14,16]; [1]Dept. of Immunology and Molecular Biology, Sch. of BioMed. Sci., Coll. of Hlth.Sci., Makerere Univ., Kampala, Uganda, [2]MRC/UVRI and LSHTM Uganda Res. Unit, Entebbe, Uganda, [3]The African Computational Genomics Res. Group, Entebbe, Uganda, [4]Clinical Res. Lab.-Genetic and Molecular Epidemiology Lab. (CRLB-GMEL), Population Hlth.Res. Inst. (PHRI) & McMaster Univ., David Braley Cardiac, Vascular and Stroke Res. Inst., 237 Barton Street East, Hamilton, L8L 2X2, ON, Canada, [5]Dept. of Med., Univ. of Cape Town & Groote Schuur Hosp., Cape Town, South Africa, [6]The African Computational Genomics (TACG) Res. Group, MRC/UVRI and LSHTM Uganda Res. Unit, Entebbe, Uganda, [7]Malawi Epidemiology and Intervention Res. Unit, Lilongwe, Malawi, [8]Med. Res. Council/ Uganda Virus Res. Inst. (MRC/UVRI) and London Sch. of Hygiene and Tropical Med. (LSHTM) Uganda research Unit, Entebbe, Uganda, [9]African Ctr. of Excellence in Bioinformatics (ACE-B), Makerere Univ., Kampala, Uganda, [10]Med. Res. Council/Wits Univ. Rural Publ. Hlth.and Hlth.Transitions Res. Unit (Agincourt), Sch. of Publ. Hlth., Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa, [11]Wits Donald Gordon Med. Ctr., Sch. of Clinical Med., Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa, [12]Div. of Nephrology and Hypertension, Dept. of Med., Vanderbilt Univ. Med. Ctr., Nashville, TN, [13]Gillings Sch. of Global Publ. Hlth., Univ. of North Carolina, Chapel Hill, NC, [14]London Sch. of Hygiene and Tropical Med., London, United Kingdom, [15]Clinical Res. Lab.-Genetic and Molecular Epidemiology Lab. (CRLB-GMEL), Population Hlth.Res. Inst. (PHRI) & McMaster Univ., David Braley Cardiac, Vascular and Stroke Res. Inst.,237 Barton Street East, Hamilton,L8L 2X2, ON, Canada, [16]Inst. of Translational Genomics, Helmholtz Zentrum München - German Res. Ctr. for Environmental Hlth., Neuherberg, Germany, [17]TUM Sch. of Med., Translational Genomics, Technical Univ. of Munich and Klinikum Rechts der Isa, Munich, Germany, [18]Ctr. for Genetics and Genomics Versus Arthritis, Ctr. for Musculoskeletal Res., Div. of Musculoskeletal and Dermatological Sci., The Univ. of Manchester, Manchester, Manchester, United Kingdom, [19]Sydney Brenner Inst. for Molecular BioSci., Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa, [20]MRC/Wits Dev.al Pathways for Hlth.Res. Unit, Dept. of Pediatrics, Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa

Abstract:

Background: Chronic kidney disease is becoming more prevalent in Africa, and its genetic determinants are poorly understood. Currently, in clinical practice, the creatinine-based estimated glomerular filtration rate (eGFRcrea) is commonly used to assess kidney function, modelling the excretion of the endogenous biomarker (creatinine). However, eGFRcrea has recently been shown to inadequately detect individuals with low kidney function in Sub-Saharan Africa, with cystatin-C estimated glomerular filtration rate (eGFRcys) exhibiting significantly superior performance. To address the limitations of eGFRcrea, we conducted the first genome-wide association study (GWAS) for eGFRcys in continental Africans. **Methods:** Using the Uganda Genome Resource (UGR) we performed a GWAS of eGFRcys in 5,877 Ugandans and evaluated replication in independent studies. Subsequently, putative causal variants were screened through Bayesian fine-mapping. Functional annotation of the GWAS loci was performed using the Functional Mapping and Annotation (FUMA) platform. **Results:** Three independent lead single nucleotide polymorphisms (SNPs) (P-value<$5x10^{-8}$) were identified; rs59288815 (ANK3), rs4277141 (OR51B5) and rs911119 (CST3). The rs911119 SNP maps to the cystatin C gene and has been previously associated with eGFRcys among Europeans. From fine-mapping, rs59288815 and rs911119 each had a posterior probability of causality of >99%. With gene-set enrichment analyses of the olfactory receptor family 51 overlapping genes, we identified an association with the G-alpha-S signalling events. **Discussion:** Our study found two novel SNPs for eGFRcys in continental Africans (rs59288815 and rs4277141) and validated a previously well-established SNP (rs911119) for eGFRcys. The identified gene-set enrichment for the G-protein signalling pathways relates to the kidney to readily adapt to an ever-changing environment. This is the first GWAS of eGFRcys in continental Africa. Additional GWASs are required to represent the diverse regions in Africa.

Title: Hypertension predicts polycystic kidney disease 11 years ahead of time in PKD1 and PKD2 carriers, with a PPV of 71%

Authors:

N. Telis[1], L. Mcewen[2], A. Bolze[3], D. Judge[4], P. Pawloski[5], J. Grzymski[6], K. Schiabor Barrett[7], N. Washington[8], E. Cirulli[1]; [1]Helix, San Mateo, CA, [2]Helix OpCo LLC, San Mateo, CA, [3]Helix, San Carlos, CA, [4]Med. Univ. of South Carolina, Charleston, SC, [5]Hlth.Partners, Minneapolis, MN, [6]DRI, Reno, NV, [7]Helix, San Diego, CA, [8]Helix, Albany, CA

Abstract:

Introduction: Autosomal dominant polycystic kidney disease (ADPKD) is the leading hereditary form of kidney disease and is characterized by the formation of fluid-filled cysts that decrease kidney function and lead to complications including cardiovascular disease, kidney failure, and other organ failure. Although causal variants in the PKD1 and PKD2 genes have been identified, penetrance, morbidity, and outcome are known to vary widely in carriers. We propose to identify early markers of disease and to characterize disease presentation at the level of biomarkers, metabolites, and comorbidity incidence.
Methods: This study is based on exomes and medical information from the UK Biobank (n=470k) and 3 American cohorts (n>70k) sequenced with Helix: the Healthy Nevada Project, In Our DNA SC, and myGenetics. We identified individuals with rare LoF variants in either PKD1 or PKD2. We subsequently analyzed these individuals' medical records, blood labs, and NMR-based metabolomics data (n~120k) to characterize the detailed clinical and biomarker-level disease presentation.
Results: We find that 0.05% of the general population has a relevant variant in PKD1 or PKD2 (1/2000 people). In this population-level sample, the prevalence of kidney disease in carriers with hypertension is 71% (74% for PKD1 and 66% for PKD2) compared to 12% for noncarriers with hypertension. For carriers without hypertension, the prevalence is 14% (2% for PKD1 and 30% for PKD2) vs. 1% for noncarriers without hypertension. Importantly, the hypertension diagnosis precedes the kidney disease diagnosis by a mean of 11 years in these carriers. We additionally identify a subset of carriers who are outliers for a set of metabolites that includes cystatin C, creatinine and urea. Assessed prospectively, these carriers have decreased eGFR and a number of other signs of early stage kidney disease a mean of 5 years prior to their KD diagnosis.
Conclusion: Identifying PKD1 and PKD2 carriers with hypertension would allow prediction of kidney disease a mean of 11 years in advance of the disease, with a PPV of 71%. Additionally, aberrant but subclinically significant biomarkers can provide an early warning of morbidity.

Session 104: Addressing barriers to accessible genetic research and services

Location: Conv Ctr/Room 202A/Level 2

Session Time: Saturday, November 4, 2023, 1:00 pm - 2:00 pm

Title: Effective recruitment and retention of a diverse pediatric population in a genome sequencing study

Authors:

K. Bonini[1], M. A. Ramos[1], N. R. Kelly[2], L. Scarimbolo[1], B. Insel[1], S. A. Suckiel[1], K. Brown[2], M. Di Biase[2], K. M. Gallagher[2], N. Ioele[3], J. Lopez[2], K. López Aguiñiga[1], P. N. Marathe[1], E. Maria[2], J. A. Odgis[1], J. E. Rodriguez[1], M. A. Rodriguez[1], N. Ruiz[1], M. Sebastin[2], N. M. Yelton[1], C. Cunningham-Rundles[1], K. Davis[4], M. Gertner[1], I. Laguerre[5], T. V. McDonald[2], P. E. McGoldrick[6], M. Robinson[7], A. Rubinstein[8], L. H. Shulman[8], T. Williams[9], S. M. Wolf[6], E. G. Yozawitz[8], R. E. Zinberg[1], N. S. Abul-Husn[10], L. J. Bauman[11], G. A. Diaz[1], B. S. Ferket[1], J. M. Greally[2], V. Jobanputra[12], B. D. Gelb[1], E. E. Kenny[1], M. Wasserstein[2], C. R. Horowitz[1]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Children's Hosp. at Montefiore/Montefiore Med. Ctr./Albert Einstein Coll. of Med., Bronx, NY, [3]Mem. Sloan Kettering Cancer Ctr., Middletown, NJ, [4]Bethany Baptist Church, New York, NY, [5]Children's Cultural Ctr. of Native America, New York, NY, [6]Boston Children's Hlth.Physicians/Maria Fareri Children's Hosp., Hawthorne, NY, [7]Bethel Gospel Assembly, New York, NY, [8]Children's Hosp. at Montefiore/Montefiore Med. Ctr./Albert Einstein Coll. of Med., Bronx, NY, [9]Midwifery Collective, Brooklyn, NY, [10]Icahn Sch. of Med. at Mount Sinai / 23andMe, Inc., New York / Sunnyvale, NY, [11]Albert Einstein Coll. of Med., Bronx, NY, [12]New York Genome Ctr., New York, NY

Abstract:

Purpose: Diverse patients have not been adequately recruited into genomics studies, limiting the generalizability of findings and benefits of discoveries. Here we describe the impact of patient-centered, culturally, and linguistically appropriate strategies for recruitment and retention in a genome sequencing study (NYCKidSeq) of pediatric patients with suspected genetic conditions. Methods: A stakeholder board composed of patients, parents, advocates, clinicians, and health system leaders guided all phases of the project. Formative research with parents whose children had undergone genetic testing, as well as feedback from stakeholders and participants enrolled in the pilot phase of NYCKidSeq, informed the development of the consent document, study materials, and approaches. Strategies for recruitment and retention included: involving culturally and linguistically congruent study staff reflecting/representing the target population; building relationships with referring clinicians; sustaining connection with participants; providing flexibility for study visits including in-person, remote, evenings and weekends with provisions for childcare on site when needed; and employing data-driven quality improvement to create an iterative approach to problem solving. The study took place in person and then remotely in the context of COVID-19 restrictions. Results: Of the 1847 pre-screened patients, 1655 (89.6%) were eligible and 1059 (70%) enrolled with a decline rate of 6.5% (N=107). Of the 1059, 60.5% had not previously undergone genetic testing and 87.3% (N=924) completed the study. Participants were 75.1% non-White; 65.5% had Medicaid/Medicare; 42.8% lived below the federal poverty level; 52.8% lived in a medically underserved area. There were no demographic differences between those who enrolled in NYCKidSeq versus those who declined, or between those who completed the study versus those who did not. Conclusion: Stakeholder-informed, patient-centered, data-driven approaches to recruitment and retention may support low rates of declining and high rates of completing pediatric genomic studies and help address racial, ethnic and sociodemographic disparities in trial participation. Future studies should compare strategies in different contexts and research areas.

Title: Asian Health Coalition (AHC) addresses cultural and linguistic barriers in genomics education for Asian Americans, Native Hawaiians, and Pacific Islanders (AANHPI) through AANHPI Genetic and Genomic Education (AGE) project.

Authors:

S. Kim[1], F. T. Randal[1], S. Wesmiller[2], K. Whitt[3], J. Xiong[2], S. Qi[1], K. E. Kim[4]; [1]Asian Hlth.Coalition, Chicago, IL, [2]Univ. of Pittsburgh, Pittsburgh, PA, [3]The George Washington Univ., Washington DC, DC, [4]Univ. of Chicago Med., Chicago, IL

Abstract:

Background: Due to health disparities resulting from differences in genetics and genomics and low health literacy rates for AANHPI and limited English proficiency populations, there is a critical need for culturally and linguistically appropriate materials to help reduce the burden of education for those underrepresented in biomedical research. Engaging and educating AANHPIs in genetics and genomics can increase interest in research and give the opportunity to achieve optimal health for future generations through the NIH All of Us Research Program (AoURP). The AoURP is a longitudinal research program to build a diverse health database for researchers to address health disparities in underrepresented groups. One benefit to participants is the genetic return of results through which participants may learn more about genetic health risks and traits. Additionally, there is a need to educate the academic community on appropriate engagement and education around genetics and genomics for AANHPI due to differing language barriers and cultural norms. Objective: Develop culturally and linguistically appropriate approaches to help AANHPI community partners and members better understand the importance of genetic and genomic information and the potential health risks associated with this. In tandem, also create an educational module educating providers and academics on best practices for engaging with AANHPIs. Methods: AHC partnered with the International Society of Nurses in Genetics and AANHPI-serving partner organizations across the US to create AGE education modules. Clinical content and cultural adaptation were garnered to culturally and linguistically adapt the modules for AANHPIs. Critical genetic and genomic differences such as increased risk for diabetes mellitus and cardiovascular disease were addressed to create education and understanding for specific AANHPI communities within each topic. Additionally, a literature review was conducted regarding pharmacogenomics impacting AANHPI communities to give tangible information on the impact of AoURP's genetic return of results for AANHPI participants. Results: An education Session was created and adapted in 5 Asian languages: Simplified Chinese, Traditional Chinese, Vietnamese, Korean, and Hindi, to educate AANHPIs across the nation about genetic and genomic information from the AoURP. Conclusion: Through partnership between academic health professionals and ethnic serving community-based organizations, culturally and linguistically appropriate resources are provided to AANHPI communities regarding genetic and genomic information.

Title: Addressing inequity in rare disease genomic research: preliminary findings from the Rare Genomes Project

Authors:

M. Wojcik[1], J. Serrano[2], G. VanNoy[2], B. Mangilog[2], G. Shah[2], E. Martinez[2], I. A. Holm[1], Y. S. Fraiman[3], M. O'Leary[2], H. L. Rehm[2], A. O'Donnell-Luria[1]; [1]Boston Children's Hosp., Boston, MA, [2]Broad Inst. of MIT and Harvard, Cambridge, MA, [3]Beth Israel Deaconess Med. Ctr., Boston, MA

Abstract:

Background: Effective strategies to address inequities in rare disease genomics remain incompletely understood. The Rare Genomes Project (RGP) is a genome sequencing study for participants with rare disease, intended to address barriers to genetic diagnosis towards improved equity. However, we identified several categories of underrepresentation within our cohort: 1) non-white race; 2) Hispanic or Latinx ethnicity; 3) limited English proficiency; 4) household income under the federal poverty line; 5) educational level of high school or less; and 6) primary residence in a rural area. To address this, we incorporated equity-focused implementation science into the RGP protocol.
Methods: We implemented a multi-faceted intervention which includes expanded outreach, enhanced non-English language support, proactive and flexible participant contact and engagement, and use of mobile phlebotomy. Implementation outcomes were assessed via a mixed-methods approach.
Results: Over the past 18 months, 56 potential participants were referred to our study who were within one or more of our categories of underrepresentation. Of these, 54 (96%) completed the study application, 36 (64%) enrolled, and 29 (48%) returned samples for genome sequencing. To facilitate this, 14/54 (26%) received assistance with the online RGP application, such as telephone support from RGP staff; 11/54 applicants (20%) received assistance to enroll, such as engagement with non-English speaking RGP staff; and 10/36 enrolled families (28%) received assistance to return samples for sequencing, such as mobile phlebotomy. This approach has resulted in a significant increase in the proportion of enrolled Hispanic/Latinx participants from 8.8% to 13.2% ($p = 0.04$), and a non-significant increase in overall enrollment of non-white participants to 11.3% (from 10.9% prior to our intervention), with continued enrollment efforts and analyses of other categories of underrepresentation ongoing. Preliminary survey data ($n = 16$) demonstrate that all participants highly value a genetic diagnosis, particularly to understand their medical problems and prepare for the future.
Conclusions: We describe a personalized, supportive approach to enrollment in a rare disease genome sequencing study, highlighting strategies enabling diversification of our study cohort. Our results suggest that deliberate outreach is critical to expanding access, as is enhanced support throughout the application and enrollment process. Future analyses are underway to further explore access barriers, motivations for participation, and outcomes after diagnosis for minoritized and underserved populations.

Title: Lessons learned in building trust and engaging underrepresented groups in direct-to-consumer genetic testing

Authors:

A. Shastri[1], E. Rios[1], A. Petrakovitz[1], T. Joseph[2], E. Ukandu[1], J. Granka[1], C. Paler[1], R. Jago[1], E. Rasmussen[1], E. Bullis[1], The 23andMe Research Team, N. Abul-Husn[1], H. Taylor[2], L. Scott[3], T. Leak-Johhnson[2], J. Tung[1]; [1]23andMe, South San Francisco, CA, [2]Morehouse Sch. of Med., Atlanta, GA, [3]Breaking the SSickle Cell Cycle, Nashville, TN

Abstract:

Participation in direct-to-consumer genetic testing is highest among individuals with European ancestry, and the resulting lack of diversity in genomic cohorts renders genetic test results less useful for underrepresented populations. Black/African American (B/AA) communities are among the most underrepresented, and a number of barriers to access exist. Here we describe a community-informed program 23andMe, Inc. has launched focused on engagement in B/AA communities that, in partnership with advocacy and academic organizations, is raising awareness for sickle cell (SC) trait and disease.

Prior to launch, the team identified community preferences, barriers, and motivators through qualitative and quantitative assessments. Surveying a general population sample of B/AAs 18 years+, we found strong interest in genetic testing (48%;1000+ respondents), including in carrier status and health predisposition information. Despite this, trust issues due to historic systemic racism in medical institutions were cited as barriers to participation, with communication from trusted sources, transparency about data use, and clear personal and community benefits being key to overcoming these.

Given this, 23andMe designed and executed a program focused on SC trait awareness, in partnership with organizations that serve the B/AA community. Immediate benefit was provided through a Health and Ancestry kit at no cost, which includes FDA-authorized genetic health risk reports, including a Sickle Cell Anemia Carrier Status report. Participants were also offered complimentary genetic consultations from community-based organizations and educational resources co-developed with advocacy groups to ensure transparency in areas such as privacy and data security. Program outreach was done at community events and B/AA community members were involved in design and execution. Finally, emphasis was placed on developing long-term partnerships and long-term benefit - examples include internships for HBCU students and research collaborations.

These efforts have resulted in the recruitment of 8,000+ participants in the first year. Despite research participation being optional, participants in the program consent to research at 89%, higher than B/AA participants in the general research cohort. Participants also engage with program communications at higher rates than in the broader cohort, with 10% more providing phenotypic data through health surveys. The success of this program underscores the importance of a community-informed approach, representation in study design and execution, transparency about motivations and outcomes, and benefit for all parties.

Session 105: Epilepsy - new germline and somatic insights

Location: Conv Ctr/Room 145A/Level 1

Session Time: Saturday, November 4, 2023, 1:00 pm - 2:00 pm

Title: Large-scale genetic studies discover an allelic spectrum of genetic risk factors for the epilepsies

Authors:

S. Chen, Epi25 Collaborative, International League Against Epilepsy Consortium on Complex Epilepsies; Massachusetts Gen. Hosp., Boston, MA

Abstract:

The epilepsies are a group of heterogeneous disorders and follow a complex pattern of inheritance where a number, and/or a combination, of rare and common variants confer risk for the disorder. Here, through an international collaborative effort via the Epi25, we present a whole-exome sequencing (WES) study of >54,000 individuals, composed of 20,979 deep-phenotyped patients with diverse epilepsies and 33,444 controls spanning six genetic ancestries. We investigated ultra-rare variant (URV) risk for the epilepsies using a hypothesis-free approach that evaluates the excess of URVs in epilepsy across the entire exome and for each major subtype of epilepsy. With the unprecedented sample size, we discovered six individual genes and three gene sets at stringent exome-wide significance, revealing both shared and distinct rare variant risk for different subtypes of epilepsies. In parallel, we performed copy number variant calling on the same dataset and genome-wide association study (GWAS) in partnership with the International League Against Epilepsy (ILAE) consortium. The GWAS meta-analyzed >29,000 subjects with common epilepsies and >52,000 controls, and discovered 26 genome-wide significant common risk loci implicating 29 likely causal genes. Collectively, these results constitute an allelic spectrum of genetic variation contributing to epilepsy, greatly expanding our view of the genetic landscape of this highly heterogenous disorder. Meanwhile, by integrating all these findings - from across rare single nucleotide/short indel-, copy number-, and common variants - we identified convergence between different genetic risk factors in the same genes, suggesting that the wide range of risk loci likely points to a much smaller set of genes with at least partially shared etiological processes. More broadly, comparing our results to other large-scale WES studies, we found strong evidence for a genetic overlapping between epilepsy and other neurodevelopmental disorders. This latest package of our studies, representing the largest-scale genetic investigation into the epilepsies to date, suggested that well-powered and well-phenotyped association analysis may help elucidate the complex genetic architecture underlying the epilepsies. The accumulated findings provide a valuable resource for the epilepsy research community; thus we have made all our discovered variants and association statistics fully publicly available through a user-friendly browser (https://epi25.broadinstitute.org/), which we believe will greatly accelerate functional and translational follow-up studies.

Title: *De novo* and selfish maternal variants in *SYTL4*, encoding a member of the synaptotagmin-like protein family, disrupt vesicle exocytosis and cause an incompletely penetrant X-linked neurobehavioral disorder with autism, epilepsy, and spasticity.

Authors:

M. Scala[1,2], L. Tianzhi[3], M. Severino[2], M. Mancardi[2], G. Nobile[2], L. Nobili[2], O. Murch[4], I. Tully[4], A-S. Denommé-Pichon[5,6], L. Faivre[7,6], O. Rieß[8], T. Haack[9], K. Kahrizi[10], H. Najmabadi[10], V. Kalscheuer[11], A. Torella[12,13], P. Striano[2,1], F. Zara[2], V. Nigro[12,13], V. Capra[2], C. Ma[3]; [1]Univ. of Genoa, Genoa, Italy, [2]IRCCS Istituto Giannina Gaslini, Genoa, Italy, [3]Huazhong Univ. of Sci. and Technology, Wuhan, China, [4]Univ. Hosp. of Wales, Cardiff, United Kingdom, [5]Dijon Bourgogne Univ. Hosp., Dijon, France, [6]INSERM, Dijon, France, [7]CHU Dijon, Dijon, France, [8]Univ. Hosp. and Med. Faculty Tübingen, Tübingen, Germany, [9]niversity Hosp. and Med. Faculty Tübingen, Tübingen, Germany, [10]Univ. of Social Welfare and Rehabilitation Sci., Tehran, Iran, Islamic Republic of, [11]Max Planck Inst. for Molecular Genetics, Berlin, Germany, [12]Univ. of Campania Luigi Vanvitelli, Naples, Italy, [13]Telethon Inst. of Genetics and Med., Pozzuoli, Italy

Abstract:

Synaptogenesis and neuronal connectivity rely on the coordinated orchestration of complex intracellular events regulating membrane trafficking. Synaptotagmin-like proteins (Slp) are lipid-binding effectors of specific small Rab-GTPases playing crucial roles as vesicle trafficking modulators. The human *SYTL4* gene encodes the Synaptotagmin-like 4 (Slp4), a member of the Slp family closely interacting with Rab27A and synapse-enriched proteins. *SYTL4* is highly expressed in brain regions determining the social behavior and *Sytl4* knockout in mice results in specific behavioral impairment. Through exome sequencing and international matchmaking platforms, we identified a range of either *de novo* or maternal variants in *SYTL4* in a cohort of seven male individuals from five unrelated families displaying overlapping neurodevelopmental phenotypes. Affected individuals presented with psychomotor delay, mild-to-severe intellectual disability, autism, kind personality, epilepsy, and spasticity. Brain MRI revealed diffuse white matter abnormalities in some cases. We detected five distinct variants in *SYTL4*, either stop-gain or missense affecting the C2A and C2B domains, that primarily mediate Slp4 lipid-binding ability. Diseases mutants (T337M, D387V, H448Q, R451G, and H580R) were introduced into *Sytl4* by site-directed mutagenesis in HEK293T and PC12 cells. Cell phenotype was analyzed by confocal imaging and characterized through liposome co-flotation and total internal reflection fluorescence microscopy (TIRFM) assays. All variants led to an impairment of Slp4 functions with variant-specific effects. Missense changes in the C2A (D387V, H448Q and R451G) and C2B domains (H580R) affected Slp4 membrane-binding ability, and dense core vesicles (DCVs) docking and fusion processes. A similar disruption was observed for R38* and W643* mutants. Several mutants (R38*, D387V, H448Q, R451G, H580R, and W643*) led to a gain of function effect, acquiring the ability of mediating DCV fusion as compared to the wild type (wt). Incomplete penetrance was observed in subjects harboring the p.(Asp387Val) and p.(Arg451Gly) variants, with mildly affected or reportedly unaffected male carriers. Additionally, we observed a transmission ratio distortion (TRD) due to the preferential transmission of *SYTL4* variants over wt alleles, suggesting that *SYTL4* act as a 'selfish gene' in human subjects. Overall, our data show that *SYTL4* variants may act as 'selfish genetic elements' and significantly impact on vesicle trafficking and exocytosis, leading to an incompletely penetrant X-linked neurodevelopmental disorder with autism, epilepsy, and spasticity in males.

Title: Tissue resampling reveals missing somatic variation in brain tissue from pediatric focal epilepsy patients

Authors:

D. Koboldt[1], S. Ramadesikan[2], A. Rivaldi[1], M. Marhabaie[1], K. E. Miller[1], T. Bedrosian[2], D. L. Thomas[1], J. Pindrik[1], A. Shaikhouni[1], D. R. Boué[1], R. Wilson[3], A. Ostendorf[4], E. R. Mardis[4]; [1]Nationwide Childrens Hosp., Columbus, OH, [2]Nationwide Children s Hosp., Columbus, OH, [3]Nationwide Children's Hosp., Columbus, OH, [4]Nationwide Children's Hosp., Columbus, OH

Abstract:

Developmental brain lesions such as focal cortical dysplasia (FCD) and low-grade epilepsy-associated tumors (LEATs) represent a major cause of drug-resistant epilepsy in the pediatric population. In the last decade, genomic studies of brain tissue from such patients have identified brain-specific somatic mutations underlying many lesions. Perhaps the best-understood etiology involves FCD type II - characterized by focally disrupted cortical lamination and dysmorphic neurons - in which ~25-55% of published cases harbor somatic mutations in MTOR pathway members. Yet the significant proportion of unexplained FCDII patients suggests that some of the genetic drivers remain hidden. To date, we have performed exome and RNA sequencing on 308 surgical brain tissues from 100 pediatric epilepsy patients. We have identified and validated brain-restricted mosaic variants in well-established genes (*MTOR*, *PTEN*, *RHEB*, and *SLC35A2*) as well as candidate novel genes (*EEF2*, *NAV2*, and *PTPN11*) in FCDs. However, 50% of FCDI/FCDII patients lack a somatic variant in the tissues sequenced. Some may harbor somatic variants below the detection threshold (~1-2% AF) or outside the regions targeted by exome sequencing. Yet another possibility is that variants may be missed due to sampling effects. We and other groups sequence nucleic acids from fresh-frozen tissue, often cut from the margins of each specimen before fixation. It is therefore possible that insufficient sampling of heterogeneous tissue may allow some genetic variants to go undetected. To test this hypothesis, we selected eight study patients (7 FCD, 1 LEAT) who met three criteria: (1) no history of prenatal concerns or birth complications; (2) highly focal lesions with clear neuropathological diagnoses suggesting genetic etiologies; and (3) negative somatic results from our study. For each patient, we obtained and sequenced sequenced DNA from a different part of the lesional tissue section. Candidate somatic variants were validated by targeted deep resequencing (10,000x depth) in all patient tissues, as well as blood and a reference DNA control. We validated six somatic coding variants in five of eight patients studied (62.5%), including variants in *TSC2*, *FIGN*, *CHEK2*, *CUBN*, and *PGBD5* in cortical dysplasia patients and *BRAF* p.V600E in the LEAT. Allele frequencies for somatic variants ranged from 0.2% to 3.3% for FCD cases, whereas a driver *BRAF* variant undetected in frozen tissue reached 36.3% AF in a macro-dissected specimen. Our results suggest that sampling effects are a significant contributor to genetic diagnostic rates among epilepsy surgery patients, particularly for FCD cases with low mosaic fractions.

Title: Low-Abundance Somatic Variants Activating Ras-MAPK Signaling Cause Drug-Resistant Focal Epilepsy in Adolescents and Adults

Authors:

S. Khoshkhoo[1], Y. Wang[2], Y. Chahine[3], A. M. Tillett[1], A. Huang[3], S. M. Robert[4], B. Chhouk[3], E. Kiziltug[4], C. Nelson-Williams[4], E. J. Stronge[3], H. W. Phillips[5], T. Adikari[6], Z. Ye[6], T. Witkowski[6], D. Lai[7], J. Lokan[8], E. Yang[3], E. L. Heinzen[7], E. C. Damisah[4], H. Moradi[9], R. Coras[10], B. Mathon[11], V. Navarro[11], F. Bielle[11], S. Alexandrescu[3], A. Huttner[4], I. E. Scheffer[6], S. Berkovic[6], M. Hildebrand[6], A. Poduri[3], N. DeLanerolle[4], D. Spencer[4], T. Valiante, A[9], I. Blümcke[10], K. T. Kahle[12], E. A. Lee[3], C. Walsh[13]; [1]Brigham and Women's Hosp., Boston, MA, [2]Harvard Univ., Cambridge, MA, [3]Boston Children's Hosp., Boston, MA, [4]Yale Sch. of Med., New Haven, CT, [5]Stanford Univ. Sch. of Med., Stanford, CA, [6]Univ. of Melbourne, Heidelberg, Australia, [7]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [8]Austin Hlth., Heidelberg, Australia, [9]Krembil Brain Inst., Univ. Hlth.Network, Toronto, ON, Canada, [10]Univ. Hosp. Erlangen, Erlangen, Germany, [11]Sorbonne Université, ICM, Salpêtrière Hosp., Paris, France, [12]Massachusetts Gen. Hosp., Boston, MA, [13]Children's Hosp Boston, Boston, MA

Abstract:

Mesial temporal lobe epilepsy (MTLE), characterized by seizures arising from the hippocampus, is the most common focal epilepsy subtype in adults. Whole-exome sequencing (WES) and gene-panel sequencing (GPS) studies of blood- and buccal-derived DNA have had minimal success in identifying genetic determinants of focal epilepsies such as MTLE. Recently, post-zygotic (i.e., somatic) variants have emerged as a major cause of pediatric focal epilepsy-associated lesions like focal cortical dysplasia, but their contribution to MTLE pathogenesis is unknown. To evaluate the role of somatic variants in drug-resistant MTLE, we recently performed high-coverage WES (depth >500X) of DNA derived from the hippocampus, and paired brain tissue and/or blood when available, of 105 surgically-treated MTLE patients and 30 neurotypical individuals. We detected 10 pathogenic somatic variants all predicted to constitutively activate Ras-MAPK signaling, in patients with MTLE and none in the neurotypical controls. Since the variant allele frequency was typically less than 2% in patients with mesial temporal sclerosis—the predominant MTLE histopathology—we hypothesized that most MTLE-associated somatic variants are present at extremely low-abundance and may be undetectable using standard WES approaches. To test this, we performed GPS paired with duplex sequencing (depth >1000X) on DNA derived from the hippocampus of 218 surgically-treated MTLE patients and 56 neurotypical individuals. This approach yielded 98 pathogenic somatic Ras-MAPK variants in 81 patients with MTLE (37% of total tested) and no neurotypical controls, overall supporting a significant association ($p<1e-5$) and a likely causal role for these variants in MTLE pathogenesis. All the MTLE-associated Ras-MAPK variants increased signaling through either gain-of-function of the activators (FGFR1, PTPN11, BRAF, KRAS, etc.) or loss-of-function of the repressors (NF1, RASA1, LZTR1, etc.) of this pathway. Most variants (n=92, 93%) had variant allele frequencies less than 1%, suggesting a late gestational or early postnatal origin, which may explain the onset of MTLE in adolescence or adulthood. Some of the novel, recurrent MTLE-associated SHP2 (protein encoded by PTPN11) variants were examined using in vitro assays that demonstrated activation of downstream Ras-MAPK signaling. In summary, low-abundance hippocampal somatic variants activating Ras-MAPK signaling likely cause MTLE in a significant subset of patients with sporadic, drug-resistant disease. Our findings provide a novel genetic mechanism for MTLE and highlight new therapeutic targets for this common form of epilepsy.

Session 106: Human genome evolving II

Location: Conv Ctr/Room 146B/Level 1

Session Time: Saturday, November 4, 2023, 1:00 pm - 2:00 pm

Title: The variation and evolution of complete human centromeres

Authors:

G. A. Logsdon[1], A. N. Rozanski[1], F. Ryabov[2], T. Potapova[3], V. Shepelev[4], Y. Mao[5], M. Rautiainen[6], S. Koren[6], S. Nurk[6], D. Porubsky[1], J. K. Lucas[7], K. Hoekzema[1], K. M. Munson[1], J. L. Gerton[3], A. M. Phillippy[6], I. A. Alexandrov[8], E. E. Eichler[1]; [1]Univ. of Washington, Seattle, WA, [2]Natl. Res. Univ. Higher Sch. of Economics, Moscow, Russian Federation, [3]Stowers Inst. for Med. Res., Kansas City, MO, [4]Inst. of Molecular Genetics, Moscow, Russian Federation, [5]Shanghai Jiao Tong Univ., Shanghai, China, [6]NIH/NHGRI, Bethesda, MD, [7]Univ. of California, Santa Cruz, Santa Cruz, CA, [8]Tel Aviv Univ., Tel Aviv, Israel

Abstract:

Advances in long-read sequencing technologies have now made it possible to resolve the complete sequence of centromeres within and between species, allowing their variation to be understood in an evolutionary context. Here, we completely sequenced and assembled all centromeres from a second human genome and used two reference sets to benchmark genetic, epigenetic, and evolutionary variation from a diversity panel of humans and apes (~625 sequence-resolved centromeres). We find that centromere single-nucleotide variation can increase by up to 4.1-fold relative to other genomic regions, with the caveat that up to 45.8% of centromeric sequence, on average, cannot be reliably aligned due to the emergence of new α-satellite higher-order repeat (HOR) structures and two to threefold differences in the length of the centromeres. The extent to which this occurs differs depending on the chromosome and haplotype. Comparing the two sets of complete human centromeres, we find that eight harbor distinctly different α-satellite HOR array structures and four contain novel α-satellite HOR variants in high abundance. DNA methylation and CENP-A chromatin immunoprecipitation experiments show that 26% of centromeres differ in their kinetochore position by at least 500 kbp—a property not readily associated with novel α-satellite HORs. To understand evolutionary change, we selected six chromosomes and sequenced and assembled 31 orthologous centromeres from the common chimpanzee, orangutan, and macaque genomes. Comparative analyses reveal nearly complete turnover of α-satellite HORs, but with idiosyncratic changes in structure characteristic to each species. We find, for example, that chimpanzee centromeres are ~67% smaller than their human counterparts on average; orangutan centromeres have distinct blocks of independently evolved α-satellite HORs that create a “mosaic patchwork” of centromeric α-satellite; and macaque centromeres are consistently larger and more homogenous than the human ones. Phylogenetic reconstruction of human haplotypes supports limited to no recombination between the p- and q-arms of human chromosomes and reveals that novel α-satellite HORs share a monophyletic origin, providing a strategy to estimate the rate of saltatory amplification and mutation of human centromeric DNA.

Title: Community-engaged research identifies genetic histories, basis of malaria resistance, and evolution of shorter stature in lower-altitude (<1400m) Himalayan populations.

Authors:

R. Fahmy[1], A. Jha[1], E. Muthukrishnan[1], M. Penjueli[1], Y. Gautam[1], D. Bhandari[2], S. Tandukar[3], S. Sutton[4], K. Spees[5], G. Wojcik[6], G. P. Gautam[7], J. B. Sherchand[2], C. Bustamante[8], A. Ioannidis[9]; [1]New York Univ. Abu Dhabi, Abu Dhabi, United Arab Emirates, [2]Publ. Hlth.Res. Lab., Inst. of Med., Maharajgunj, Kathmandu, Nepal, [3]Organization for Publ. Hlth.and Environment Management, Latipur, Nepal, [4]Stanford Univ, Stanford, CA, [5]Dept. of BioMed. Data Sci., Stanford, CA, [6]Johns Hopkins Bloomberg Sch. of Publ. Hlth., Baltimore, MD, [7]Dept. of Geography, Tribhuvan Univ. Nepalgunj, Nepalgunj, Nepal, [8]Stanford Univ. Sch. of Med., Stanford, CA, [9]Stanford Univ., Stanford, CA

Abstract:

Most genomic studies in the Himalaya have focused on high-altitude populations, revealing adaptations to hypoxia. The lower-altitude regions - home to over 200 ethnolinguistic groups spanning four major language families - have remained under-explored in genomics. We conducted community-engaged research to generate, merge, and analyze genome-wide genotyping data from 1,205 individuals belonging to 74 Himalayan populations from Tibet, Bhutan, and Nepal and 4,000 individuals from global reference populations. Using multiple global ancestry approaches, we detected East Asian and South Asian ancestries in all Himalayan populations. Differences in proportions of these ancestries are associated with genetic differentiation along an East-West longitudinal cline and the North-South altitudinal gradient. Using various local ancestry-based approaches, we corroborate these results, determine the date of admixtures using analysis of tract lengths, and identify source populations contributing different ancestries in the Himalayan peoples. Genome-wide scans for positive selection on representative indigenous populations of the hills (1000-4000m) and the Terai plains (<400m) revealed several genomic regions evolving under strong positive selection. Candidate single nucleotide polymorphisms (SNPs) were associated with height and immunity in previous genome-wide association studies (GWAS). We detected positive selection on SNPs associated with potential malaria resistance in the natives of the malaria-endemic Terai plains. These signals replicated in multiple independently-evolving populations cohabiting the Terai. We also identified signals of selection on several height-related genes in two indigenous Tibeto-Burman populations of the middle hills, which have shorter stature based on our measurements. These SNPs with strongest signals of selection are nearly fixed in both Himalayan populations for the allele associated with shortness in previous GWAS. However, very few of these candidate SNPs or genes overlap between the two populations, indicating convergent evolution of shorter stature- a polygenic trait - in Himalaya.

Title: Genetic architecture and fitness costs of meiotic recombination across 69,223 in vitro fertilized embryos

Authors:

A. Biddanda[1], S. A. Carioscia[1], I. Vogel[2], E. Hoffman[2], R. McCoy[1]; [1]Johns Hopkins Univ., Baltimore, MD, [2]Univ. of Copenhagen, Copenhagen, Denmark

Abstract:

Crossovers between homologous chromosomes during meiosis are a key source of human genetic diversity and are also essential for ensuring proper chromosome segregation. However, the process of crossover formation is also mutagenic, inducing point mutations and structural variation with potential deleterious effects. Understanding these fitness tradeoffs would clarify the evolutionary mechanisms shaping rates and patterns of recombination in humans.

Much of the current knowledge about the frequency and distribution of meiotic recombination along the human genome comes from samples of living individuals, thereby solely capturing crossovers that have survived gametogenesis and fertilization, as well as prenatal and postnatal development . As fewer than half of all conceptions survive to live birth in humans, this ascertainment bias may obscure variation in the crossover landscape that is subsequently removed by purifying selection and may also contribute to variation in human fertility.

Data from preimplantation genetic testing (PGT) of in vitro fertilized (IVF) embryos offer an ideal resource for mapping meiotic crossovers at extraordinary scale by comparing haplotypes of sibling samples. Using PGT data from 69,223 chromosomally-normal (i.e., euploid) blastocyst-stage embryo biopsies and 12,101 sets of parents, we mapped a total of 4,255,078 crossovers across the embryo genomes (2,619,439 maternal, 1,635,639 paternal) with a median resolution of 8.14 kilobases. We also replicate the stronger maternal versus paternal age effect on crossovers (1.1848 ± 0.003 vs 0.7715 ± 0.002 additional crossovers per year).

The large number of sibling embryos per IVF cycle (median of 5 embryos per case) offers powerful insight into variance in recombination phenotypes and its potential genetic underpinnings. We observe higher variance in the total number of crossovers across the embryo biopsies relative to those observed in living human trios (256 vs. 92, Levene's test $p<10\text{-}128$). The fraction of variance in total crossover number due to covariance across chromosomes is also higher among IVF embryos compared to living parent-child duos (0.53 ± 0.005 vs. 0.22 ± 0.004 paternal, 0.64 ± 0.002 vs. 0.63 ± 0.001 maternal), reflecting individual-specific environmental and genetic modulators and supporting a model whereby rates of crossovers vary in a nucleus-wide fashion. Overall our results highlight the utility of data from IVF embryos for improving fundamental understanding of the fitness tradeoffs that shape human reproductive biology.

Title: The landscape of inverted Alu repeats in the human genome and their evolutionary implications in RNA splicing.

Authors:

D. Denisko[1,2], J. Kim[3], B. Zhao[2], E. Lee[2,1,4]; [1]Harvard Med. Sch., Boston, MA, [2]Boston Children's Hosp., Boston, MA, [3]Dana-Farber Cancer Inst., Boston, MA, [4]Broad Inst., Cambridge, MA

Abstract:

Exon skipping, described as the complete exclusion of an exon from an RNA transcript, makes up the largest fraction of alternative splicing events. Generally, exon skipping arises from disruption of splice donor or acceptor motifs, or nearby cis-acting elements, in turn impacting spliceosome machinery. However, a recent study on the evolutionary loss of tails in hominoids revealed a novel potential mechanism of exon skipping involving inverted repeats (Xia et al. bioRxiv 2021). Specifically, pairs of inverted Alu repeat elements flanking exons may promote the formation of pre-mRNA stem-loops and facilitate exon excision. Here, we present a systematic genome-wide investigation of Alu repeat pairs to understand their contribution to exon skipping, especially, to hominoid-specific events through comparative genomic analysis.

Using skippable and constitutive exons from HEXEvent and ExonSkipDB exon splicing annotations, we found significant enrichment of inverted Alu pairs flanking skippable exons compared to constitutive exons across flanking windows from 500bp to 5kb, with particular enrichment of elements from AluS and AluY subfamilies. Minimum free energy calculations for all sequences of Alu pairs flanking exons revealed that exons flanked by inverted Alu pairs were predicted to form more stable secondary (stem-loop) structures, compared to exons flanked by non-inverted pairs and size-matched sets of random genic sequences.

We also performed comparative genomics analysis using genomes of monkeys, apes, and the human to identify inverted Alu pairs that may impact human evolution. We identified over 50,000 hominoid-specific Alu insertions (49% are AluY, 40% AluS, 9% AluJ), approximately 600 of which can form inverted Alu pairs flanking skippable exons. Pathway analyses of such hominoid-specific inverted pairs flanking skippable exons highlighted several DNA repair processes, including mismatch repair. Experimental validation confirmed 11 out of 11 hominoid-specific skipping events mediated by inverted Alus. Our findings suggest the significance of deep intronic transposon insertions in RNA splicing and genome evolution. Future work aims to further elucidate genomic contexts of inverted Alu pairs in mediating exon skipping and the functional implications in human evolution.

Session 107: Integrating functional annotation data in genetic association studies

Location: Conv Ctr/Ballroom C/Level 3

Session Time: Saturday, November 4, 2023, 1:00 pm - 2:00 pm

Title: A probabilistic framework for joint analysis of genetically-predicted molecular gene product levels and colocalization to identify complex trait mechanisms.

Authors:

J. Okamoto[1], X. Yin[2], B. Ryan[1], J. Chiou[3], F. Luca[4], R. Pique-Regi[4], H. Im[5], J. Morrison[1], E. Fauman[3], M. Laakso[6], M. Boehnke[1], X. Wen[7]; [1]Univ. of Michigan, Ann Arbor, MI, [2]Nanjing Med. Univ., Nanjing, China, [3]Pfizer, Cambridge, MA, [4]Wayne State Univ., Detroit, MI, [5]Univ. of Chicago, Chicago, IL, [6]Univ. of Eastern Finland and Kuopio Univ. Hosp., Kuopio, Finland, [7]Univ. of Michigan, Ann Arbor, MI

Abstract:

Transcriptome-wide association studies (TWASs) and colocalization analysis are popular classes of methods that have shown promise for nominating putative causal genes underlying complex trait biology by linking genome-wide association study (GWAS) variants with expression quantitative trait loci (QTL). A recent work proposed INTACT, an algorithm that exploits the complementary nature of TWAS and colocalization by combining their results in a probabilistic framework. Although this method achieves higher power than colocalization alone and better false discovery control than TWAS alone, it is not robust to scenarios in which a causal gene impacts the complex trait, but not through gene expression. Motivated by recent protein and splicing QTL studies suggesting these scenarios are quite common, we introduce Multi-INTACT, an empirical Bayes framework that extends the INTACT model to jointly consider multiple molecular gene products (e.g. encoded RNA transcript levels and protein levels). Multi-INTACT comprises two stages: a scanning stage and model selection stage. The scanning stage integrates GWAS results and QTL data for two or more molecular gene products to implicate putative causal genes. The model selection stage aims to determine which of the measured molecular gene products for a gene implicated in the scanning stage directly affects the complex trait-of-interest. We employ a Bayesian procedure to compare possible underlying molecular mechanisms using probabilistic evidence. While leveraging additional molecular gene product information, Multi-INTACT shares strengths with INTACT such as computational efficiency and probabilistic uncertainty quantification for causal gene nomination. We show that in a range of simulated scenarios, the Multi-INTACT scanning stage maintains proper false discovery rate control and higher power than methods that consider only one molecular gene product at a time such as TWAS, colocalization analysis, and INTACT. Additionally, the Multi-INTACT model selection procedure consistently provides quantitative evidence that points towards the true underlying molecular mechanism. Finally, we apply Multi-INTACT to 1,540 plasma metabolite GWASs from the METSIM study, integrating the GTEx multi-tissue expression QTL and UK Biobank plasma protein QTL data sets to nominate putative causal genes and mechanisms. The scanning stage identifies many known causal genes underlying metabolite levels, while the model selection stage provides insights concerning whether each nominated gene's effect is likely to be mediated by expression levels or protein levels.

Title: Fine-mapping causal tissues and genes at disease-associated loci

Authors:

B. Strober[1], M. Zhang[2], A. Price[3]; [1]Harvard Sch. of Publ. Hlth., Brookline, MA, [2]Harvard Univ., Roxbury Crossing, MA, [3]Harvard Sch. of Publ. Hlth., Boston, MA

Abstract:

Heritable diseases often manifest in a highly tissue-specific manner, motivating intense efforts to elucidate tissue-specific mechanisms of disease (Hekselman et al. 2019 Nat Rev Genet). Recent studies have identified disease-critical tissues/cell types based on genome-wide patterns, and have deeply dissected a limited number of GWAS loci. However, different GWAS loci may be mediated by different tissues, motivating genome-wide efforts to fine-map causal tissues and genes at individual GWAS loci.

Here, we introduce a new method, Tissue-Gene Fine-Mapping (TGFM), that infers the posterior probability (PIP) for each gene-tissue pair to mediate a disease association at a given locus, using GWAS summary statistics, eQTL summary statistics from diverse tissues, and in-sample LD; TGFM also assigns PIPs to causal SNPs that are not mediated by gene expression in assayed tissues and genes. TGFM models both gene-tissue pairs (using cis-predicted expression) and non-mediated SNPs as potential causal elements, and accounts for both co-regulation across genes and tissues and LD between SNPs, generalizing existing fine-mapping methods. TGFM incorporates genome-wide estimates of each tissue's contribution to disease as tissue-level priors, and employs a sampling approach to account for uncertainty in cis-predicted expression, avoiding false positives that may arise from imperfect predictions of cis-genetic expression. In extensive simulations using real genotypes, TGFM powerfully identifies causal gene-tissue pairs while maintaining correct calibration (whereas methods that do not account for uncertainty in cis-predicted expression suffer inflated type I error).

We applied TGFM to 35 UK Biobank diseases/traits (average N = 318K) and 38 GTEx tissues. TGFM identified an average of 135 PIP>0.5 causal elements per trait, of which 12% were gene-tissue pairs (consistent with estimates of disease heritability mediated by gene expression in GTEx tissues; Yao et al. 2020 Nat Genet) and 88% were non-mediated SNPs. Causal gene-tissue pairs were concentrated in known disease-causal tissues (e.g. 52% in liver and 19% in whole blood for Total cholesterol, 46% in artery and 14% in heart left ventricle for Diastolic blood pressure). TGFM identified the well-studied *PCSK9* gene in whole blood for Total cholesterol (PIP: 0.54;) and *PADI1* in skin tissue for Vitamin D level (PIP: 0.74); *PADI1* is known to interact with keratins in late stage epidermal differentiation. In conclusion, TGFM is a robust and powerful method for fine-mapping causal tissues and genes using eQTL data; ongoing development of richer eQTL data is likely to produce additional discoveries.

Title: Fine-mapping causal cell-types of human diseases and disease variants

Authors:

A. Kim, C. Legros, Z. Zhang, Z. Lu, A. de Smith, N. Mancuso, S. Gazal; Univ. of Southern California, Los Angeles, CA

Abstract:

Genome-wide association studies (GWAS) have identified thousands of genetic variants contributing to human diseases, but their cell types of action remain largely unknown. Existing methods, such as stratified LD score regression (S-LDSC), link cell types to traits using cell-type-specific (CTS) epigenomic annotations but are limited by their capacity to 1) distinguish causal cell types among many significant associations and 2) interrogate individual variants. These limitations hinder our understanding of how genetic variants act at the cellular level to confer disease risk.

Here, we propose CT-FM and CT-FM-SNP, the first methods that fine-map causal cell types for a trait and for its candidate variants. Our methods take as input GWAS summary statistics, a reference panel, a set of CTS annotations (N=1112 for this study), and (for CT-FM-SNP) a list of candidate variants. CT-FM estimates the posterior inclusion probabilities (PIP) for a cell-type to act on disease and outputs credible sets reflecting the number of independent causal cell types while accounting for uncertainty in cell type prioritization.

We benchmarked CT-FM and CT-FM-SNP on GWAS of blood cell traits with known causal cell types. Focusing on Lymphocyte Count GWAS, CT-FM identified two credible sets (CS) corresponding to expected causal cell types (lymphocytes B and T). Applied to 310 candidate variants, CT-FM-SNP assigned a causal cell type to 228 (73%) SNPs, including 176 SNPs assigned to lymphocytes B and/or T, illustrating consistency between the results of our two methods.

Next, we applied CT-FM to 114 GWAS, and identified, on average, 1.4 CS per trait with 11.9 CTS annotations per CS, representing a ~14x decrease from significant associations identified by S-LDSC. CT-FM identified 120 high-confidence (PIP > 0.5) CTS annotations, including expected cell types such as CD4+ T cells for Multiple Sclerosis and Glutamatergic neurons for Bipolar Disorder. We observed 48 traits with at least two CS, suggesting that GWAS variants confer disease risk by impacting independent cell types. Notably, we identified 4 CS for Schizophrenia, including cell types related to brain, immune and liver tissues.

Finally, we applied CT-FM-SNP to 8,897 fine-mapped variants of 41 UK Biobank traits and identified a cell type with PIP>0.5 for 5,627 (63%) variants. For 25 traits with at least 2 CS identified by CT-FM, CT-FM-SNP estimated that ~3/4 of SNPs impact disease risk through a single cell-type, while ~1/4 impacted disease risk through multiple distinct cell-types.

In conclusion, our approaches fine-map causal cell-types of human diseases and disease variants and provide insights into complex trait architectures.

Title: CWAS-Plus: Estimating genome-wide evaluation of noncoding variation from whole genome sequencing data.

Authors:

Y. Kim[1], M. Jeong[1], R. Yurko[2], J. Kim[1], I. Koh[1], H. Lee[1], D. Werling[3], S. Sanders[4,5], B. Devlin[6], J-Y. An[1]; [1]Korea Univ., Seoul, Korea, Republic of, [2]Carnegie Mellon Univ., Pittsburgh, PA, [3]Univ. of Wisconsin-Madison, Madison, WI, [4]Univ California San Francisco, San Francisco, CA, [5]Univ. of Oxford, Oxford, United Kingdom, [6]Univ. of Pittsburgh, Pittsburgh, PA

Abstract:

Background: The noncoding genome contains regulatory elements that play a critical role in human development. Due to advancements in whole genome sequencing (WGS) technologies, we are now able to explore mutations within these regulatory elements. In the meantime, the recently accumulated single-cell data allows the investigation of cell-type-specific regulatory elements, facilitating the identification of cell-type-specific noncoding mutations associated with diseases. To perform a genome-wide evaluation of noncoding mutations using WGS data, an analytic framework that enables fast and easy integration of diverse functional annotations to the WGS data and empowers multiple testing comparisons is essential. **Objectives:** Our study aims to develop CWAS-Plus, a statistical framework to perform a category-wide association test for noncoding variants and provides an efficient analysis of genome-wide noncoding associations. **Methods:** CWAS-Plus conducts genome-wide assessment of noncoding associations using WGS data by integrating functional annotation datasets, including cell-type-specific enhancers and promoters. Variants are categorized into functional annotation combinations, referred to as categories in CWAS-Plus, allowing an enrichment test for qualifying variants. For multiple testing comparison, we computed an effective number of tests based on correlations between the categories. CWAS-Plus explores of relationships between noncoding categories and disease risk through network analysis. To evaluate the performance of CWAS-Plus, a thorough assessment was conducted using WGS data obtained from 1,991 families with autism spectrum disorder (ASD). **Results:** We developed CWAS-Plus, a fast and user-friendly Python package for efficiently detecting risks associated with diverse functional annotations through effective multiple hypotheses testing (https://cwas-plus.readthedocs.io/en/latest/). From annotation to noncoding association testing, CWAS-Plus can process WGS data from approximately 4,000 individuals, along with various functional annotations, including single-cell epigenome data, within 2 hours. Our findings successfully identified noncoding categories enriched for ASD, particularly highlighting regulatory elements of excitatory neurons. **Conclusions:** Our findings showed that CWAS-Plus efficiently processes large-scale WGS data and functional annotations, enabling multiple comparisons of hypotheses. Hence, we present CWAS-Plus as an analytic framework applicable for investigating diverse genomic disorders in future studies.

Session 108: Modifiers in neurological disease

Location: Conv Ctr/Ballroom B/Level 3

Session Time: Saturday, November 4, 2023, 1:00 pm - 2:00 pm

Title: Genetic modifiers in myotonic dystrophy type 1: Insights from a region with a strong founder effect

Authors:

C. Moreau[1], J. Bouchard[1], C. Légaré[1], C. Laprise[1], H. Vézina[1], C. Gagnon[2], É. Duchesne[1], **S. Girard**[1]; [1]Université du Québec à Chicoutimi, Chicoutimi, QC, Canada, [2]Université de Sherbrooke, Sherbrooke, QC, Canada

Abstract:

Myotonic dystrophy type 1 (DM1) is an autosomal dominant inherited disorder for which symptoms, severity and progression are highly variable among patients. Despite large interindividual variations, it is thought to be caused by the expansion of a CTG repeat in the *DMPK* gene. However, this expansion only partly explains variable symptoms onset and progression of the disease. In consequence, other genetic mechanisms such as modifier genes, might be involved and could help to better explain the CTG instability and the disease severity. The Saguenay-Lac-St-Jean (SLSJ) region of Quebec (Canada) has the highest incidence of DM1 worldwide, with a frequency of around 1/630, due to its strong founder effect. We hypothesize that the high prevalence and the availability of data from different sources will allow us to identify modifier genes more easily in this population. We are currently generating genotypes for 200 DM1 patients from the SLSJ population and will combine these with the deep genealogical data provided by BALSAC. BALSAC is a population database built at UQAC based on vital records allowing the reconstruction of the entire genealogy of the Quebec population of European descent from the beginning of the colonization in the 17th century until the present day. Genealogical analyses were conducted using the R package Genlib and ISgen to identify the most probable introducer of the disease in the colony. Since most of the SLSJ DM1 cases have a common haplotype that was inherited from a single common ancestor, this resulted in a big multigenerational family of DM1 in the SLSJ region. Using genotypes, we will infer identical-by-descent (IBD) segments shared among patients with refinedIBD. The genealogical data will be combined with genotypes and patients' clusters sharing IBD segments and related through common ancestors will be identified. With this combined data we expect to find shared genomic regions IBD around the *DMPK* gene, but also around modifier genes that could have been inherited along with the expanded CTG repeat on *DMPK* gene. The identification of genetic modifiers shared by subgroups of genealogically linked patients could be the first step in initiating therapies or drug repurposing. Additionally, the identification of genetic modifiers associated with earlier onset of the disease could lead to early preventive therapies being offered to patients in the SLSJ region, but also elsewhere in the world.

Title: Genetic modifiers of seizures in fragile X syndrome

Authors:

A. Lee, J. Lim, Y. Zhou, M. Epstein, D. Cutler, E. Allen, Z. Wen, P. Jin; Emory Univ., Atlanta, GA

Abstract:

Fragile X syndrome (FXS), the most common inherited cause of intellectual disability, is caused by a CGG trinucleotide repeat expansion in the 5'-UTR of the fragile X messenger ribonucleoprotein 1 (*FMR1*) gene that leads to the loss of the functional fragile X messenger ribonucleoprotein (FMRP). FXS increases the risk of epilepsy. The loss of FMRP may be considered a susceptibility locus but insufficient alone to trigger epilepsy. Hence, finding genetic variants influencing epilepsy in FXS patients would provide important insights into disease mechanisms and therapeutic targets. Here, **we performed whole genome sequencing (WGS) to identify potential genetic modifiers of seizures in FXS followed by functional testing of those genes in *Drosophila* and human-induced pluripotent stem cell (hiPSC)-derived models.** We analyzed WGS data of a cohort of 148 FXS patients (67 with seizure and 81 without seizure). We identified mutations in several genes associated with epilepsy, including *CHD2*, *SLC9A6*, *KCNA1*, *DCX*, and *DEPDC5*. We also identified multiple novel potential modifiers using rare-variant association testing—the burden test and sequence kernel association test (SKAT), which included *GK5*, *RGPD2*, *TTLL4*, and *PDE10A*. To functionally test the effect of identified genes on seizures associated with FXS, we employed the Bang-Sensitivity Behavioral Assay (vortex assay) and measured seizure phenotypes in a *Drosophila* model of FXS. Compared to controls, FXS flies display a significant increase in seizure rate. We found that the loss of *CHD2*, *SLC9A6*, *DEPDC5*, *GK5*, *RGPD2*, and *TTLL4* could all independently ameliorate seizure phenotypes in FXS flies, while the loss of *KCNA1* interestingly led to increased seizures. To further test the potential seizure modifier genes in a human model system, we used neurons generated from FXS patient-derived iPSCs. We measured the electrophysiological activity of these neurons using microelectrode array. FXS neurons displayed hyperexcitability measured by increased firing rate and average number of bursts compared to control neurons. We found that the reduced expression of *CHD2*, a known FMRP mRNA target, rescued the phenotype of increased firing rate in FXS neurons with no significant difference from those of control neurons, suggesting that the loss of *CHD2* leads to the amelioration of seizure phenotypes in human neuronal models of FXS as was in flies. Our integrated analyses have **identified potential genetic modifiers of seizures in FXS. Examining the specific modifiers of seizures in the setting of FXS can help better understand the underlying mechanisms as well as identify novel treatment targets.**

Title: Natural partial loss of function variants in *SEL1L* rescue NGLY1 deficiency by altering ERAD and improving proteasome function.

Authors:

T. Tu'ifua, C. Chow; Univ. of Utah, Salt Lake City, UT

Abstract:

N-glycanase 1 (NGLY1) deficiency is a rare disease caused by autosomal recessive loss of function mutations in the *NGLY1* gene. Patients suffer from movement disorder, developmental delay, liver dysfunction, and alacrima. NGLY1 removes N-linked glycans from glycoproteins in the cytoplasm and is thought to help clear misfolded glycoproteins from the endoplasmic reticulum (ER) through the ER associated degradation (ERAD) pathway. Despite this, the physiological significance of NGLY1 in ERAD is not understood. The best characterized substrate of NGLY1 is NRF1, a transcription factor that upregulates proteasome expression and the proteasome bounce back response.

Our lab created a *Drosophila* model of NGLY1 deficiency that faithfully recapitulates disease phenotypes in patients, including movement disorder, seizures, and lethality. We performed a *Drosophila* modifier screen using this model of NGLY1 deficiency and identified a number of modifiers that reduce the lethality of the model. In particular, we identified two putative loss of function variants in *SEL1L*: S780P and Δ806-809. Both variants are localized in the SEL1L cytoplasmic tail, an uncharacterized domain of the protein. SEL1L is a component of the ERAD complex that retrotranslocates misfolded proteins from the ER to the cytoplasm for degradation.

To test the interaction between these *SEL1L* variants and *NGLY1*, we created CRISPR mutant fly lines that carry these *SEL1L* variants in a common genetic background and tested them with our model of NGLY1 deficiency. We determined that these variants are partial loss of function. Validating our screen, the $SEL1L^{P780}$ and $SEL1L^{\Delta 806\text{-}809}$ variants increase the survival of the NGLY1 deficiency model, compared to the $SEL1L^{S780}$ variant. Further, we found that the $SEL1L^{P780}$ and $SEL1L^{\Delta 806\text{-}809}$ variants are protective against proteasome inhibition, leading to increased survival, in heterozygous *NGLY1* null flies. We also find that these variants modify general ERAD function. We hypothesize that these *SEL1L* variants modify NGLY1 deficiency through NRF1 signaling and ERAD. These results will provide new insights into the role of NGLY1 in ERAD and the etiology of NGLY1 deficiency. *SEL1L* is a strong candidate modifier gene in patients, where variability in presentation is common.

Title: Assessing the functional effect of the Presenilin-1 G206A variant on age of onset of Alzheimer Disease in the Puerto Rican population

Authors:

K. Celis[1], F. Rajabli[2], A. Griswold[2], L. Adams[2], S. Tejada[2], J. J. Sanchez[1], C. Silva[3], S. Simon[2], K. Hamilton-Nelson[4], K. Scott[1], P. Whitehead[2], P. Mena[2], J. Arvizu[5], C. Golightly[2], C. Duarte[6], H. Acosta[7], K. McInerney[6], G. Beecham[2], B. Feliciano-Astacio[6], M. Cuccaro[8], J. Vance[2], D. Dykxhoorn[2], J. Young[2], M. Pericak-Vance[4]; [1]Univ Miami, Miami, FL, [2]Univ. of Miami, Miami, FL, [3]universidad central del caribe, san juan, PR, [4]Univ. of Miami Miller Sch. of Med., Miami, FL, [5]Univ. Of Miami, Miami, FL, [6]Univ. of Miami, miami, FL, [7]Clinica de la memoria, san juan, PR, [8]John P. Hussman Inst. for Human Genomics, Miami, FL

Abstract:

Background: The variant G206A in the Presenilin-1 (*PSEN1*) gene has been identified almost exclusively in Alzheimer Disease (AD) Puerto Rican families. The G206A variant is associated with extreme variability in age of onset (AOO), ranging from 30 to 90 years. We aim to identify the mechanisms involved in the AOO variability between carriers by transcriptomic and functional analysis from brains and induced pluripotent stem cells (iPSCs) and plasma biomarkers. **Methods**: Genotyping data were phased using SHAPEIT to identify local ancestry and RFMix to estimate genetic ancestry. p-Tau181 levels were tested from plasma using Simoa (Quanterix). iPSCs of G206A carriers were reprogrammed using non-integrating Sendai virus. These lines were assessed for pluripotency and chromosomal stability. Isogenic construction of G206A iPSC lines is currently in progress. Two brains samples from G206A carriers were identified. **Results:** We identified 43 G206A carriers (39 AD and 4 cognitively unimpaired <65 years) using whole genome sequencing from 182 families. 55% of AD G206A carriers had early AOO (<65). A single African haplotype (79,995bp) was identified in all carriers. We observed higher pTau181 levels in AD G206A carriers with early AOO (<65) (*p* value= 0.01) vs late (>65). Association analysis did not identify *APOE4* or genes suggested to affect *PSEN1* AOO *(SYNJ1*, *SNX25*, *PDLIM3* and *SORBS2)* as contributors to AOO variability within G206A carriers. We selected three individuals with early and three with late AOO AD to generate iPSC lines. G-band Karyotype, factor loss analyses, genetic finger printing, immunocytochemistry (ICC) and qRT-PCR for intracellular and surface pluripotency markers were confirmed in the iPSC lines. Single nuclei RNA sequencing from brain samples and functional analysis from iPSC lines is in progress. **Conclusion:** The *PSEN1* G206A variant, originating in a shared African ancestry haplotype revealing a founder effect, exhibits variable AOO. Our data suggest that this variability is not mediated by *APOE4* genotype or polymorphisms in nominated *PSEN1* G206A modifiers. Higher levels of p-Tau181 in plasma were found in AD G206A carriers with early vs late AOO, a significant clinical finding not reported before. Thus, identification of additional differences in biomarker levels, transcriptomic profiles, and functional consequences between early and late AOO AD G206A carriers from brains and iPSC, will provide insight in AD pathogenesis, mechanisms influencing AOO variation and identification of potential novel therapeutic targets.

Session 109: New insights in inborn errors of metabolism

Location: Conv Ctr/Room 147A/Level 1

Session Time: Saturday, November 4, 2023, 1:00 pm - 2:00 pm

Title: Impaired nuclear glycogen metabolism affects liver homeostasis in Argininosuccinic aciduria.

Authors:

N. Brunetti-Pierri[1], A. M. D'Alessio[1], C. Perna[1], R. De Cegli[1], B. Julien[2], Y. Lee[3], E. Polishchuk[1], L. Soria[4]; [1]Telethon Inst Genetics & Med., Pozzuoli, Italy, [2]UCL Great Ormond Street Inst. of Child Hlth., London, United Kingdom, [3]Univ. of Connecticut Sch. of Med., Farmington, CT, [4]Telethon Inst. of Genetics and Med. (TIGEM), Pozzuoli, Italy

Abstract:

Glycogen is synthesized de novo in the nucleus, and nuclear glycogenolysis provides a carbon pool for histone acetylation. Abnormal nuclear glycogenolysis results in epigenetic changes and cancer progression. We previously found hepatic nuclear glycogen storage in *Asl*$^{Neo/Neo}$ mice, a mouse model of Argininosuccinic Aciduria (ASA). Here we investigated the impact of nuclear glycogen storage on metabolic homeostasis of *Asl*$^{Neo/Neo}$ mice. While cytosolic glycogen storage was detected in livers of *Asl*$^{Neo/Neo}$ as well as other models of glycogen storage (*Pygl*$^{-/-}$ and *Gaa*$^{-/-}$ mice), nuclear glycogen storage was only detected in *Asl*$^{Neo/Neo}$ mouse livers. In the nuclei of *Asl*$^{Neo/Neo}$ mouse livers, quantification of glycogen metabolizing enzymes suggested increased glycogen synthesis and reduced glycogenolysis, because the protein amounts of glycogen phosphorylase was reduced and glycogen synthase was increased. Notably, marked decrease in lysine acetylation of several histones, including H3K9, was found in ASA livers. Transcriptomic and metabolomic analyses of *Asl*$^{Neo/Neo}$ mouse livers revealed H3K9-related and energy metabolism changes, respectively. Treatment of *Asl*$^{Neo/Neo}$ mice with vorinostat, a pan-HDAC inhibitor improved mouse survival. In conclusion, in ASA livers, we found nuclear glycogen storage, likely as a result of defective nuclear glycogen mobilization. Nuclear glycogen was associated with histone hypoacetylation and changes in gene expression. Interestingly, treatment of *Asl*$^{Neo/Neo}$ mice with an HDAC inhibitor restored histone acetylation and improved mouse survival, providing evidence for a key role of glycogen storage and histone hypoacetylation in the disease pathogenesis.

Title: Multi-omics approach to characterize the TANGO2 function: dance partners and potential therapies

Authors:

A. Kashiparekh[1,2], A. Powers[1,2], A-W. Mohsen[1,2], J. Vockley[1,2], L. Ghaloul Gonzalez[1,2]; [1]Univ. of Pittsburgh, Pittsburgh, PA, [2]UPMC Children's Hosp., Pittsburgh, PA

Abstract:

TANGO2-deficiency disorder (TDD) is caused by mutations in the *TANGO2* (transport and Golgi organization protein 2 homolog) gene. A broad spectrum of clinical manifestations- including metabolic crises with rhabdomyolysis, developmental delays, and cardiac arrhythmias are observed in individuals with the same mutations. Several studies, including ours, show secondary mitochondrial dysfunction and impaired fatty acid oxidation. The broad phenotypic range and no specific phenotypic-genotypic correlation lead to the postulation of modulating genes or the presence of post-translational modifications. We aim to understand the pathogenesis of the disease, as well as elucidate the TANGO2 transcriptional and protein network, using a multi-omics approach. The current treatment recommendation for TDD is pantothenic acid (B5 vitamin) at the recommended daily dose, showing decreased seizures and metabolic crises. Due to mitochondrial abnormalities observed, we also hypothesize that mitochondrial stabilizing compounds (MSCs) in addition to the B5 vitamin can improve the clinical manifestations further in TANGO2 patient fibroblasts.

To achieve these goals, we will perform co-immunoprecipitation studies to pull down TANGO2 and associating proteins. Baseline characterization of the TANGO2 patient fibroblasts using RNAseq and metabolomics, and a downstream correlation, will help pinpoint the biochemical pathways impaired in TDD. Three TANGO2 deficient patient skin fibroblasts will also be treated with Vitamin B5 and analyzed at a multi-omic level to understand the role of pantothenic acid in rescuing some phenotype. TANGO2 deficient patient skin fibroblasts will also be tested with and without MSCs. Functional studies include the quantification of oxygen consumption rate (OCR), ATP production as well as reactive oxygen species (ROS).

Preliminary data shows an efficient pull-down of the TANGO2 protein and associated proteins- which will be sent for further characterization using mass spectrometry. Early transcriptomic data as well as published studies show a decrease in cardiolipin formation in patient cells, postulating a role of TANGO2 protein in lipid trafficking. Functional studies on TANGO2 patient fibroblasts show a decrease in the TANGO2 protein as well as OCR and mitochondrial ATP production in the patient cell lines as compared to the control cells. We also observe an increase in ROS production in patient cell lines. Different concentrations of the most promising MSCs will be examined to see a rescue in these biomarkers.

Title: Examining genotype-phenotype correlations in Cobalamin B-Type(*cblB*)Methylmalonic Acidemia.

Authors:

A. Gebremariam, C. S. Hall, J. L. Sloan, S. McCoy, A. R. Pass, S. Ferry, C. Van Ryzin, O. Shchelochkov, I. Manoli, C. P. Venditti; NHGRI, NIH, Bethesda, MD

Abstract:

Cobalamin B-type (cblB) methylmalonic acidemia (MMA) is an ultra-rare autosomal recessive inborn error caused by pathogenic variants in the MMAB gene, which encodes for the enzyme ATP:cob(I)alamin adenosyltransferase (ATR). ATR catalyzes the synthesis and delivery of 5'-deoxyadenosylcobalamin (AdoCbl), an essential cofactor, to the methylmalonyl-CoA mutase enzyme (MMUT). Both enzymatic defects result in metabolic instability and multisystem injury that are typically non-responsive to B12 supplementation. Supportive care includes dietary management, carnitine supplementation, and intermittent antibiotics, while elective liver transplantation is offered to severely affected patients. We evaluated the presentation, course, complications, and effects of liver transplantation on biomarkers and outcomes in 17 cblB patients, 12 male (70.5%), ranging from 7 to 49 years of age, including eight with a liver or combined liver and kidney transplant (LT/LKT) under our natural history protocol (NCT00078078). We utilized gnomAD, and cross-referenced with HGMD, ClinVar, our NIH database of deeply phenotyped patients, and review by a molecular geneticist, to estimate the incidence of cblB to be 1:1,855,717, and confirm previous reports that c.556C>T p.(Arg186Trp) is the most common variant in individuals of European ancestry (MAF: 0.00015). This variant results in severe phenotype, and was present in 58.8% of alleles in our cohort (8 homozygous). Five patients were compound heterozygotes for c.700C>T p.(Gln234Ter), a frequently identified nonsense variant associated with a higher residual enzymatic activity and B12 responsiveness in vitro. Half of the patients with the homozygous severe variant presented in the neonatal period, while 80% of the patients with c.700C>T p.(Gln234Ter) were late onset including a patient that presented at 6 years with a basal ganglia metabolic stroke. 1-13C-propionate oxidation breath test was developed as a safe, non-invasive, in vivo indirect enzyme activity measurement. The breath test confirmed low enzyme activity in a patient homozygous for p.Arg186Trp (16% of the isotope dose oxidized in 120 min) in contrast to near normal enzymatic activity (47%) in a patient compound heterozygous for p.Arg186Trp and p.Gln234Ter variant. Mild patients maintained lower levels of serum methylmalonic acid concentrations and mitochondrial dysfunction markers, FGF21 and GDF15, than those with severe disease. These genotype-phenotype correlations informed the generation of murine models that are used to enable AAV gene therapy for MMAB, in collaboration with NCATS via the Platform Vector Gene Therapy (PaVe-GT) project.

Title: Integration of genomic sequencing and metabolomic screening for inborn errors of GABA metabolism highlight use of blood-based biomarkers for early diagnosis and access to care across populations

Authors:

C. Gijavanekar[1,2], K. Glinton[1,3], A. Rajagopal[1], T. A. Wilson[1], K. A. Martin[1], V. R. Sutton[1,3,2], S. H. Elsea[1,2]; [1]Baylor Coll. of Med., Houston, TX, [2]Baylor Genetics, Houston, TX, [3]Texas Children's Hosp., Houston, TX

Abstract:

Estimates suggest that ~1000 individuals in the United States may have an autosomal recessive primary enzyme deficiency involving catabolism of GABA (gamma-aminobutyric acid). GABA is catabolized by a 2-step enzymatic pathway involving GABA-transaminase (GABA-T) and succinic semialdehyde dehydrogenase (SSADH) in the astrocyte. Deficiency of either enzyme leads to accumulation of GABA. Early and accurate diagnosis of these rare neurometabolic genetic conditions is crucial for effective treatment. Both conditions are thought to be underdiagnosed due to limited knowledge of pathogenicity of genomic variants in both *ABAT* and *ALDH5A1* that encode GABA-T and SSADH, respectively. The neurodevelopmental endophenotypes can range from moderate-to-severe presentations with early death in the most severe cases. Accurate genetic diagnosis requires comprehensive knowledge of disease - including pathogenic genetic variants and associated biomarkers. Both metabolite and genomic testing may be required for accurate diagnosis but are currently clinically underutilized due to invasiveness of cerebrospinal fluid neurotransmitter analysis and poor characterization of genomic variants. To improve the diagnostic process, we first curated all reported cases of SSADHD to date identifying >150 pathogenic and likely pathogenic variants in *ALDH5A1*, and then estimated the pan-ethnic prevalence of SSADHD to be ~1/460,000, with higher prevalence in Asian populations. We performed plasma untargeted metabolomic analysis and genome sequencing to identify key biomarkers of altered GABA catabolism and genomic variants. We identified plasma biomarkers for GABA-TD (2-pyrrolidinone, succinamic acid, 4-guanidinobutanoate) and SSADHD (2-pyrrolidinone, 4-guanidinobutanoate, argininate) that also distinguish individuals treated with vigabatrin, an anti-seizure medication that blocks GABA-T. From the period 2014-2022, our laboratory assessed >4,000 clinical plasma samples using untargeted metabolomics, identified 7 individuals with confirmed SSADHD or GABA-TD, while genome sequencing (exome, genome, Sanger) identified 15 individuals with bi-allelic pathogenic or likely pathogenic variants in *ALDH5A1* or *ABAT.* Metabolomic profiling demonstrated that blood-based biomarkers effectively interrogate the GABA pathway, and CSF is not required for accurate diagnosis. The prevalence of these two conditions in our diagnostic laboratory database is ~ 1/1000 in a pediatric population with neurological phenotypes. Approaching the diagnostic process in a comprehensive, integrated manner will reduce time to diagnosis and improve access to treatment.

Session 110: Omics, omics everywhere

Location: Conv Ctr/Ballroom A/Level 3

Session Time: Saturday, November 4, 2023, 1:00 pm - 2:00 pm

Title: 'Space Omics' Program of Significant Scale at BCM-HGSC

Authors:

H. Doddapaneni[1], E. Urquieta[2], M-C. Gingras[1], J. Wu[2], J. Rogers[3], H. Chao[1], V. Korchina[1], M. Murugan[1], G. Metcalf[1], J. Posey[4], D. Muzny[1], R. Gibbs[3]; [1]Human Genome Sequencing Ctr., Baylor Coll. Med., Houston, TX, [2]Ctr. for Space Med. and Translational Res. Inst. for Space Hlth.,Baylor Coll. Med., Houston, TX, [3]Human Genome Sequencing Ctr. and Dept. of Molecular and Human Genetics, Baylor Coll. of Med., Houston, TX, [4]Dept. of Molecular and Human Genetics, Baylor Coll. of Med., Baylor Coll. of Med., Houston, TX

Abstract:

The surge in commercial and civilian spaceflight enables for the first time, systematic, large-scale data collection to understand prospective effects of space travel on human health. The Genomics and Space Medicine project (‘Space Omics’) at the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) is designed to study pre- , in-, and post-flight biological specimens using an array of omics assays, including clinical Whole Genome Sequencing (WGS), research assays (RNA-Seq, microbiome, proteomics among others), and biobanking for future use to gain insights into the impact of space travel. These comparisons can also be valuable for advancing terrestrial health care under extreme environments. This ‘Space Omics’ program is an integral component of the larger Enhancing eXploration Platforms and Analog Definition (EXPAND) Program initiated through a partnership between the BCM Translational Research Institute for Space Health (TRISH), and NASA Human Research Program. This is a program aimed at collecting specimens from 8 or more individuals/year as well as gender and age matched controls. Rigorous Standard Operating Procedures (SOPs) for specimen collection have been developed along with a LIMS-enabled biobank that tracks specimen details and storage information. Development of assay standards for pre/ post flight travel are under way. For the first time in the space human research program, participants will be analyzed for reportable genetic variants in 205 genes including the ACMG59 genes. Pathogenic and likely pathogenic variants of these 205 genes are consistent with genetic diagnoses of a monogenic disorder or carrier status of an autosomal recessive or X-linked condition. Additionally, this panel also reports two risk alleles in the LPA gene and a limited set of pharmacogenomic alleles (star alleles) in 7 genes. A ‘Space Omics’ clinical-grade report will be returned to the participants unless the participants specifically opt-out of the data return. The ‘Space Omics’ program has so far enrolled four participants each from Axiom-2 and Polaris Dawn missions. Sample collections from crew, including whole blood, saliva, stool, body swabs and urine from Axiom-2 mission to ISS (5/21 2023 - 5/31/2023) are nearing completion. An overview of the ‘Space Omics’ program and results from multiple Omics assays will be presented. This project is supported by the Translational Research Institute for Space Health through NASA Cooperative Agreement NNX16AO69A.

Title: Metabolomic and genomic prediction of 12 common diseases in half a million individuals from three national biobanks

Authors:

J. Barrett, Nightingale Health Biobank Collaborative Group; Nightingale Hlth., Helsinki, Finland

Abstract:

Life expectancy is increasing faster than healthy life expectancy, leading to individuals in high-income countries living more years restricted by chronic diseases. Healthcare systems face a cost crisis as they try to provide an increasingly wide range of transformative but expensive treatments to the older, sicker populations. Better prevention, to complement new treatments, is essential to provide healthier lives and sustainable healthcare. Polygenic scores have received most attention, but there is also an opportunity to combine genetics with other easily measured biomarkers.

Here we present nuclear magnetic resonance metabolomics from half a million blood samples from three national biobanks. We built metabolomic risk scores that identify a high-risk group for each of 12 diseases that cause the most morbidity in high-income countries and show consistent cross-biobank replication of the relative risk of disease for these groups. For seven diseases (heart attack, stroke, lung cancer, type 2 diabetes, COPD, alcoholic liver disease, and liver cirrhosis) this high risk group has >3-fold increased risk of disease (ranging up to 10-fold for diabetes and liver disease), and there are existing medical and lifestyle interventions available to reduce future risk.

We compared these metabolomic predictions to best available polygenic scores and found that for all diseases except Alzheimer's and colon cancer, metabolomic scores are more strongly associated with future disease risk than polygenic scores. These two scores are additive, except for a weak negative ineraction in diabetes, and in the cases with good performance of both scores (e.g. diabetes, heart attack) a combined score is significantly better than either alone. Metabolomic scores also show stable association to disease incidence for 10 years of follow up, demonstrating that scores composed of multiple biomarkers capture stable long-term risk, and not just immediate fluctuations.

Several factors suggest that metabolomic, or metabolomic+polygenic risk prediction could be used in the real world. First, because our measurements are in absolute units, no within-cohort normalization or batch correction is required, in contrast to some other omic technologies. Second, the absolute calibration of our predictions in separate biobanks/countries from those where we train the models is similar to the portability of existing risk tools, like Framingham or QRISK. Finally, 28% of individuals were at high risk for one or more of the seven diseases with good prediction and available intervention, suggesting that population screening in a health-check setting could provide value.

Title: Genetic basis of missing metabolites in the general population overlaps with inborn errors of metabolism

Authors:

T. Faquih[1,2,3], M. Imtiaz[4], D. Mook-Kanamori[3], E. N. Landstra[4], A. N. Aziz[4], A. Van Hylckama Vlieg[3], R. Li-Gao[3], R. Noordam[3], D. W. van Heemst[3], M. Breteler[4], K. Willems Van Dijk[3]; [1]Brigham and Women's Hosp., Boston, MA, [2]Harvard Med. Sch., Boston, MA, [3]Leiden Univ Med Ctr, Leiden, Netherlands, [4]German Ctr. for Neurodegenerative Diseases (DZNE), Bonn, Germany

Abstract:

A commonly assumed reason for missing measurements in metabolomics measurements is levels of the metabolite below the limit of detection. These metabolites are often removed from the analysis or imputed. However, missing measurements of metabolites could also be caused by genetic mutations that disrupt metabolic pathways, such as observed in patients that suffer from inborn errors of metabolism (IEM). In this study, we aimed to identify SNPs associated with the presence of missing values in metabolites measured by the Metabolon untargeted metabolomic platform by performing a meta-analysis of genome wide association studies (GWAS) in two European cohorts (N = 4,759). We included metabolites likely missing due to genetic factors by limiting our analysis to 224 metabolites with missingness percentages between 10% and 90% in each study population. We identified 5455 significant metabolome-wide SNP-metabolite associations (P-value cutoff = $1.59x10^{-10}$) involving 41 independent SNPs associated with 33 metabolites. Among these results, 27 associations with 12 metabolites involved novel loci, not reported in previous large metabolomics GWAS. Notable examples were the association of rs2147896 in the *PYROXD2* gene with the relatively novel N2-acetyl,N6,N6-dimethyllysine metabolite (Beta=1.27; P-value=$3.1x10^{-85}$), and the association of rs34976817 and rs1454247 near the *UGT2B17* gene with cholic acid glucuronide (Beta=−0.94, P-value=$4.9x10^{-32}$; Beta=−1.038; P-value= $2.0x10^{-75}$). Furthermore, the reported 12 metabolites were primarily related to acylation and beta oxidation pathways. Importantly, based on our findings and by applying the “prioritization of candidate causal genes at molecular QTLs” (ProGeM) framework, we found that these novel associations are likely linked to IEMs and chronic kidney disease. In conclusion, we report that the missingness in 33 metabolites was associated with 41 genetic loci. These findings indicate that genetic factors can contribute to the presence of missing values in certain metabolites. We also report novel loci associations with 12 of the 33 metabolites, indicating potential causal relationship related to errors in metabolism.

Title: Catalog and Genetic Architecture of Circulating Metabolites from Trans-Omics for Precision Medicine (TOPMed) Program

Authors:

N. Wang[1], **T. Alkis**[1], T. Blackwell[2], R. P. Bowler[3], C. B. Clish[4], A. M. Evans[5], R. E. Gerszten[4], M. Grove[1], S. Hutton[5], R. Kelly[6], C. Kooperberg[7], M. G. Larson[8], D. Meyers[9], L. Raffield[10], V. Ramachandran[11], A. Reiner[12], S. Rich[13], J. Rotter[14], E. Silverman[6], A. Smith[2], J. Lasky-Su[6], K. Wong[5], NHLBI Trans-Omics for Precision Medicine (TOPMed)Metabolomics Working Group, H. Chen[1], B. Yu[1]; [1]The Univ. of Texas Hlth.Sci. Ctr. at Houston, Houston, TX, [2]Univ. of Michigan, Ann Arbor, MI, [3]Natl. Jewish Hlth., Denver, CO, [4]Beth Israel Deaconess Med. Ctr., Boston, MA, [5]Metabolon, Morrisville, NC, [6]Brigham & Women's Hosp., Boston, MA, [7]Fred Hutchinson Cancer Res. Ctr., Seattle, WA, [8]Boston Univ., Boston, MA, [9]Univ. of Arizona, Tucson, AZ, [10]UNC - Chapel Hill, CHAPEL HILL, NC, [11]Boston Univ Sch Med, Boston, MA, [12]Univ of Washington, Seattle, WA, [13]Univ. of Virginia, Scottsville, VA, [14]Lundquist Inst., Harbor-UCLA Med Ctr, Torrance, CA

Abstract:

Circulating metabolite levels reflecting the state of human health and disease can be impacted by genetic effects. The NHLBI Trans-Omics for Precision Medicine (TOPMed) Program has sponsored metabolomic measures in ~100,000 samples across multiple studies to promote discovery of causal molecular pathways and therapeutic targets. We initiated a standard operating procedure (SOP) to harmonize metabolite data across TOPMed studies. In Phase 1, we catalogued 1,730 circulating metabolites from two metabolomics cores (25,058 samples; 53% females) and made them accessible through TOPMed portal.

Metabolite levels are heritable. However, their genetic architectures are not fully understood, including the generalizability of findings from European ancestry dominant studies, and the identification of sex-specific metabolic signatures. Whole genome sequencing (WGS) data were available in 16,359 samples (54% females) who had metabolite data from eight studies, including African, Asian, European, and American ancestries. We performed single variant analyses (minor allele frequency ≥ 0.5%) on 1,135 circulating metabolites (missing rate < 50%), using sex-pooled and sex-stratified approaches by GMMAT pipeline on BioData Catalyst (BDC).

We discovered 147,160 variant-metabolite pairs of associations (1,429 independent loci across 667 metabolites with $P < 4.4 \times 10^{-11}$). Among the associations mapped to well-known genes, four significant loci (*CPS1*, *ALDH1L1*, *PSPH*, *GCSH*) play critical roles on glycine metabolism. We also identified potential novel loci that require further investigation, e.g., *SLC22A24* was associated with 11-beta-hydroxyetiocholanolone glucuronide levels. We observed sex-specific genetic associations in ~10% metabolites. Sex-stratified analysis identified 2,414 variant-metabolite pairs involving 194 independent loci and 74 metabolites (at a Bonferroni-corrected $P = 2.5 \times 10^{-4}$). We confirmed that *CPS1* has a stronger effect on glycine levels in females than males. We also identified potential novel sex-specific loci with genetic effects in only one sex group, e.g., *GSPT1* for N-acetylglycine in females, *ABCC1* for Glutarylcarnitine (C5-DC) in males. We also detected potential novel association on chromosome X, i.e., *ARSD* for ascorbic acid 3-sulfate in females. The analytical pipeline is accessible through BDC, while sex-pooled and sex-stratified summary statistics are accessible through dbGaP Exchange Area.

In summary, we created a catalog for TOPMed metabolomics data and identified potential novel sex-pooled and sex-specific genetic associations contributing to our understanding of human circulating metabolites.

Session 111: Polygenic scores: Mediators, modifiers, and meanings

Location: Conv Ctr/Room 207A/Level 2

Session Time: Saturday, November 4, 2023, 1:00 pm - 2:00 pm

Title: Metabolite mediators of polygenic risk for obesity traits in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).

Authors:

V. Buchanan[1], H. Highland[2], C. L. Avery[1], A. Howard[3], S. Buyske[4], J. Cai[3], M. Daviglus[5], C. Isasi[6], R. Kaplan[6], D. Kim[1], M. Graff[7], S. Berndt[8], R. Smit[9], J. Arias[10], R. Loos[11], Q. Qi[12], R. Rohde[3], J. Rotter[13], L. Van Horn[14], B. Yu[15], C. Kooperberg[16], E. Boerwinkle[17], A. Justice[18], G. Chittoor[18], E. Wilson[1], K. North[7], K. Young[7]; [1]Univ. of North Carolina - Chapel Hill, Chapel Hill, NC, [2]Univ North Carolina at Chapel Hill, Chapel Hill, NC, [3]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [4]Rutgers Univ, Piscataway, NJ, [5]Inst. for Minority Hlth.Res., Univ. of Illinois Coll. of Med., Chicago, IL, [6]Dept. of Epidemiology and Population Hlth., Albert Einstein Coll. of Med., Bronx, NY, [7]Univ North Carolina, Chapel Hill, NC, [8]Natl. Cancer Inst., Rockville, MD, [9]Icahn Sch. of Med. at Mount Sinai, New York, NY, [10]Div. of Cancer Epidemiology and Genetics, Natl. Cancer Inst., Bethesda, MD, [11]Univ. of Copenhagen, Copenhagen, Denmark, [12]Albert Einstein Coll. of Med., Bronx, NY, [13]Lundquist Inst., Harbor-UCLA Med Ctr, Torrance, CA, [14]Dept. of Preventive Med., Feinberg Sch. of Med., Northwestern Univ., Chicago, Chicago, IL, [15]Dept. of Epidemiology, Univ. of Texas Hlth.Sci. Ctr. at Houston, Houston, TX, [16]Fred Hutchinson Cancer Ctr., Seattle, WA, [17]Univ. of Texas Hlth.Sci. Ctr. at Houston, Houston, TX, [18]Geisinger, Danville, PA

Abstract:

Obesity is a substantial public health burden but underlying metabolic characteristics and links to genetic risk remain unknown. US Hispanic/Latino adults are disproportionately affected yet understudied. We aim to identify metabolites associated with body mass index (BMI) and waist-to-hip ratio (WHR) and assess metabolite mediation of polygenic risk scores for these traits (PRS_{BMI} and PRS_{WHR}) in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). In 2,382 individuals (59.6% female), we tested associations between 640 untargeted baseline metabolites (Metabolon) with BMI and WHR at follow-up, to establish temporality. We carried forward significant metabolites ($p<0.05/640$) for mediation analyses of both PRS with their respective traits. For BMI, 203 metabolites were significant in females, 125 in males, and 188 in the sex-combined meta-analysis. For WHR, 54 were significant in females, 2 in males, and 43 in the meta-analyses (Bonferroni-corrected $p<0.05/640=7.81x10^{-5}$). Results were robust to diet and physical activity adjustment. PRS_{BMI} was significantly associated with BMI in females ($p=8.40x10^{-41}$, $R^2=10.9\%$) and males ($p=2.17x10^{-25}$, $R^2=9.8\%$); 17 of the 203 BMI metabolites in females and 4 of the 125 in males significantly mediated the PRS_{BMI} association with BMI. Three were mediators in both sexes: **3beta-hydroxy-5-cholestenoate** is a primary bile acid biosynthesis lipid that has been negatively associated with obesity and insulin resistance. **Glutamate** is an amino acid positively associated with BMI; dietary consumption in the form of MSG (a common additive) has been associated with weight gain in animal and human studies. **Mannose** is a carbohydrate that can be converted to and from glucose that has been positively associated with BMI. PRS_{WHR} was significantly associated with WHR in females ($p=9.32x10^{-22}$, $R^2=4.8\%$) and males ($p=1.13x10^{-11}$, $R^2=2.7\%$). Three WHR metabolites in females significantly mediated the PRS_{WHR}-WHR association: (**1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)**, **gamma-glutamylvaline**, and **sphingomyelin (d18:2/24:2)**). Gamma-glutamylvaline, a peptide with antidiabetic and anti-inflammatory effects, was associated with higher WHR and BMI and significantly mediated the PRS_{BMI} association with BMI in females, while the other two metabolites did not. The metabolome is strongly associated with obesity traits in HCHS/SOL, and some metabolites mediate PRS associations with these traits, helping inform causal paths from genetics to phenotypes, and pointing to potential modifiable targets for obesity prevention and intervention.

Title: The interplay between Long QT variants and QT prolonging medications on the risk for Long QT syndrome.

Authors:

L. Stalbow[1,2,3,4], M. Lui[4,1], R. Smit[5,2,3,4,6], R. Loos[5,2,3,4,6], B. Gelb[4,7,8], A. Kontorovich[4,9]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]The Charles Bronfman Inst. for Personalized Med. at the Icahn Sch. of Med. at Mount Sinai, New York, NY, [3]The Genetics of Obesity and Related Metabolic Traits Program at the Icahn Sch. of Med. at Mount Sinai, New York, NY, [4]The Mindich Child Hlth.and Dev. Inst. at the Icahn Sch. of Med. at Mount Sinai, New York, NY, [5]The Novo Nordisk Fndn. Ctr. for Basic Metabolic Res., Faculty of Hlth.and Med. at the Univ. of Copenhagen, Copenhagen, Denmark, [6]The Dept. of Environmental Med. and Publ. Hlth.at the Icahn Sch. of Med. at Mount Sinai, New York, NY, [7]The Dept. of Pediatrics at the Icahn Sch. of Med. at Mount Sinai, New York, NY, [8]The Dept. of Genetics and Genomic Sci. at the Icahn Sch. of Med. at Mount Sinai, New York, NY, [9]The Cardiovascular Res. Inst. at the Icahn Sch. of Med. at Mount Sinai, New York, NY

Abstract:

Introduction: With increasing access to large-scale population and hospital based biobanks, it is possible to study rare diseases in unselected populations. Long QT syndrome (LQTS) is a disease that can be congenital or acquired and, if not correctly identified and managed, can cause sudden cardiac death. Variants in three genes (*KCNQ1/KCNH2/SCN5A*) cause the majority of congenital LQTS. Here, we examine the effect of such monogenic variants on LQTS risk, and whether their effect is influenced by a polygenic susceptibility to a longer QT interval on the electrocardiogram, or medications known to increase the risk for QT prolongation.
Methods: Through exome sequencing of 31,057 individuals in the Mount Sinai Bio*Me* biobank, we identified carriers of ACMG (likely) pathogenic LQTS variants. Using results from a multi-ancestry QT interval GWAS (Young 2022, N > 250k), we built and standardized a QT polygenic risk score (PRS) for Bio*Me* participants, using PRS-CS. We then identified use of medications known to cause QT prolongation. We generated and standardized a phenotype risk score (PheRS) to quantify expression of LQTS phenotypes, as measured through ICD codes. Using linear regression, we estimated the association between LQTS variant carrier status, the PRS, and exposure to QT prolonging medications with the risk of LQTS (standardized PheRS score), separately, as well as combined.
Results: Fifty seven individuals (0.2%) harbored a LQTS variant (N_{KCNQ1} = 37, N_{KCNH2} = 14 and N_{SCN5A} = 6). Of those, only one person had an ICD diagnosis of LQTS. Genotype-positive individuals had higher LQTS PheRS than genotype-negative (0.48 S.D., p = 0.0002). KCNH2+ individuals had the highest scores (0.63, p = 0.018), followed by KCNQ1+ (0.47, p = 0.004), and no significant effect was seen in the SCN5A+ group (0.22, p = 0.58) as compared to the genotype-negative group. The PRS was not associated with the PheRS score (-0.036, p = 0.18), indicating that common variation linked to QT duration does not increase risk for LQTS. Use of QT prolonging medications was associated with higher PheRS; per increase in medication risk class (none, low, medium, high), there was a 0.19 ($p < 2x10^{-16}$) increase in the score. We observed an interaction effect between monogenic genotypes and medications (p = 0.036), such that per increase in medication risk class, LQTS variant carriers had a 0.34 (p = 0.026) increase in LQTS PheRS.
Conclusion: Both monogenic LQTS variants and QT prolonging medications increase the risk of manifesting LQTS features, and when combined, interact to exacerbate LQTS risk. Clinically, these findings may have implications for pharmacovigilance in individuals harboring high risk LQTS variants.

Title: Have economic evaluations of PRS considered the impact of uncertainty on patient outcomes and value?

Authors:

S. Jiang[1], G. F. Guzauskas[1], S. Garbett[2], J. A. Graves[2], M. S. Williams[3], J. Hao[3], L. K. Jones[3], J. Zhu[2], J. F. Peterson[2], D. Veenstra[1]; [1]Univ. of Washington, Seattle, WA, [2]Vanderbilt Univ. Med. Ctr., Nashville, TN, [3]Geisinger, Danville, PA

Abstract:

Introduction: Polygenic risk scores (PRS) have shown promising predictive power in identifying individuals at high-risk for disease to guide personalized screening and treatment. However, the accuracy and precision of PRS are limited by issues inherent in observational genome-wide association studies, such as sample size and analytical techniques, and by heritability. Understanding how uncertainty in estimated disease risk influences projected clinical and economic outcomes is needed before clinical adoption and reimbursement of PRS. Yet it is unclear whether and how published economic evaluations to date have incorporated these considerations.
Methods: We conducted a targeted review for PRS decision modeling studies published between January 2018 and June 2023 by searching the PubMed database using the keywords “polygenic/genetic risk score” and “cost-benefit/utility/effectiveness”. Our analysis focused on whether and how these studies considered an uncertain PRS-predicted risk in one-way and probabilistic sensitivity analyses and examined how such uncertainty impacted the value of PRS.
Results: We identified 16 studies meeting our inclusion criteria: 4 in colorectal cancer screening, 4 in breast cancer screening, 4 in prostate cancer screening, 2 in guiding statin therapy to prevent cardiovascular diseases, 1 in guiding glucose control treatment to prevent nephropathy, and 1 in glaucoma screening. Among the 16 studies, we found that only 2 articles explicitly incorporated an uncertain PRS-predicted risk in sensitivity analyses. Of these, 1 study indicated uncertainty in PRS-predicted risk was the 9th most influential factor on the value of using PRS to guide statin therapy to prevent coronary artery disease, and the other study showed that uncertainty in PRS-predicted risk was the 3rd most influential factor on the value of using PRS to guide glucose control to prevent nephropathy.
Conclusion: Examining uncertainty, particularly for key risk variables, is recommended good practice for economic evaluation in healthcare. To date, published economic evaluations have not consistently incorporated the uncertainty in PRS-predicted risk or examined how this uncertainty impacts the value of PRS. Future economic modeling studies should explicitly incorporate these considerations to facilitate evidence-based policy making for PRS.

Title: Characterizing the interaction between sexual trauma and polygenic scores across mental health conditions

Authors:

A. Lake[1], L. Davis[2]; [1]Vanderbilt Univ. Sch. of Med., Nashville, TN, [2]Vanderbilt Univ Med Ctr., Nashville, TN

Abstract:

Sexual trauma is a major non-genetic risk factor for many physical and mental health conditions, including highly heritable and polygenic psychiatric diseases such as schizophrenia, major depressive disorder (MDD), and bipolar disorder (BPD). Few studies have examined the joint impact of sexual trauma and genetic predisposition on psychiatric disease risk. Here, in a study of 77,566 individuals with genotyping data linked to electronic health records from the Vanderbilt University Medical Center biobank (BioVU), we conducted trauma-by-polygenic score (PGS) interaction analyses using logistic regression with schizophrenia diagnosis as the outcome. Sexual trauma disclosures were identified using a validated phenotyping algorithm applied to de-identified clinical notes. We conducted separate analyses for individuals with European (N=65,261) and African (N=12,305) genetic ancestry. European-ancestry analysis revealed a strong main effect of trauma (OR=95.3, 95% CI: 62.5, 145.2). We observed a significant trauma-by-PGS interaction effect on schizophrenia diagnosis (interaction coefficient $p<0.05$), with a stronger main effect of schizophrenia PGS in individuals with no disclosures of sexual trauma (OR=1.84, 95% CI: 1.58, 2.14) compared with those disclosing sexual trauma (OR=1.25, 95% CI: 0.81,1.92). Sexual trauma was associated with increased odds of schizophrenia in the African-ancestry cohort (OR=42.4, 95% CI: 21.7, 82.9), but no significant interaction was detected in this group. This work suggests that schizophrenia PGS may be a greater risk factor in the absence of additional risk factors including traumatic exposures such as sexual trauma. Ongoing work will use polygenic scores for MDD and BPD in PGS-by-trauma interaction analyses of the respective conditions to assess whether the independent, joint, and interacting contributions of genetic risk and trauma disclosures are similar across other major heritable mental health conditions with higher prevalence.

Session 128: Advances in applied ancestry and admixture

Location: Conv Ctr/Room 147A/Level 1

Session Time: Sunday, November 5, 2023, 8:30 am - 9:30 am

Title: The role of admixture in the rare variant contribution to inflammatory bowel disease

Authors:

C. Astore[1], S. Sharma[1], S. Nagpal[1], D. Cutler[2], S. Brant[3], S. Kugathasan[2], D. McGovern[4], I. Jordan[1], G. Gibson[1]; [1]Georgia Inst. of Technology, Atlanta, GA, [2]Emory Univ., Atlanta, GA, [3]Rutgers Robert Wood Johnson Med. Sch., New Brunswick, NJ, [4]Cedars-Sinai Med. Ctr., Los Angeles, CA

Abstract:

Identification of rare variants involved in complex, polygenic diseases like Crohn's disease (CD) has accelerated with the introduction of whole exome/genome sequencing association studies. Rare variants can be used in both diagnostic and therapeutic assessments; however, since they are likely to be restricted to specific ancestry groups, their contributions to risk assessment need to be evaluated outside the discovery population. Prior studies implied that the three known rare variants in *NOD2* are absent in West African and Asian populations, and only contribute in African Americans via admixture. Whole genome sequencing (WGS) data from 3,418 African American individuals, 1,774 inflammatory bowel disease (IBD) cases and 1,644 controls, were used to assess odds ratios and allele frequencies (AF), as well as haplotype-specific ancestral origins of European-derived CD variants discovered in a large exome-wide association study (ExWAS). Local and global ancestry was performed to assess the contribution of admixture to IBD contrasting European and African American cohorts. Twenty-five rare variants associated with CD in European discovery cohorts are typically five-fold lower frequency in African Americans. Correspondingly, where comparisons could be made, the rare variants were found to have a predicted four-fold reduced burden for IBD in African Americans, when compared to European individuals. Almost all the rare CD European variants were found on European haplotypes in the African American cohort, implying that they contribute to disease risk in African Americans primarily due to recent admixture. In addition, proportion of European ancestry correlates the number of rare CD European variants each African American individual carry, as well as their polygenic risk of disease. Similar findings were observed for 23 mutations affecting 10 other common complex diseases for which the rare variants were discovered in European cohorts. European-derived Crohn's disease rare variants are ultrarare in African Americans and contribute to disease risk mainly due to admixture, which needs to be accounted for when performing cross-ancestry genetic assessments.

Title: Approximate haplotype-based methods reveal fine-scale population structure, ancestry, and admixture in 70,000 Chinese.

Authors:

S. Morris, K. Lin, Z. Chen, R. G. Walters; Univ. of Oxford, Oxford, United Kingdom

Abstract:

To unravel the complete history of a population, it is crucial to have an understanding of fine-scale population structure, ancestry proportions, and admixture events, particularly when investigating localised demographic events. Haplotype-based methods can be particularly useful at this, as they capture information on rare variants which have not been genotyped. However, extending these methods to inform on population history in large datasets poses challenges due to their substantial computational complexity. Therefore, employing more efficient approximate methods to explore the population history of biobank-scale datasets is necessary. One example is the method PBWT-paint, which leverages the Burrows-Wheeler transform to rapidly assign specific haplotypes at each genomic position to individuals in large cohorts with only a fractional reduction in accuracy over exact methods.

We have used PBWT-paint to investigate fine-scale population structure, ancestry, and admixture in a cohort of 71,591 samples from the China Kadoorie Biobank (CKB), achieving resolution not previously reported for an East Asian population. We identify fine-scale genetic population structure, to a resolution of less than 5km, discriminating even between neighbouring recruitment centres in both rural and urban regions. We also identify historical admixture events likely associated with Silk Road migrations around 1000 AD and with more recent migrations from South-East Asia to Southern China in the 18th Century. Additionally, we observe evidence of shared genetic ancestry between ethnic groups (such as the Tu, Tujia, Uyghur, and Dai) and Han Chinese groups in different regions across China.

Previous studies have shown that admixture and genetic stratification can confound association studies - our work shows that even in a relatively homogenous cohort such as CKB, substantial fine-scale genetic structure exists. Current work in CKB is exploring the extent to which the above findings can help improve the performance of future GWAS.

Title: Identifying acute lymphoblastic leukemia risk loci in latino children via admixture mapping.

Authors:

J. Langie[1], S. Jeon[1], X. Ma[2], C. Metayer[3], A. J. de Smith[1], J. L. Wiemels[1], C. W. K. Chiang[1]; [1]Univ. of Southern California, Los Angeles, CA, [2]Yale Univ., New Haven, CT, [3]Univ. of California Berkeley, Berkeley, CA

Abstract:

Acute lymphoblastic leukemia (ALL) is the most common childhood cancer, with Latino children having up to 1.4 times the rate of ALL compared to non-Latino White children. The etiology of ALL is complex and this disparity in risk has not been fully explained by environmental factors, suggesting the role of genetic variants and ancestries. For instance, variants in *IKZF1*, *ARID5B*, *GATA3*, *PIP42KA*, and *ERG* have been found to have greater risk allele frequencies or population-specific risk effects in Latinos. Indigenous American ancestries have also been previously associated with ALL risk, relapse, and poor prognosis. To elucidate the genetic and ancestral etiology of ALL, we conducted the first admixture mapping analysis of childhood ALL including 1930 cases and 8520 controls of self-identified Latino ethnicity. We uncovered putative admixture associations on chr 2, 7, 10, and 15 ($P=3.21 \times 10^{-5}$ to 7.32×10^{-8}) among top associations across the four loci). Loci on chr 2 and 15 were previously not known to be associated with ALL. Following imputation with the TOPMed reference panel and fine-mapping, the top associated variants at each locus showed substantial frequency differences between ancestries. For example, the top associated variant on chr 2 has a risk allele frequency of 43% in Native Americans from the Human Genome Diversity Panel, but a frequency of 0.10% in Non-Finnish Europeans in gnomAD. Including these variants in the admixture mapping model greatly attenuated the admixture signals at each locus, suggesting that they likely include the causal SNPs or their close proxies. The top associated variants on chr 2, 7, and 10 passed correction for regional multiple testing burden and mapped to genes *MGAT5*, *IKFZ1* and *ARID5B*, respectively. The top two SNPs on chr 15, which narrowly missed the regional threshold, mapped to genes *RAB11A* and *MEGF11*. In a separate Latino ALL cohort, we replicated the signals on chr 7 and 10; replication efforts in additional Latino cohorts are ongoing. In performing replication, we also implemented a critical quality control measure to address ancestry-driven imputation quality discrepancies that can arise from use of external controls. Our results suggest that ALL risk variants with higher frequencies in individuals with Indigenous American ancestries may contribute to the observed increased risk of ALL in Latino children. Investigation of such risk loci and method development for admixed populations can contribute to the identification of new target genes for ALL prediction and therapeutics and new insights for precision medicine, which will reduce the burden of ALL.

Title: Leveraging multiple fine-tuning datasets and genetic ancestry continuum information to harmonize PRS for admixed populations

Authors:

R. Bhukar[1], Y. Ruan[1], P. O'Reilly[2], C. Hoggart[3], W. Hornsby[4], P. Natarajan[5]; [1]Broad Inst., Cambridge, MA, [2]Icahn Sch. of Med., Mount Sinai, New York, NY, [3]Icahn Sch. of Med. at Mount Sinai, New York, NY, [4]Univ. of Michigan, Ann Arbor, MI, [5]Massachusetts Gen. Hosp., Boston, MA

Abstract:

Background: Admixture, or genetic overlap of multiple continental genetic ancestry clusters, is increasingly common. The accuracy of PRS for admixed populations are limited by the vast genetic heterogeneity of admixed populations and lack of the representation in both discovery GWAS and fine-tuning datasets for optimal PRS hyper-parameters. Previous research has shown that leveraging multiple GWAS from multiple populations can partly remedy the low accuracy caused by discovery GWAS mismatch. However, the performance of multi-GWAS PRS models relies on a fine-tuning dataset of typically a single matched population with the testing dataset. Thus, current PRS methods rely on ancestry discretization and insufficiently account for the continuum of genetic diversity that exists.
Methods: Here we introduce a novel method to generate a harmonized PRS for admixed populations. By leveraging multiple PRS fine-tuned with single-ancestry datasets and genetic ancestry continuum information like PCA, we build a PRS model combining multiple single-ancestry information tailored to each individual's ancestry composition. We tested our method on simulated data and large-scale biobank data, UK Biobank and All of Us, and generated PRS for multiple traits.
Results: Even without matched fine-tuning data, our method generates PRS for admixed populations of accuracy similar to fine-tuned using the matched fine-tuning datasets. For example, when predicting lipid traits in UKBB samples, PRS fine-tuned in matched admixed samples or generated using our method has a marginal (relative increase of mean partial R^2 is 5.0%) but statistically significant ($p<0.05$) increase compared with PRS fine-tuned in European samples. In summary, our method yields a harmonized PRS for single-ancestry and admixed individuals.

Session 129: Deep thoughts on brain genomics

Location: Conv Ctr/Room 207A/Level 2

Session Time: Sunday, November 5, 2023, 8:30 am - 9:30 am

Title: Decoding the Genetic Underpinnings of Brain Connectivity: An Integrative Genomic, Transcriptomic, and Proteomic Approach

Authors:

Y. Yang[1], H. Zhu[1], W. Guan[1], B. Zhao[2], Y. Li[1], J. Stein[1]; [1]UNC Chapel Hill, Chapel Hill, NC, [2]Univ. of Pennsylvania, Dept. of Statistics and Data Sci., Philadelphia, PA

Abstract:

Genome-wide association study (GWAS) of neuroimaging traits has identified many single nucleotide polymorphisms (SNPs) associated with brain white matter microstructures and brain intrinsic activity. However, the molecular mechanisms leading from genetic variation to brain structure alterations remain unclear. To identify molecular profiles associated with white matter microstructural connectivity and resting state functional connectivity phenotypes, we conducted a large-scale regulome-wide association study (RWAS), transcriptome-wide association study (TWAS), and a proteome-wide association study (PWAS), together to link brain structure and function associated SNPs to chromatin accessibility, messenger RNA expression, and protein abundance. Subsequently, we applied Bayesian colocalization and Mendelian randomization (MR) to identify potential causal regulatory elements and genes that contribute to white matter microstructural connectivity and resting state functional connectivity phenotypes. We identified 247 accessible peaks in RWAS, 553 significant genes in TWAS, and 40 proteins in PWAS associated with interindividual differences in brain structure and function. Our integrative analysis revealed that ICA1L was associated with white matter microstructural connectivity and resting state functional connectivity phenotypes, according to robust evidence at transcriptional, protein, and regulatory levels based on brain-derived data. We also identified NBEAL1 that was causally related to white matter microstructural connectivity and resting state functional connectivity phenotypes via its cis-regulated brain expression level. Altogether, these genes provide novel, promising targets for mechanistic studies to find new biomarkers for white matter microstructural connectivity and functional connectivity traits.

Title: A genome-wide association study of 10000 Japanese whole genome sequencing data to identify the risk genes for brain structure changes.

Authors:

P. Yang[1], T. Monjo[1], Y. Ling[1], T. Saito[2], S. Mugikura[2], S. Ogishima[2], F. Nagami[2], K. Kinoshita[2], S. Kuriyama[2], M. Yamamoto[2], T. Ando[1]; [1]Takeda Pharmaceutical Company Limited, Fujisawa, Kanagawa, Japan, [2]Tohoku Med. Megabank Organization, Tohoku Univ., Sendai, Miyagi, Japan

Abstract:

This study analyzed whole genome sequencing (WGS) data of more than 10,250 Japanese individuals from the Tohoku Medical Megabank Organization (ToMMo) dataset, integrating magnetic resonance imaging (MRI) brain scans. By performing genome-wide association studies (GWAS) between common variants (minor allele frequency > 1%) and 1,324 T1-brain imaging phenotypes, we identified more than 50 novel genetic loci that were significantly associated with various brain structures. WGS also enables us to detect rare deleterious variants related to various phenotypes. To identify potential risk genes of abnormal brain structures, we conducted association tests between rare protein-truncating variants (e.g. frameshift variant, stop gained variants; minor allele frequency < 1%) and 1,324 T1-brain imaging phenotypes by using a scalable linear-mixed model (i.e. Regenie) to increase statistical power. As a result, we found a considerable number of novel genes that were significantly associated with brain imaging phenotypes after Bonferroni correction. These findings provide new insights into the genetic basis of brain development. In addition, we compared our results with the UK Biobank dataset to explore the differences of the genetic associations of brain structure between Japanese and European populations. We found that the genetic loci associated with the brain structures were generally similar between the two populations, but there were also some notable differences. Specifically, we identified several genetic variants that were more strongly associated with brain structure in the Japanese population than in the European population, including a locus at 12q24 associated with caudate volume, which has been previously implicated in neurological disorders.
Overall, our study represents a major step forward in understanding the genetic basis of brain structure and its relationship to diseases in East Asian. By leveraging the unique resources of the ToMMo dataset, we have identified several novel genetic loci that may serve as promising targets for future research and drug development. Our findings also highlight the importance of genetic diversity in brain research across populations, as genetically diversified populations could identify novel genetic associations underlying brain structure and function.

Title: BigBrain: Genetic analysis of novel transcript expression in 13,061 human brain transcriptomes.

Authors:

K. BP[1,2,3,4], B. Z. Muller[1,2,3,4], W. H. Cuddleston[1,2,3,4], A. Réal[5], J. Humphrey[1,2,3,4], D. A. Knowles[5,6,7,8], T. Raj[1,2,3,4]; [1]Dept. of Genetics and Genomic Sci., Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Nash Family Dept. of NeuroSci. & Friedman Brain Inst., Icahn Sch. of Med. at Mount Sinai, New York, NY, [3]Ronald M. Loeb Ctr. for Alzheimer's Disease, Icahn Sch. of Med. at Mount Sinai, New York, NY, [4]Icahn Inst. for Data Sci. and Genomic Technology, Icahn Sch. of Med. at Mount Sinai, New York, NY, [5]New York Genome Ctr., New York, NY, [6]Data Sci. Inst., Columbia Univ., New York, NY, [7]Dept. of Computer Sci., Columbia Univ., New York, NY, [8]Dept. of Systems Biology, Columbia Univ., New York, NY

Abstract:

Background: The formation of a mRNA transcript is a nuanced process, involving multiple transcription initiation, splice and polyadenylation sites. The human brain's remarkable transcriptional complexity is genetically controlled, with ties to neurological disease risk. However, our grasp of genetic variants affecting transcript expression in the brain is limited, partly due to incomplete annotation. While current methods quantify novel splicing in short-read RNA-seq, they are unable to place the splicing event within the full transcript and thus fail to fully capture the complexity of splicing in the human brain. In contrast, long-read RNA-seq can characterize millions of full-length mRNA molecules and identify novel transcripts. Our study incorporates short-read and long-read RNA-seq of the human brain to catalog genetic variants impacting gene and transcript expression in *cis* and *trans*.
Methods: We have assembled a new genetic and transcriptomic resource called "BigBrain" containing 13,061 uniformly processed RNA-seq samples from 12 cohorts of various brain disorders. We quantified gene expression and splice junction abundance using standard approaches. We also created a transcript reference by integrating novel transcripts discovered in PacBio long-read RNA-seq from 9 dorsolateral prefrontal cortex samples (ENCODE) with annotated transcripts from GENCODE. We used Salmon for transcript expression quantification, LeafCutter for junction usage and tensorQTL to map expression QTLs (eQTLs), splicing QTLs (sQTLs) and transcript expression QTLs (teQTLs). We combined summary statistics across cohorts using random-effects meta-analysis.
Results: We have completed *cis* meta-analysis for eQTLs and sQTLs in 4,342 European samples, and started mapping teQTLs in 412 samples. We found 22,120 eGenes and 12,285 sGenes, in *cis*, at a qvalue < 0.05, indicating our meta-analysis boosts discovery rate 2-fold for eQTLs and 3-fold for sQTLs, compared to the average discovery rate per cohort. We discovered 5,917 genes harboring a significant teQTL, of which 447 genes were identified only in the novel long-read RNA transcripts. We found at least one transcript harboring a teQTL for 57 out of 287 schizophrenia loci with PPH4 > 0.7.
Conclusion: Our study presents the most extensive collection of bulk brain RNA-seq data to date. Utilizing a long-read reference, we have identified and prioritized previously unknown transcripts associated with loci linked to neurological diseases. We are actively expanding our analysis to generate a comprehensive catalog of genetic variants that alter transcript expression in both *cis* and *trans* across multiple cohorts.

Title: Single-cell multiomics integrative analysis identified shared and distinct genetic etiologies across neurodegenerative diseases

Authors:

O. Chiba-Falek, E. Shwab, D. Gingerich, Z. Man, J. Gamache, M. Garrett, G. Crawford, A. Ashley-Koch, M. Lutz; Duke Univ, Durham, NC

Abstract:

The human brain is composed of multiple heterogeneous cell subtypes. Age related neurodegenerative diseases (NDDs), such as late-onset Alzheimer's disease (LOAD) and Parkinson's disease (PD), are characterized by massive neuronal loss, accompanied by gliosis. Furthermore, it has been suggested that specific neuronal and glial cell subtypes and their interplay are involved in NDDs pathophysiology. Thus, it is imperative to study the brain on a cell subtype specific level. Although numerous loci have been associated with each of the different NDDs via GWAS, the genetic etiologies are yet to be fully elucidated and the causal genes and variants remain largely unknown. Our overarching goal is to untangle the genetic complexity of these diseases and to identify shared and distinct biological pathways across NDDs. We applied *parallel* single-nucleus (sn)multi-omics approach to profile gene expression (snRNA-seq) and chromatin accessibility (snATAC-seq) using simultaneously the same nuclei samples from LOAD, DLB, PD and normal brains (12/group). Differential expression analysis by Nebula identified common and unique cell-subtype specific differential expressed genes (DEGs) among NDDs, and discovered common impaired biological pathways in specific cell-types. These included pathways related to brain development in oligodendrocytes and to RNA metabolism, protein processes and catabolism in excitatory neurons. Next, we investigated the gene-regulatory landscape in the brain of each pathology by integrative single-cell genomic analysis. The results discovered disease-associated *cis* co-accessibility networks (CCAN), candidate *cis* regulatory elements (cCRE), their candidate target genes, and the enriched transcription factors (TF). Finally, we focused on a subset of these cell subtype-specific cCRE-DEG links that overlap with known GWAS loci and catalogued functional SNPs and short structural variants (SSVs) that change the binding of TFs to the cCRE sequences. To our knowledge this study represents the first systematic interrogation of gene dysregulation networks across NDDs in an unprecedented cell-subtype resolution. The results enhance our understanding of the shared and distinct genetic factors and biological processes underlying age related NDDs.

Session 130: Ode to tandem repeats in common disorders

Location: Conv Ctr/Room 202A/Level 2

Session Time: Sunday, November 5, 2023, 8:30 am - 9:30 am

Title: A phenome-wide association study of large tandem repeats and multicopy genes in the UK Biobank

Authors:

A. Sharp[1], P. Garg[1], B. Jadhav[1], W. Lee[1], E. Dolzhenko[2], A. Martin-Trujillo[1]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Pacific BioSci.s of California, Menlo Park, CA

Abstract:

The human genome contains tens of thousands of large tandem repeats (also called Variable Number Tandem Repeats, VNTRs) and hundreds of genes and large coding repeats that show common and highly variable copy number variation (CNV). Due to their large size and repetitive nature, these VNTRs and multicopy genes are generally recalcitrant to standard genotyping approaches, and as a result this class of variation is poorly characterized. In order to identify effects of multi-allelic CNVs on human traits, we used read depth from Illumina whole genome sequencing to profile copy number and performed a phenome-wide association study (PheWAS) of VNTRs and multicopy genes in ~168,000 unrelated European individuals from the UK Biobank. We identified 341 different traits associated with copy number of 632 VNTRs and multicopy genes (Bonferroni $p<0.05$). In support of our results, VNTRs showing trait associations were significantly enriched for expression and methylation QTLs and we replicated dozens of previously reported associations from the literature, including 83 traits related to the kringle coding repeat within *LPA*, all of which were consistent with its known role in regulating lipoprotein levels and risk of cardiovascular disease. We also observed many novel associations of biological interest, examples of which include increased copy number of *AMY1* genes (Salivary amylase, which converts starch to sugar in the saliva) with denture use and reduced copy number of a coding VNTR in *MUC1* (Mucin 1, which forms the mucus lining of the digestive tract) with duodenal polyps. Using fine mapping and conditional analyses approaches, we identified 240 pairwise instances where large multi-allelic CNVs were scored as the mostly likely causal variant for the observed associations. Of these, we observed multiple instances where the lead SNVs at these loci reported in prior GWAS of the same trait were predictive for copy number of the causal repeat, indicating that a subset of GWAS signals are explained by SNVs that are actually proxies in linkage with nearby multi-allelic repeats. Based on this, we performed a large-scale comparison of our results to the GWAS catalog and estimate that at ~2% of GWAS signals are best explained by SNVs that tag underlying multi-allelic CNVs. Our study indicates that VNTRs and multicopy genes that are missed by standard GWAS approaches contribute to diverse human traits and suggests that complex structural variants that are weakly tagged by cis-linked SNVs explain some of the so-called “missing heritability” of GWAS.

Title: Large tandem duplications in cancer result from transcription and DNA replication collision

Authors:

L. Yang[1], Y. Yang[1], J. Chou[2]; [1]Univ. of Chicago, Chicago, IL, [2]Univ. of California, San Francisco, San Francisco, CA

Abstract:

Somatic structural variations (SVs) are common in cancer. Although a small fraction of SVs in breast and ovarian cancers can be attributed to homologous recombination deficiency, the underlying molecular mechanisms for the vast majority of somatic SVs remain unclear. Here, we focus on the roles of transcription and DNA replication collisions in genomic instability in cancer. Such collisions are unavoidable in cells since both transcription and replication use the same DNA as template. We hypothesized that transcription replication collisions (TRCs), if not properly repaired, would lead to collapsed replication forks and result in SVs. To this end, we studied somatic SVs in 5994 high-coverage whole-genome sequenced primary and metastatic tumors from three independent pan-cancer cohorts. A total of 12 conserved SV signatures, representing independent molecular mechanisms, were deconvoluted from these cohorts using non-negative matrix factorization approach. We detected replicated-strand bias, the expected footprint of transcription-replication collision, in large tandem duplications (TDs) across multiple cohorts. This bias was only observed in expressed genes, consistent with TRCs depending on transcription activity. Large TDs were abundant in female-specific (breast, ovarian and uterus), upper gastric-intestinal tract and prostate cancers. They were associated with worse patient survival and TP53 and CDK12 mutations. CDK12 is a cyclin-dependent kinase (CDK) and a key regulator of transcription elongation. Deleting or suppressing CDK12 using CRISPR-Cas9 in prostate cancer cell lines not only increased RNA:DNA hybrids (R-loops), but also promoted TRCs, suggesting a mechanism by which dysregulation of a transcriptional CDK may lead to genomic instability. Finally, using existing large-scale drug screening data, we found that cancer cell lines with abundant large TDs were significantly more sensitive to the WEE1 inhibitor, MK-1775, which we experimentally validated in prostate cancer cells lacking CDK12. In summary, our data suggest that large TDs in cancer form due to impaired TRC repair and can be used as a biomarker for prognosis and treatment.

Title: Short tandem repeats and their phenotypic associations in 200,000 participants in the UK Biobank

Authors:

L. Fearnley[1,2,3], M. F. Bennett[1,2], H. Rafehi[1,2,3], M. Bahlo[1,2]; [1]Walter and Eliza Hall Inst. of Med. Res., Parkville, Australia, [2]Univ. of Melbourne, Parkville, Australia, [3]Murdoch Children's Res. Inst., Parkville, Australia

Abstract:

Introduction: Short tandem repeats (STRs) and their repeat expansions (REs) play a role in over 50 disorders, including frontotemporal dementia (FTD), amyotrophic lateral sclerosis (ALS), spinocerebellar ataxias (SCA), Huntington's disease (HD). Their detection and sizing in short-read next-generation sequencing data using bioinformatic methods is a challenging task in large cohorts.
One such cohort is the 200,000 whole genomes sequenced in UK Biobank (UKBB) participants. Participants in the UKBB have extensive medical histories, phenotype and demographic data allowing for unprecedented studies of STRs and REs in both the context of known disease and non-disease phenotype.
Methods: We used ExpansionHunter and ExpansionHunter Denovo to genotype STRs from 200,000 UKBB whole genomes, focusing on 62 known disease loci, novel STRs over 150nt in length, and a larger set of polymorphic loci in individuals with neurodegenerative diseases. The motifs of each STR genotype were then systematically checked for accuracy using k-mer counts in reads at each locus. Novel and known REs were tested for association with various phenotypes sourced from medical records, MRI data, and other UKBB measures.
Results: We found 5,043 individuals with REs in the pathogenic range for multiple neurodegenerative diseases, of which a minority were previously diagnosed with the appropriate disorder. Only 0.5% of C9ORF72-expanded individuals are diagnosed with FTD/ALS, 0.9% of autosomal dominant SCA RE carriers with ataxia, 6.2% of HTT-expanded individuals for HD, and 3.9% of *TCF4*-expanded individuals for FECD3. We present 421,007 loci with REs >150nt in length and their novel associations with phenotypes. We also found that diagnosis of neurological disorders including HD, FTD, ALS, and Parkinson's disease was complicated by co-occurrent REs causing similar symptoms.
Conclusions: Our findings reveal underdiagnosis, impending burdens of RE diseases, and multifaceted STR-phenotype variation in the UKBB. This is likely caused by a complex interplay between medical record granularity and completeness, the age of participants, access to testing services and specialists, and our evolving understanding of these disorders. This highlights the influence of underexplored STR variation on health and phenotype at population scale, and the complexity of RE interpretation and the penetrance of these variants. Our results provide evidence for a spectrum of RE linked neurodegenerative disease in large observational cohorts.

Title: Short tandem repeat calling in patients and populations: opportunities and challenges.

Authors:

H. Dashnow[1], L. Hiatt[1], B. Weisburd[2], A. Quinlan[1]; [1]Univ. of Utah, Salt Lake City, UT, [2]Program in Med. and Population Genetics, Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

Short Tandem Repeats (STRs) are 1-6 bp repetitive DNA sequences, while Variable Number Tandem Repeats (VNTRs) have larger repeating motifs. STRs and VNTR variants at more than 60 loci are implicated in Mendelian diseases such as Huntington's disease, Fragile X syndrome and spinocerebellar ataxia. Variation at many STR loci is associated with more common phenotypes such as autism, intelligence, and depression. Until recently, STR genotyping in whole genome sequencing data has been prohibitively challenging due to the size and instability of these variants and limited read lengths. Even with recent algorithmic improvements, precisely sizing large repetitive alleles remains error-prone. Additionally, for more recently discovered STR disease loci, pathogenic allele size thresholds and penetrance data are limited, making clinical interpretation difficult. For these reasons, STR expansions are currently under-explored in rare disease diagnosis, especially in pediatric cases. There is a clear need to establish strategies for the routine assessment of STR loci in rare disease genomics in the context of their population allele frequencies.Recent population-scale tandem repeat allele frequency resources have allowed us to generate new insights into the variation at these loci. In particular, gnomAD v3.1.2 added ExpansionHunter allele size estimates at 59 disease associated STR loci from more than 19,000 individuals. We show that several established pathogenic repeat expansions appear at higher allele frequency than expected based on their previously reported prevalence, suggesting their penetrance may be lower than expected. This is consistent with our experiences in the Undiagnosed Diseases Network, where we have observed and in many cases orthogonally verified pathogenic expansions in individuals without a relevant disease phenotype. Several of these individuals are unaffected adult family members, suggesting later onset or lower penetrance than previously reported. This indicates a need to interpret STR variants in the context of their allele frequencies in addition to clinical features.To this end, here we introduce a new STR reference website and machine-readable database that will house the latest information from the literature regarding metrics such as age of onset, penetrance, population allele frequencies, pathogenic motifs etc. Drawing on our experiences in the Undiagnosed Diseases Network, I will describe how to use this data to mitigate many of the existing challenges in interpreting STR results at known disease-associated loci.

Session 131: RNA in action: Form and function

Location: Conv Ctr/Room 145A/Level 1

Session Time: Sunday, November 5, 2023, 8:30 am - 9:30 am

Title: Compromised THOC2, a TREX mRNA export complex scaffold protein, leads to accummulation of R-loops, DNA damage and adverse neurodevelopmental outcomes in mice and men.

Authors:

J. Gecz[1], R. Bhattacharjee[1], L. Jolly[2], M. Corbett[3], I. Wee[1], S. Rao[1], A. Gardner[1], T. Ritchie[1], E. J. H. van Hugte[4], U. H. Ciptasari[5], S. Pilz[1], N. Nazri[1], C. van Eyk[3], J. E. Noll[1], M. White[1], D. Fornarino[1], C. Poulton[6], G. Baynam[7], L. E. Collins-Praino[1], M. F. Snel[1], N. N. Kasri[8], K. Hemsley[9], P. Q. Thomas[1], R. Kumar[1]; [1]The Univ. of Adelaide, Adelaide, Australia, [2]Adelaide Univ., Adelaide, Australia, [3]Univ. of Adelaide, Adelaide, Australia, [4]6Dept. of Human Genetics, Nijmegen, Netherlands, [5]Dept. of Human Genetics,, Nijmegen, Netherlands, [6]King Edward Mem. Hosp., Subiaco, Australia, [7]King Edward Mem. Hosp, Subiaco, Australia, [8]Radboudumc, Nijmegen, Netherlands, [9]SA Pathology (WCH Campus), North Adelaide, Australia

Abstract:

TREX (Transcription-Export) is a highly conserved multi-subunit protein complex that, besides its established role in mRNA export from nucleus to cytoplasm, is involved in transcriptional regulation, stem cell maintenance, 3` mRNA processing, mitotic progression, and genome stability in eukaryotic cells. TREX is composed of a six subunit THO core complex (THOC1, THOC2, THOC3, THOC5, THOC6 and THOC7) and eight accessory proteins. We have implicated the X-chromosome *THOC2* gene, which encodes the scaffold protein of the THO complex, in a clinically variable neurodevelopmental disorder (NDD) with intellectual disability (ID) as the core phenotype. To study the function of this essential eukaryotic gene and the mechanisms underlying its NDD outcomes, we generated a clinically-relevant mouse model based on a hypomorphic *Thoc2* exon37-38 deletion variant of a patient with ID, speech delay, hypotonia, and microcephaly. The *Thoc2* exon37-38 deletion male ($Thoc2^{\Delta/Y}$) mice recapitulate the core phenotypes of *THOC2* syndrome including smaller size and weight, and significant deficits in spatial learning, working memory and sensorimotor functions. The $Thoc2^{\Delta/Y}$ mice brain development is significantly impacted by compromised THOC2/TREX function resulting in R-loop accumulation, DNA damage and consequent cell death. Differentiated cortical neurons of $Thoc2^{\Delta/Y}$ mice show perturbed synaptic and electrophysiological properties underlined by increased DNA damage comparable to the primary cells of patients with *THOC2* pathogenic variants. Altogether, our data suggest that perturbed R-loop homeostasis, in stem cells and/or differentiated cells in mice and men, and DNA damage-associated functional alterations are at the root of *THOC2* syndrome. Overall, we suggest that *THOC2* is a major, conserved player in the efficient resolution of R-loops, a process essential for normal brain development and function.

Title: Genome-wide dysregulation of R-loops in Ataxia Telangiectasia neurological pathogenesis

Authors:

K. Westover[1], Y. Hou[2], Y. Li[2], B. Yao[1]; [1]Emory Univ., Atlanta, GA, [2]Emory Univeristy, Atlanta, GA

Abstract:

Ataxia Telangiectasia (AT), a neurodegenerative disease characterized by cerebellar degeneration of Purkinje cells that control balance and movement, affects up to 1 in 40,000 to 100,000 people worldwide. A recessive early childhood onset disorder, AT is caused by mutations within the ataxia telangiectasia mutated (ATM) threonine/serine kinase which plays crucial roles within the DNA damage response (DDR). However, the precise molecular mechanisms underlying AT pathogenesis and how ATM loss-of-function leads to deficient DDR remain elusive. R-loops, three stranded RNA-DNA structures composed of an DNA-RNA hybrid and a non-template DNA strand, have emerged as key components of double strand break (DSB)-induced DDR. Mounting evidence has documented critical roles of R-loops in both causing and responding to DSBs. As DSBs and the failure of their repair play major roles in the pathology of AT, R-loop dysregulation is likely to contribute to AT pathogenesis. One recently identified kinase substrate of ATM is methyltransferase like 3 (METTL3) protein, a N^6-methyladenosine (m6A) methyltransferase. m6A on the RNA strand of R-loops is present inside nuclei and affects R-loop formation during DSB repair. The relationship between ATM-METTL3 phosphorylation in response to DNA damage and regulation of R-loop formation through m6A deposition, which could play crucial roles in AT pathogenesis, has yet to be defined. Our preliminary data has demonstrated that induction of DNA damage in the form of DSBs, either through a radiomimetic chemical or irradiation, results in an increased accumulation of R-loops in 293T cells. In AT patient-derived neurons, we see a global trend towards a loss of R-loop loci compared to healthy controls. Many of these lost loci are rescued when the pathogenic mutation is corrected by genome editing and some are associated with AT symptom processes. We have obtained several AT patient-derived fibroblasts and lymphoblastoid cell lines and have successfully reprogrammed them into induced pluripotent stem cells (iPSCs). Currently we are investigating how loss of ATM in iPSC derived neuronal progenitor cells (NPCs) impacts DDR and R-loop accumulation in response to DSBs. We hypothesize that in AT, the lack of METTL3 phosphorylation by ATM could globally dysregulate R-loop formation and underly AT progression.

Title: G4mer: mapping the human RNA G-Quadruplexes and genetic variants that affect them

Authors:

F. Zhuang, D. Gutman, D. Wu, S. Jewell, Y. Barash; Univ. of Pennsylvania, Philadelphia, PA

Abstract:

RNA G-quadruplexes (rG4s) are RNA secondary structures known to play an important role in gene regulation. Despite its importance, the effects of genetic variants on rG4 formations and functions remain unexplored. To address this challenge, we introduce G4mer, a deep learning transformer-based model that predicts transcriptome-wide rG4 formations. We trained G4mer on the experimental data produced by the recent advances in sequencing technology detailing rG4 formations in the transcriptome. While existing computational methods have been developed to predict whether rG4s are likely to form on a given sequence, we show that G4mer outperforms other state-of-the-art models in predicting rG4 sequences and their sub-categories. Additionally, G4mer offers a computational approach to study the effect of variants on rG4 formation and the association of these variants with diseases. With G4mer, we map variants in the 5' and 3' untranslated regions that are predicted to break rG4 formations. We show that rG4-breaking variants, compared to those variants with little to no effect on rG4s, are under negative selection. Hence, suggesting the functional importance of rG4s. Further, using variants that are predicted to alter rG4 formations, we map and validate gene-disease associations in a biobank containing exome sequencing from approximately 44,000 individuals. Specifically, we find that 17.2% of rG4-altering variants that are strongly associated with diseases have strong associations with neoplasms. Out of these variants, 8 have strong associations with breast cancer and are found in both 5' and 3' UTR regions of both known genes associated with breast cancer and genes with new associations. Finally, to study the effect of disease-associated rG4-altering variants on protein expression, we use luciferase assays in two different cell lines. Our results show that rG4-altering variants that are significantly associated with diseases and are located upstream of the luciferase reporter gene affect reporter protein expression levels, alluding to a functional role of rG4s in gene regulation. With G4mer, we show that computational modeling, genetic analysis, and experimental validations can create a high quality map of functional rG4s in the human transcriptome and help shed light on their functional role

Title: dsRID: Editing-free in silico identification of dsRNA region using long-read RNA-seq data

Authors:

R. Yamamoto, L. Zhiheng, M. Choudhury, X. Xiao; UCLA, Los Angeles, CA

Abstract:

Endogenous dsRNAs, recognized by sensor proteins, falsely activate innate immune responses. Unwanted antiviral signaling can be prevented by Adenosine-to-Inosine (A-to-I) RNA editing. Accumulating evidence suggests that A-to-I editing on dsRNAs affects their immunogenicity, and is implicated in autoimmune diseases such as Alzheimer's disease (AD).Since the only known mechanism for A-to-I editing is through enzymes binding to dsRNA, A-to-I editing sites have been used to identify dsRNA regions. However, dsRNA structures that undergo low-level RNA editing may escape from identification by editing-based methods. For example, brain samples from AD patients exhibit a lower level of A-to-I editing globally. Such disease-specific dsRNAs with less editing may be potent activators for immune response impacting the disease condition.

To overcome these limitations, we developed a new approach, named double-stranded RNA Identifier (dsRID), to detect dsRNA regions in an editing-agnostic manner. This method is built upon a previous observation that dsRNA structures may induce region-skipping in RNA-seq reads, an artifact likely reflecting intramolecular template switching in reverse transcription. Leveraging this observation and long-read RNA-seq data, we constructed a machine-learning model that extracts features from mapped reads and outputs predictions of dsRNA regions.

dsRID achieved in-silico identification of dsRNA regions independent of editing with high accuracy and precision (Average AUC of 0.95, AUPRC of 0.94 across 11 datasets). By applying this method, we identified 32391 novel dsRNA regions, 1.51 times more than dsRNA identified in the editing-based approach. We applied this method to long-read RNA-seq datasets derived from AD and control samples, which predicted novel dsRNAs with low RNA editing levels. Interestingly, there are higher fractions of dsRNA predicted in AD samples compared to control samples ($p=0.017$) and the overall expression of dsRNAs was higher in AD. This suggests that higher total production of dsRNAs might affect innate immune responses in the AD brain. Furthermore, we observed that dsRNA found specifically in AD samples have significantly lower editing levels (FC = 0.76, $p=4.9e-6$), showing that hypoediting happens not only in well characterized dsRNAs but also in dsRNAs that are not detected by the editing-based approach. Our findings emphasize the importance of identifying dsRNAs without relying on editing and demonstrate the utility of our editing-free dsRNA identification approach for studying dsRNAs associated with immune response and disease.

Session 132: Using proteomics to elucidate mechanisms underlying cardiometabolic traits and T2D

Location: Conv Ctr/Room 146B/Level 1

Session Time: Sunday, November 5, 2023, 8:30 am - 9:30 am

Title: Genetic analyses of the plasma proteome and metabolome in the same cohort from participants with African and European ancestry identify ancestry-specific type 2 diabetes risk associated molecular traits

Authors:

C. Yang, P. Gorijala, J. Timsina, L. Wang, C. Wang, E. Liu, J. C. Morris, Y. Sung, C. Cruchaga; Washington Univ. in St. Louis, Saint Louis, MO

Abstract:

Background: Type 2 diabetes (T2D) has different prevalence across different genetic ancestries. However, the genetic factors underlying T2D risk between diverse populations remains elusive. The latest population-scale T2D genome-wide association study (GWAS) (Mahajan et al 2022) reported 283 significant genetic loci in participants with European (EUR) ancestry. Another recent study (Vujkovic et al 2020) identified 28 loci in participants with African (AFR) ancestry. It will be important to map these genetic loci into the molecular traits, e.g. RNA expression, protein abundance, metabolite level. Multiple studies integrated such molecular traits with diseases and coined them as TWAS, PWAS, and MWAS. However, few studies covered more than one post-translational molecular trait and in multiple ancestries. **Method:** We first identified the plasma proteomic (SomaScan 7K) and metabolomic (Metabolon HD4) quantitative trait loci (QTLs) from participants with AFR (N=400) and EUR (N=2,300) ancestry respectively. We next associated the genetic variants with the ancestry-matched T2D risk GWAS (AFR-Vujkovic2020 & EUR-Mahajan2022) using the protein or metabolite as the intermediate molecular trait. We extended the TWAS framework (Gusev et al 2016) to test the disease association with the genetically predicted levels of protein (PWAS) and metabolite (MWAS). Moreover, we integrated the colocalization and Mendelian randomization (MR) results for each trait-T2D pair. **Result:** We identified hundreds to thousands of QTLs. Out of these findings, approximately 40% were novel QTLs compared to the previous studies. We uncovered ancestry-specific genetic architectures comparing results between EUR and AFR. We found ancestry-specific proteins and metabolites on T2D risk. For example, integrating the genetic architecture of proteomics with disease, proteins NFL and ZW10 were only significant in EUR-PWAS results but not AFR-PWAS using the ancestry-matched disease associations. Vice versa, FAM3D and TSEAR were only significant in AFR-PWAS. For metabolomics, metabolites GPE and N-acetyl-isoputreanine were only significant in EUR-MWAS, whereas GlcNAc sulfate was only significant in AFR-MWAS. By integrating PWAS and colocalization, we found 36 and four proteins were associated with T2D-risk in EUR and AFR, respectively. After accounting for MR results filtering out pleiotropic variants, five and one proteins remained. Four and one metabolites were associated with T2D-risk in EUR and AFR, respectively. **Conclusion:** Our study serves the first multi-omic multi-ancestry genetic study on T2D risk and may guide the future more precise ancestry-specific therapy.

Title: Proteome-wide Mendelian randomization identifies COL6A3-derived endotrophin as a clinically actionable mediator for the effect of obesity on coronary artery disease

Authors:

S. Yoshiji[1], T. Lu[2], G. Butler-Laporte[3], J. Carrasco-Zanini-Sanchez[4], Y. Chen[5], K. Liang[1], J. Willett[1], C-Y. Su[1], S. Wang[6], D. Adra[7], Y. Ilbudo[1], S. Takayoshi[1], V. Forgetta[7], Y. Farjoun[7], H. Zeberg[8], S. Zhou[5], M. Hultström[9], M. Machiela[10], V. Mooser[11], N. Wareham[4], N. Timpson[12], C. Langenberg[13], J. B. Richards[1]; [1]McGill Univ., Montreal, QC, Canada, [2]Univ. of Toronto, Toronto, ON, Canada, [3]Univ. of Oxford, Oxford, United Kingdom, [4]Univ. of Cambridge, Cambridge, United Kingdom, [5]McGill Univ., Montréal, QC, Canada, [6]SomaLogic, Boulder, CO, [7]Lady Davis Inst., Montreal, QC, Canada, [8]Karolinska Inst.t, Stockholm, Sweden, [9]Uppsala Univ., Uppsala, Sweden, [10]Natl. Cancer Inst., Rockville, MD, [11]McGill Genome Ctr., Montreal, QC, Canada, [12]Univ. of Bristol, Bristol, United Kingdom, [13]Queen Mary Univ. of London, London, United Kingdom

Abstract:

Obesity strongly increases the risk of cardiometabolic diseases; however, the underlying mediators of this relationship are not fully understood. Since obesity influences the human plasma proteome, we sought to identify circulating proteins mediating the effects of obesity on cardiometabolic diseases by integrating proteome-wide Mendelian randomization (MR), mediation analysis, colocalization, and single-cell RNA-sequencing analysis.
The study consists of four sections:
1) Step 1 MR: We performed a two-sample MR to estimate the effect of body mass index (BMI) on 4,907 circulating protein levels. For this, we used a BMI of 694,649 participants as the exposure and GWAS of 4,907 plasma protein levels from 35,559 individuals as the outcome. We found that 2,714 proteins were influenced by BMI (false discovery rate <0.5%) with no detectable horizontal pleiotropy or reverse causation.
2) Step 2 MR: Next, we evaluated the causal effects of the above-identified proteins (BMI-driven proteins) on coronary artery disease, stroke, and type 2 diabetes outcomes, again using MR. We used cis-pQTLs of BMI-driven proteins and the largest GWAS of these cardiometabolic outcomes. Following further filtering with colocalization and mediation analysis, we identified seven plasma protein mediators, including collagen type VI alpha-3 (COL6A3). COL6A3 was strongly increased by BMI ($\beta = 0.32$, 95% CI: 0.26-0.38, $P = 3.7 \times 10^{-8}$) and increased the risk of coronary artery disease (odds ratio = 1.47, 95% CI:1.26-1.70, $P = 4.5 \times 10^{-7}$) per s.d. increase in COL6A3 level.
3) Follow-up analysis for COL6A3: Domain-aware MR evaluating the effect of the C- and N-terminal effect of COL6A3 on CAD found that the C-terminal fragment of COL6A3 known as “endotrophin” mediated the effect. In single-cell RNA sequencing of adipose tissues and coronary arteries in humans, *COL6A3* was highly expressed in cell types involved in metabolic dysfunction and fibrosis.
4) Assessment of clinical actionability: Finally, multivariable MR revealed that body fat reduction can lower plasma levels of COL6A3-derived endotrophin and other protein mediators and reduce cardiometabolic risk, highlighting clinical translation of these findings.
In summary, we provide actionable insights into how circulating proteins mediate the effect of obesity on cardiometabolic diseases and prioritize endotrophin as a potential therapeutic target.

Title: Proteome-Wide Association Study Using Local and Distal SNPS and applied to Blood Cell and Lipid-Related Traitsin the Women's Health Initiative Study

Authors:

B. Chen[1], C. Lee[1], A. Tapia[2], A. Reiner[3], H. Tang[4], C. Kooperberg[5], Y. Li[6], L. Raffield[7]; [1]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [2]UNC - Chapel Hill, Durham, NC, [3]Univ of Washington, Seattle, WA, [4]Stanford Univ. Sch. of Med., Stanford, CA, [5]Fred Hutchinson Cancer Ctr., Seattle, WA, [6]Univ North Carolina, Chapel Hill, NC, [7]UNC - Chapel Hill, CHAPEL HILL, NC

Abstract:

Studying the genetically regulated component of protein abundance in the blood can improve our knowledge of biological processes through which genetic variants impact complex traits and diseases. In most Proteome-Wide Association Studies (PWAS), genetic variants near the protein-coding gene (+/- 1 Mb), also known as cis or local SNPs, are used to predict the protein level. These predicted protein levels are then associated with phenotypes of interest. However, proteins can be regulated through variants outside of the local region. Thus, we propose an intermediate GWAS step to identify protein quantitative trait loci (pQTL) throughout the genome and select variants for consideration in PWAS model training. This allows for the inclusion of trans or distal SNPs in protein-level prediction models. Here, we assess the prediction of 540 proteins measured using targeted Olink panels in 1,002 individuals from the Women's Health Initiative (WHI) with complete proteomics data across all panels and whole genome sequencing data. We split the individuals equally into a GWAS set (for identification of pQTL), an elastic net training set using all nominal pQTLs genome-wide ($p<0.0001$), and a testing set, for comparison of predicted and inferred protein levels. We compared the correlation between inferred and predicted protein levels in held-out testing set data (testing r^2) using this proposed approach, to the testing r^2 using only local SNPs in elastic net model training. The two approaches resulted in similar testing r^2's for most proteins (despite the increased sample size for model training using local SNPs only), but some proteins showed a noticeable increase in testing r^2 with our method compared to using local-only SNPs. For example, for cartilage acidic protein 1 (CRTAC1) the testing r^2 increased from 0.101 to 0.351 due to several large effect distal pQTLs. We demonstrate reproducible findings for imputed protein association with lipid and blood cell traits in both WHI participants without proteomics data and in UK Biobank utilizing our PWAS weights. Namely, angiopoietin-related protein 3 was significantly associated with triglyceride levels in the predicted and measured protein levels in WHI and predicted protein levels in UK Biobank. Our results suggest that consideration of distal SNPs may substantively increase prediction power for some circulating protein levels, even in modest sample sizes, with little reduction in prediction accuracy for proteins mostly influenced by local SNPs.

Title: Identification of plasma proteomic markers associated with polygenetic risk of type 2 diabetes and related complications

Authors:

D. Loesch[1], X. Jiang[1], B. Sun[2], H. Runz[3], C. Whelan[4], R. Holman[5], R. Mentz[6], F. Moura[7], S. Wiviott[7], M. Udler[8], I. Gause-Nilsson[9], J. Oscarsson[9], S. Petrovski[1], A. Nag[1], D. Paul[1], M. Inouye[10]; [1]AstraZeneca, Cambridge, United Kingdom, [2]Biogen, CAMBRIDGE, MA, [3]Biogen Inc., Cambridge, MA, [4]Janssen Res. & Dev., Boston, MA, [5]Diabetes Trials Unit, Radcliffe Dept. of Med., Univ. of Oxford, Oxford, United Kingdom, [6]Div. of Cardiology, Duke Univ. Sch. of Med., Durham, NC, [7]Thrombolysis in Myocardial Infarction (TIMI) Study Group, Brigham and Women's Hosp. and Harvard Med. Sch., Boston, MA, [8]Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [9]AstraZeneca, Gothenburg, Sweden, [10]Dept. of Publ. Hlth.and Primary Care, Univ. of Cambridge, Cambridge, United Kingdom

Abstract:

Introduction: Type 2 diabetes (T2D) is a heterogeneous disorder for which disease-causing pathways are incompletely understood. Here, we mapped genetic risk for T2D and its complications to proteins, mechanistic pathways and clinical consequences using proteogenomic data from population-scale biobanks and two cardiovascular outcome trials in T2D patients.
Methods: In 54,306 participants with genomics and proteomics data from the UK Biobank Pharma Proteomics Project (UKB-PPP), we generated polygenic scores (PGS) for T2D, body mass index, coronary artery disease (CAD) and chronic kidney disease (CKD), and 5 partitioned T2D PGS (pPS) that were based on clustering of known T2D risk variants into specific mechanistic pathways (βcell, lipodystrophy, liver lipid, obesity, and proinsulin; Udler, 2018). We tested each PGS and T2D pPS for association with the levels of 2,922 Olink plasma proteins in 36,642 unrelated UKB-PPP participants without prevalent T2D, CAD, or CKD. Then, we tested the PGS/pPS-associated proteins for association with the risk of T2D complications in 6,044 UKB-PPP participants with prevalent T2D, and the risk of incident complications in two cardiovascular outcome trials with proteogenomic data: DECLARE-TIMI58 (934 T2D patients/276 Olink proteins) and EXSCEL (3,026 T2D patients/4,746 SomaScan proteins). To assess causality, we evaluated the PGS/pPS-associated proteins within two-sample Mendelian randomization (MR) and statistical colocalization frameworks. Finally, we performed pathway enrichment analysis in the sets of PGS/pPS-associated proteins.
Results: We identified 926 unique proteins significantly associated with at least one PGS/pPS in the European-ancestry discovery subset of UKB-PPP, of which 554 unique proteins remained significant in the multi-ancestry replication subset. Most PGS/pPS-protein associations were polygenic with one notable exception - many (114/121) of the associations for the liver lipid T2D pPS were attenuated by *GCKR*-rs1260326, affirming the locus' pleiotropic effect. PGS/pPS-protein association patterns revealed shared pathways, e.g., complement cascade, cholesterol metabolism, IGF signaling. We found that the proteins underlying these shared pathways were consistently implicated in time-to-outcome-event analyses in the clinical trials, and many had a causal relationship with T2D or a complication in MR analyses.
Conclusions: Our proteogenomic study revealed novel proteins and mechanistic pathways underlying T2D and related complications, advancing our understanding of T2D pathobiology and identifying putative biomarkers for precision medicine approaches.

Session 133: AI meets human genetics

Location: Conv Ctr/Room 146B/Level 1

Session Time: Sunday, November 5, 2023, 10:00 am - 11:00 am

Title: AI-Enhanced Integration of Genetic and Medical Imaging Data for Risk Assessment of Type 2 Diabetes

Authors:

H-C. Yang[1,2], Y-J. Huang[2], C-h. Chen[1]; [1]Academia Sinica, Taipei, Taiwan, [2]Natl. Yang Ming Chiao Tung Univ., Taipei, Taiwan

Abstract:

Type 2 diabetes (T2D) is a global public health concern due to its increasing prevalence. Risk assessment and early detection of T2D play a vital role in improving individuals' health, reducing the burden on national health insurance, and enhancing well-being. This study leverages artificial intelligence, specifically eXtreme Gradient Boosting (XGBoost), to develop predictive models for T2D based on genetic and medical imaging data. The models aim to establish a prediction model and identify high-risk subgroups for T2D within a cohort of 68,769 Taiwan Biobank (TWB) participants. The approach integrates Polygenic Risk Score (PRS) and Multi-image Risk Score (MRS) with demographic factors and environmental exposures to assess T2D risk. The model's performance is evaluated using the Area Under the Receiver Operating Curve (AUC). Results demonstrate that genetic information alone is insufficient for accurate T2D prediction (AUC = 0.73), whereas medical imaging data, including abdominal ultrasonography, vertebral artery ultrasonography, bone density scan, and electrocardiography, significantly improves prediction accuracy (AUC = 0.89). The best-performing model integrates genetic, medical imaging, and demographic variables (AUC = 0.94), successfully identifying subgroups at high risk of developing T2D. The study also presents an online risk assessment website for T2D. In summary, this research represents the first integration of whole-genome and medical imaging data for T2D risk assessment. The genetic-only model outperforms previous genetic prediction studies, and integrating genetic and medical imaging information significantly enhances AUC. By utilizing artificial intelligence to analyze genetic, medical imaging, and demographic factors, this study contributes to early detection, precision health, and prevention of T2D.

Title: Inferring the distribution of fitness effects from genetic variation with convolutional neural network.

Authors:

L. Tran, R. Gutenkunst; Univ. of Arizona, Tucson, AZ

Abstract:

The distribution of fitness effects (DFE) of new mutations quantifies the input of mutations with certain selective effects (deleterious, neutral, advantageous) into natural populations and is fundamental to evolution by natural selection. In humans, the DFE is particularly important for understanding quantitative traits, such as the genetic architecture of complex diseases. The DFE had been estimated from genetic variation data in humans by several studies, all of which inferred the DFE from summary statistics describing some aspects of the data, such as the allele frequency spectrum and/or linkage disequilibrium statistics. While powerful, these existing approaches either neglect substantial information from the data not included in the chosen summary statistics or demand careful curation of the appropriate set of summary statistics.

Here we introduce the first deep learning approach for DFE inference using convolutional neural network (CNN) and a novel representation of genetic variation data. Single nucleotide polymorphism (SNP) alignment is represented by a three-dimensional tensor in which each layer corresponds to a functional class, such as synonymous and nonsynonymous coding variants. With this representation, our approach allows direct processing of genetic variation data in which the CNN implicitly learns the most informative features in the data, bypassing summary statistics selection and assumption.

We trained and validated the CNN using data simulated under varied gamma DFE distributions and human demographic histories. We found that the trained CNN achieves better inference accuracy on simulated data than a commonly used summary-statistics-based DFE inference method. Our result demonstrates that deep learning is a promising and powerful approach for fully harnessing the information available in modern genomic data sets to infer the DFE with higher precision.

Title: Deep learning enables gene discovery of ultra-rare coding variants for rare diseases

Authors:

D. Jordan, I. Forrest, B. Petrazzini, G. Rocheleau, R. Do; Icahn Sch. of Med. at Mount Sinai, New York, NY

Abstract:

Causal genes for many rare diseases have been discovered using exome or genome sequencing in Mendelian pedigrees, but this approach still fails to produce diagnoses in a majority of cases. One reason for this is the inherently small sample sizes of rare disease studies. Resources like Matchmaker Exchange have been developed to boost this sample size, and large population scale biobanks such as the UK Biobank and the NIH All of Us Initiative may boost it further. However, powerful state-of-the-art statistical genetics approaches remain out of reach. As one example, recent studies have performed gene discovery analysis in ultra-rare variants (minor allele frequency < 10^{-5}) in the UK Biobank, an approach that is only feasible with sample sizes in the tens or hundreds of thousands and is therefore inaccessible to even the most common rare diseases. In this study, we propose an approach to apply the enormous sample sizes of these biobanks towards gene discovery of ultra-rare coding variants for rare diseases using phenome-wide deep learning. We trained a convolutional neural network to impute rare disease diagnoses from structured electronic health record data in the UK Biobank, using a “masked phenotype modeling” approach analogous to the masked language modeling approach used to train language models like BERT. We then used this model to produce continuous diagnosis confidence scores for rare diseases, representing the phenotypic similarity of each participant to patients diagnosed with a particular rare disease. Unlike the corresponding case-control phenotypes, which may have fewer than 100 available cases even in an enormous biobank, these continuous scores can be calculated for hundreds of thousands of participants. Using these scores, we performed a phenome-wide gene discovery analysis of ultra-rare variants in 469,032 exomes from the UK Biobank using REGENIE. This analysis recapitulated known associations, including *LDLR* for lipid metabolism disorders and *BRCA2* for rare cancer syndromes; identified a new role for clonal hematopoiesis of indeterminate potential (CHIP) genes (*DNMT3A, TET2*, and *ASXL1*) in rare autoimmune, inflammatory, and infectious disease; and discovered causal genes for rare disorders with no previously known genetic etiology, including *LHFPL3* (congenital anomalies of lower limbs) and *ADAP1* (noninfectious uveitis). This study demonstrates the potential of deep learning in electronic health record data to power gene discovery for rare disease.

Title: A Graph Neural Network Approach for Disease Prediction based on Nation-wide Pedigree in over 6 Million Individuals with Extensive Health Information

Authors:

Z. Yang[1], M. Ferro[1], S. Wharrie[2], A. Liu[3], V. Anapaz[1], F. Wang[4], Z. Zheng[5], S. Kaski[2], A. Ganna[3]; [1]Univ. of Helsinki, Helsinki, Finland, [2]Aalto Univ., Helsinki, Finland, [3]Inst. for Molecular Med. Finland, Helsinki, Finland, [4]Finnish Inst. for Molecular Med., Helsinki, Finland, [5]Broad Inst. of MIT and Harvard, Boston, MA

Abstract:

Family history is an important risk factor for common diseases, but most current approaches only consider the presence of a single disease among close relatives to estimate disease risk in a target individual. Familial members share both genetics and common environmental effects for multiple diseases. Therefore, we hypothesised that by leveraging the totality of phenotypic information across an extended pedigree, we could reach higher prediction performance than considering only history for disease of interest among first degree relatives. Using Finnish registers, we reconstructed a nationwide pedigree including individuals born between 1831 and 2020 spanning eight generations. The largest family covers 6,255,587 individuals, connected by 9,222,737 pairwise relationships. For most individuals, health, medication purchase and socioeconomic information, more than 4000 features, were available since the 70's. By viewing each individual as a node and each relationship pair as an edge, we encoded the pedigree structure into network graphs, with expected genetic relationships as edges' weight priors and ~1500 most common health related measurements as node features. We subsequently trained graph neural networks (GNN) with the goal to predict 10-year risk for 5 diseases, including coronary artery disease (CHD), depression, asthma, type ii diabetes (T2D) and colorectal cancer, in a 20% held-out set. We show that, even without any information except for age and sex, the GNN approach could provide a good prediction of target individuals' 10 year CHD risk (AUC=0.80), better than a conventional clinical baseline, where only target individual's age and sex, and CHD history of first-degree relatives were considered (AUC=0.70). We used GNN-based interpretability approaches as well as logistic regression to understand contribution of features from each relative type on disease risk. We found while CHD related features among parents were strong predictors with typically larger maternal effects (e.g. mother's myocardial infarction events OR=1.15, $p = 2.13\text{x}10^{-121}$, father's OR=1.01, $p = 2.37\text{x}10^{-2}$), features from more distant relatives can also be strongly associated with target's CHD risk for even non-health related features (e.g.marriage status of aunt or uncles OR=1.83, p=$3.58\text{x}10^{-126}$). Novel deep learning approaches allow computationally trackable modelling on totality of phenotypic information across an extended pedigree to better understand how genetic along with shared environmental effects can impact disease risk. Such abundant pedigree information also makes it possible to carry out prediction with established genetic approaches such as BLUP.

Session 134: Dollars and DNA: Financial considerations of genetic testing

Location: Conv Ctr/Room 147A/Level 1

Session Time: Sunday, November 5, 2023, 10:00 am - 11:00 am

Title: Awareness and use of genetic testing in 2022 in the United States

Authors:

S. Makhnoon, M. Lee, T. Prasad, A. Badalamenti, T. Gurley, C. S. Skinner; UT Southwestern Med. Ctr., Dallas, TX

Abstract:

Introduction: Expectations that burgeoning genetic knowledge would revolutionize healthcare have long been stymied, but public interest in genetic testing - especially direct-to-consumer (DTC) tests - continues to grow. Since 2000, researchers have estimated awareness and use of genetic testing in the US. Here, we use data from the 2022 US population-based Health Information National Trends Survey (HINTS 6) to estimate awareness and use of germline genetic tests and add to the literature by reporting on four previously unreported types of genetic tests including reasons for testing. **Methods:** HINTS6 - a nationally representative survey of civilian, non-institutionalized US adults ages ≥18 - was fielded March-November, 2022 via web and on paper via mail, in English and Spanish. **Result:** Response rate was 28.1%. Of the 6,252 respondents, 81.4% were aware of genetic testing and 40% had undergone some type of testing themselves. Common information sources for genetic tests were internet/social media (60.5%), other media (53.8%), and family or friends (52.4%). Uptake was highest for ancestry testing (22.6%), followed by disease risk testing (15.9%), and prenatal testing (7.8%) and personal trait testing (6.2%). Notably, most testing occurred without involvement of genetic counselors (ranging from 9.8%-15%); tests were commonly ordered by a healthcare provider other than a genetic counselor (42.3% to 71.3%) or via direct order from a lab (14.8% to 47.9%). Common reasons for testing were to understand family ancestry (24.8%), follow doctor's recommendation (34.4%) and to learn more about disease risk (15.9%). Separate logistic regression models found use of ancestry testing was positively associated with being older, having a college degree (OR=2.05, p=0.04), earning ≥$75,000 (OR=2.05, p=0.01), and being multiracial (OR=2.23, p=0.02), but negatively associated with being Asian (OR=0.42, p<0.01). Use of personal trait testing was negatively associated with age 65-75 years (OR=0.44, p=0.04), having a family history of cancer (OR=0.51, p<0.001), and being Black (OR=0.55, p=0.04), but positively associated with being multiracial (OR=2.83, p<0.001). Use of specific disease testing was positively associated with having a personal history of cancer (OR=2.49, p<0.001). **Discussion:** Population awareness of genetic tests has grown since 2000s, perhaps in part due to growing popularity of DTC tests. In comparison, awareness of clinical genetic tests has remained steady fluctuating between 35% and 44% over the last 20 years. These metrics serve as indicators of the diffusion of genetic discoveries into communities and clinics.

Title: BRCA Testing Utilization and Cost among Privately Insured Adults Aged 18-64 Years in the United States, 2014-2021

Authors:

L. Shi[1], Z. Chen[1], S. Grosse[1], K. Kolor[1], C. Lu[2], J. Rodriguez[1]; [1]Ctr. of Disease Control and Prevention, Atlanta, GA, [2]Univ. of Sydney, Sydney, Australia

Abstract:

Objective: We estimate changes in utilization and cost for BRCA tests in privately insured US adults. **Methods**: Using claims data from employer-sponsored health plans and procedure codes, we identified BRCA1/2 testing during 2014-2021. We estimated the annual utilization rate, median total cost per enrollee, and the percentage of zero out-of-pocket (OOP) cost by sex among individuals aged 18-64 who were continuously enrolled for the calendar year. **Results**: Annual utilization rate of any BRCA testing among women increased by 14% from 294.2/100,000 in 2014 to 335.7/100,000 in 2015, with an additional 8% increase by 2019 to 361.1/100,000. Testing among women declined 29% in 2020 to 256.4/100,000 and increased by 9% to 278.2/100,000 in 2021. The utilization of any BRCA testing among men increased from 12.2/100,000 to 20.0/100,000 between 2014 and 2021. Among women who had any BRCA testing, approximately 92% received a full sequencing test, peaking at 95.3% in 2018. Among men who had any BRCA testing, the percentage of full sequencing tests increased from 69.8% in 2014 to 88.7% in 2021. Median total costs for BRCA testing decreased by 62.6% ($3,970 to $1,483) from 2014 to 2021. Approximately 80% of the women and 70% of the men had zero OOP spending for the testing. In addition, approximately 94% of the women and 67% of the men, on average, had diagnosis codes indicating either family history or personal history of breast, ovarian, or prostate cancer across all years, with approximately 61% of the women and 48% of the men reporting a family history of breast, ovarian or prostate cancer. Personal history of pancreatic cancer accounted for an additional 5% of men on average across the study period receiving BRCA testing. **Conclusion**: The substantial increase in BRCA testing rates among men reflects increased recognition that all individuals with personal or family history indicating an elevated risk of cancer should undergo testing, regardless of sex. Significant reductions in total payments were noted after the invalidation of BRCA patents in 2013 and in the context of new testing platforms and service delivery models, such as the introduction of panel tests and direct-to-consumer genetic tests. The reduction in cost was not associated with a marked increase in the number of women who were tested, which is not surprising because most individuals do not pay for the test. Testing rates in both sexes experienced sharp declines during the initial year of the COVID-19 pandemic and only partially rebounded in the subsequent year.

Title: Revealing the hidden costs: Exploring the financial toxicity of hereditary cancer syndromes

Authors:

S. Rajeziesfahani[1], **J. Sam**[2], C. Butkowsky[3], E. Reble[3], M. Clausen[3], R. Gopalakrishnan[3], B. Sparkes[1], V. Aguda[4], M. Aronson[5], D. Bishop[6], L. Dawson[1], A. Eisen[7], T. Graham[7], J. Green[8], C. Mighton[9], J. Pauling[10], C. Pavao[11], P. Pechlivanoglou[12], C. Remocker[11], S. Savas[1], S. Sun[13], T. Tiano[10], A. Tilley[6], K. Thorpe[1], K. Schrader[14], Y. Bombard[15], H. Etchegary[16]; [1]Mem. Univ. of Newfoundland, St. John's, NL, Canada, [2]Unity Hlth.Toronto - St. Michael's Hosp., North York, ON, Canada, [3]St. Michael's Hosp., Toronto, ON, Canada, [4]Unity Hlth.- St. Michael's Hosp., Toronto, ON, Canada, [5]Zane Cohen Ctr., Sinai Hlth.System, Toronto, ON, Canada, [6]Patient Partner, St. John's, NL, Canada, [7]Sunnybrook Hlth.Sci. Ctr., Toronto, ON, Canada, [8]Mem. Univ. of Newfoundland, Middle Cove, NL, Canada, [9]Univ. of Toronto & St. Michaels Hosp., Toronto, ON, Canada, [10]Patient Partner, Toronto, ON, Canada, [11]Patient Partner, Vancouver, BC, Canada, [12]The Hosp. for Sick Children, Toronto, ON, Canada, [13]BC Cancer, Vancouver, BC, Canada, [14]BC Cancer / Univ British Columbia, Co-Principal Investigators, Vancouver, BC, Canada, [15]Univ. of Toronto & St. Michaels Hosp., Co-Principal Investigators, Toronto, ON, Canada, [16]Mem. Univ. of Newfoundland, Co-Principal Investigators, St. John's, NL, Canada

Abstract:

Introduction: Financial toxicity, also known as financial distress or burden, refers to the harmful effects caused by the high costs of treatment on patients' well-being. Although financial toxicity is widely reported in cancer patients, there is limited research on its specific implications in patients with hereditary cancer syndromes (HCS), which account for approximately 10% of all cancer cases. This study aims to explore the direct and indirect financial toxicity associated with the two most common types of HCS: Hereditary Breast and Ovarian Cancer Syndrome (HBOC) and Lynch Syndrome (LS). Methods: Patients across 3 provinces in Canada with a confirmed molecular diagnosis of HBOC or LS were invited to participate in semi-structured qualitative interviews. Interpretive description was used to analyze the data. Results: Qualitative interviews were conducted with 73 patients (51 females, 21 males, 1 gender-diverse; age range 25-80yrs) diagnosed with HBOC (n= 39) or LS (n= 34). Participants described a range of financial toxicity. For many, accessing treatment and routine screenings within their province posed economic challenges. These difficulties mostly included travel costs and lost wages from time away from work. A small subset of patients faced substantial travel costs if travel to a distant medical facility or another province was required to access specialized equipment or health professionals (e.g., larger MRI machines, genetic professionals). Other financial impacts of HSC included expenses for fertility preservation procedures, reconstructive surgeries, and psychotherapy. Concerns about unidentified financial prospects in the future weighed heavily on the minds of many participants. The possibility of being unable to return to their jobs due to health limitations, choosing a different career path, or the potential financial impact of their passing created financial uncertainty and strain. Participants often relied on their families to help cope with financial challenges. This involved sharing the costs, alternate living arrangements, family members accompanying them to medical appointments, and seeking assistance with childcare. The role of health advocates and supportive employers emerged as crucial factors in mitigating financial burdens. Conclusion: Findings provide novel insights about the existence of financial toxicity for HCS in Canada, particularly regarding access to screening and treatment options, ongoing out-of-pocket expenses, and future economic uncertainties. Results highlight the need for the development of solutions to help address the financial toxicity of HCS.

Title: The estimated healthcare spend to identify 1 hereditary cancer syndrome patient with a deep intronic pathogenic variant by RNA sequencing is nearly 500x more costly than identifying 1 patient with multi-gene panel DNA testing

Authors:

E. Esplin, B. Heald, L. Fosler, D. Pineda-Alvarez, B. Johnson, S. Nielsen, N. Kamps-Hughes, L. Fresard, E. Borras, H. Kang, K. Nykamp, R. Nussbaum, C. Moretz; Invitae, San Francisco, CA

Abstract:

Background In a study of mRNA analysis, (Kamps-Hughes et al. J Mol Diagn 2022), 40/20,317 (0.2%) unselected patients undergoing germline oncology testing were found to have altered splicing without a variant detected in the reportable DNA range; of these, 8 (0.04%) were shown to have a pathogenic or likely pathogenic variant (PLPV). No studies have evaluated the potential economic impact of mRNA sequencing to detect deep intronic variants on the healthcare system. The aim of this study was to model the financial implications of mRNA sequencing for variant discovery.
Methods The median cost for a DNA multi-cancer gene panel test (MGPT) including an average cost of phlebotomy was estimated at $1501.75 (lower to upper quartile range $1,174.75-2,619.75, 2022 Genetic Test Price Transparency Report by Concert Genetics). Without a CPT code for mRNA analysis, the cost of adding RNA to MGPT was assumed to be the same as for MGPT. The total cost to the healthcare system of including mRNA testing in MGPT with phlebotomy was therefore $3003.50 (lower to upper quartile range $2,349.50-5,239.50) per patient. The average PLPV detection rate for MGPT in patients referred for germline oncology testing was estimated to be 10%.
Results Utilizing the 0.04% detection rate, the number of patients needed to screen to identify one patient with a deep intronic PLPV was 2,500, which means the median cost to the healthcare system to identify one patient with a deep intronic PLPV would be $7,508,750 (lower to upper quartile range $5,873,750-13,098,750). For this amount of money, a median of 5,000 (range 3,911-8,722) patients could have DNA MGPT with phlebotomy, which would result in the identification of 500 (391-872) patients with PLPVs.
Conclusions Recognizing the limitation that the cost of adding RNA to MGPT could only be estimated in our modeling, we find that the cost added by RNA analysis in order to find one patient with a deep intronic PLPV, is the same as the cost of finding ~500 patients with PLPVs by MGPT alone. This is an important consideration in a resource constrained healthcare system. It should be noted, this economic model assumed a universal testing approach and only addressed the use of RNA in hereditary oncology testing. The yield and economic considerations could differ in other clinical areas (e.g. hereditary retinal disease). Further studies are warranted to determine if mRNA sequencing can be more cost effectively targeted for variant discovery, which might offset these costs.

Session 135: Genetic diagnosis of neurodevelopmental disorders beyond the exome

Location: Conv Ctr/Room 207A/Level 2

Session Time: Sunday, November 5, 2023, 10:00 am - 11:00 am

Title: Deep intronic GAA repeat expansions in *FGF14* are a common cause of downbeat nystagmus syndromes.

Authors:

D. Pellerin[1,2], C. Wilke[3,4], A. Traschütz[3,4], F. Heindl[5], C. Ashton[1,6], M. Danzi[7], M-J. Dicaire[1], A. M. Hartmann[8], D. Rujescu[8], S. Zuchner[7], B. Brais[1], M. Strupp[5], M. Synofzik[3,4]; [1]McGill Univ., Montreal, QC, Canada, [2]UCL Queen Square Inst. of Neurology, London, United Kingdom, [3]Univ. of Tübingen, Tübingen, Germany, [4]German Ctr. for Neurodegenerative Diseases, Tübingen, Germany, [5]Ludwig-Maximilians Univ., Munich, Germany, [6]Royal Perth Hosp., Perth, Australia, [7]Univ. of Miami Miller Sch. of Med., Miami, FL, [8]Med. Univ. of Vienna, Vienna, Austria

Abstract:

Background: Downbeat nystagmus (DBN) is the most common form of acquired persisting nystagmus. The cause of DBN remains unknown in 30% of cases (idiopathic DBN). Intronic GAA repeat expansions in *FGF14* have recently been shown to cause spinocerebellar ataxia 27B, which is characterized by a late-onset slowly progressive cerebellar syndrome that is frequently associated with DBN. Furthermore, a previous GWAS has shown an association between a variation in *FGF14* and idiopathic DBN. We hypothesized that *FGF14* repeat expansions may represent a recurrent genetic cause of DBN syndromes in patients previously diagnosed with idiopathic DBN. ***Methods:*** We recruited 170 patients with idiopathic DBN and genotyped the *FGF14* repeat locus. Expansions of at least 250 GAA repeat units were considered pathogenic. We performed in-depth ocular motor, neurological, and disease evolution phenotyping. We also assessed treatment response to 4-aminopyridine (4-AP). ***Results:*** Frequency of *FGF14* GAA expansions was 48% (82/170) in the entire idiopathic DBN cohort, and 7% (1/14) in the subgroup with clinically isolated DBN, 61% (42/69) with cerebellar ocular motor dysfunction, 72% (23/32) with cerebellar ataxia (CA), 43% (6/14) with bilateral vestibulopathy (BVP), 33% (8/24) with neuropathy, and 12% (2/17) with CA, BVP, and neuropathy. Intermediate alleles between 200 and 249 repeats in 20/170 (12%) patients were enriched compared to control populations (~1-2%). The median age at onset was 67 years (range, 30-81) and was inversely correlated to the size of the expansion (Pearson r, -0.30; p=0.01). In patients with CA, oscillopsia developed on average less than a year after onset of gait ataxia, although it predated it by up to 8 years in some cases. Family history of DBN (38% vs 19%, p=0.01), impaired visual fixation suppression of the vestibulo-ocular reflex (86% vs 62%, p=0.002), and response to 4-AP treatment (81% vs 43%, p=0.009) were significantly more frequent in patients carrying an *FGF14* expansion. A meaningful benefit was reported by 79% of patients who had an objective response to 4-AP. The median FARS disability score at last examination (mean disease duration of 6.9 +/- 5.34 years) was 3 (range, 1.5-5), indicating the need to use a cane for walking. Functional impairment increased slowly as shown by the need for bilateral walking aid after 10 years in 9% of patients and an average increase of 0.09 FARS point/year of disease evolution. ***Conclusion:*** This study shows that *FGF14* GAA expansions are a common cause of DBN syndromes, especially when associated with cerebellar ocular motor dysfunction and CA, and defines a group of patients that are highly responsive to 4-AP treatment.

Title: Integrating polygenic scores and long-read genome sequencing to identify clinically-relevant structural variants in rare diseases

Authors:

C. LeMaster[1], C. Schwendinger-Schreck[1], B. Ge[2], W. Cheung[3], J. Johnston[3], T. Pastinen[3], C. Smail[3]; [1]Children's Mercy Kansas City, Kansas City, MO, [2]McGill Univ., Montreal, QC, Canada, [3]Children's Mercy Kansas City, Kansas City, MO

Abstract:

Recent studies have revealed the pervasive landscape of rare structural variants (rSVs) present in human genomes. rSVs can have extreme effects on proximal genes and, in a rare disease context, have been implicated in unexplained patient cases where no diagnostic single nucleotide variant (SNV) was found. Approaches for integrating rSVs to date have focused on targeted approaches in known Mendelian rare disease genes. This approach is intractable for rare diseases with many causal loci or patients with complex, multi-phenotype syndromes. We hypothesized that integrating trait-relevant polygenic scores (PGS) would provide both a substantial reduction in the number of candidate rare disease genes in which to assess rSV effects, as well as prioritization of these genes by leveraging a novel scoring metric using PGS gene weights. Among a subset of patients enrolled in the Genomic Answers for Kids (GA4K) study (N=497), we identified rSVs from PacBio HiFi long-read sequencing intersecting genes included in trait-relevant PGS across rare disease cohorts defined by Human Phenotype Ontology (HPO) terms; namely: autism (N=54); seizure (N=86); short stature (N=56); generalized hypotonia (N=62); and global developmental delay (N=138). Illustrating our approach in autism, we observed a 0.5 standard deviation (SD) increase in PGS liability for cases versus controls (P = 8x10-04), recapitulating previously reported associations in other autism cohorts. rSVs were stratified by a summarized PGS gene weight score, ranking relative disease impacts using PGS variants proximal to a given gene. We identified 252 deletions and 151 duplications overlapping genes in the top quantile of PGS gene effect. rSV length was on average longer in autism cases than controls. For individuals with &gt=1 deletion intersecting genes in the top PGS quantile, case PGS liability was on average 0.64 SD greater than controls (P = 8x10-04). This effect was not observed for deletions intersecting genes in the bottom PGS quantile (P = 0.17). We will further present an approach for leveraging phenotype-specific structural variant reference pangenome graphs to define rare/depleted regions for use in variant filtering strategies. Overall, our study provides a framework for the comprehensive integration and assessment of long-read rSVs, enabling expanded genome assessment and the potential for improved diagnosis of rare diseases.

Title: Systematic identification of recessive developmental disorders caused by noncoding variants in trans with coding variants in 4,073 trios.

Authors:

N. Whiffin[1], J. Lord[2], C. J. Oquendo[2], A. Martin-Geary[1], A. J. Blakes[3], E. Arciero[4], Genomics England Research Consortium, D. Baralle[5], H. C. Martin[4]; [1]Univ. of Oxford, Oxford, United Kingdom, [2]Univ. of Southampton, Southampton, United Kingdom, [3]Univ. of Manchester, Manchester, United Kingdom, [4]Wellcome Sanger Inst., Hinxton, United Kingdom, [5]Univ. of Southampton, Faculty of Med., Southampton, United Kingdom

Abstract:

In large exome and genome sequencing studies, 60-70% of individuals with developmental disorders (DDs) typically fail to get a genetic diagnosis. The contribution of noncoding variants to these undiagnosed cases remains to be fully explored, largely due to the challenge of the vast search space, and our relative lack of knowledge of the 'grammar' of noncoding regions, in contrast to protein-coding regions.

We systematically explored one mechanism by which noncoding variants may contribute to Mendelian disorders: by acting as a 'second hit' *in trans* with a coding variant in a recessive gene. Whilst it is common to find a single protein-altering variant in a recessive gene in individuals with DD, the prevalence of pathogenic noncoding second hits is unknown.

From 4,073 genetically undiagnosed rare disease trio probands in the 100,000 Genomes project, we identified one or more rare heterozygous predicted loss-of-function (pLoF) or ClinVar pathogenic variant in one of 793 recessive DD genes in 2,430 probands (59.7%), for a total of 3,761 unique proband-variant pairs. For each implicated gene, we defined a noncoding search space comprising the introns, the 5' and 3' untranslated regions (UTRs), an upstream promoter region, and accessible regions in fetal brain identified by sci-ATACseq. For 1,366 of the 3,761 proband-variant pairs (36.3%) we identified at least one rare noncoding variant within these regions *in trans*. Of these, 52 remained after stringent region-specific filtering. On review, 4 were deemed a probable phenotypic fit (and 13 more 'possible'). For the probable matches, we contacted the recruiting clinician to confirm the clinical fit, and where possible obtain RNA for functional validation.

In trans with a LoF variant, we identified intronic variants in *LAMA2*, *IGHMBP2*, and *NPHP3* in probands with congenital myopathy, Charcot-Marie-Tooth disease and proteinuric renal disease, respectively, and a promoter variant 182bps upstream of the transcription start site of *GAA* in a proband with limb girdle muscular dystrophy (LGMD). We subsequently identified two further LGMD probands with both the *GAA* promoter variant and a rare pLoF variant. All also carried a known pathogenic 5'UTR splice-site variant *in cis* with the promoter variant, that RNA-seq confirmed as a more likely causative second hit. This variant was initially excluded due to a high gnomAD frequency.

In summary, we developed a pipeline to systematically identify and annotate noncoding 'second hit' variants and used this to uncover new diagnoses. Whilst we acknowledge multiple reasons why our approach will underestimate the true prevalence, we conclude that this mechanism is likely a rare cause of DDs.

Title: Clustering of de novo noncoding variants in the genome in autism

Authors:

J. Ng, H. B. Heins, J. G. Manuel, T. N. Turner; Washington Univ. Sch. of Med., Saint Louis, MO

Abstract:

Noncoding *de novo* variants (DNVs) in promoters and enhancers have been implicated in autism. In the human genome, 98.5% of the sequence is noncoding and the identification of signals in promoters and enhancers is intriguing. However, the identification of enrichment in these regions is annotation-dependent for known noncoding annotations. Enhancers can function in a spatiotemporal manner, and it is likely that not all enhancer sequences are known at this time due to the vast diversity of cells and regions in the human brain. There are also other types of noncoding regions (e.g., silencers). We developed an annotation-free computational tool (GRUMP, GeneRalized clUstering by Mutation Position) for detecting enrichment of clustered variation in regions of the genome. This approach enables a variant-first based strategy for identifying enrichment in the noncoding genome. To detect DNVs quickly for the downstream GRUMP analysis, we recently developed a GPU-accelerated DNV caller (HAT). This DNV caller works on short-read whole-genome sequencing (WGS), short-read whole-exome sequencing (WES), and long-read WGS datasets. For our discovery cohort, we called DNVs from WGS data using HAT in 4,216 parent-child trios. We than applied GRUMP in a case-control implementation to examine clustering of DNVs in topologically associating domains (TADs) or genes, respectively. We identified several genes and TADs with genome-wide significant differential clustering in cases versus controls including those relevant to known autism genes (e.g., *KCNH1*, *PPP2R1A, SMARCA2*). DNV calling in the replication cohort consisting of 5,713 parent-child trios is currently underway. We are also in the process of designing a massively parallel reporter assay (MPRA) to test variation detected in the genome-wide significant GRUMP results from the discovery cohort. The MPRA will be tested in multiple neuronal cell lines that we are deeply characterizing by karyotype, Illumina whole-genome sequencing (WGS), PacBio WGS, Hi-C, ATAC-seq, Illumina RNA-seq and PacBio Iso-Seq. Overall, our work provides new insights into noncoding variation in autism.

Session 136: Manipulating metabolism: Therapeutics strategies for inborn errors of metabolism

Location: Conv Ctr/Room 145A/Level 1

Session Time: Sunday, November 5, 2023, 10:00 am - 11:00 am

Title: Development of splice-switching small molecule RECTAS2.0 to treat cardiac Fabry disease caused by GLA c.639+919G>A Taiwanese variant.

Authors:

T. Awaya[1], M. Ajiro[1,2], K. Iida[3,4], M. Hagiwara[1]; [1]Kyoto Univ., Graduate Sch. of Med., Kyoto, Japan, [2]Natl. Cancar Ctr. Res. Inst., Tokyo, Japan, [3]Kyoto Univ., Med. Res. Support Ctr., Kyoto, Japan, [4]Kindai Univ., Faculty of Sci. and Engineering, Osaka, Japan

Abstract:

Fabry disease (FD) is an X-linked lysosomal storage disorder resulting from decreased activity of α-galactosidase A (α-Gal A). The enzymatic defect leads to lysosomal accumulation of globotriaosylceramide (Gb3) and multi-systemic symptoms. The c.604-801G>A (also known as c.639+919G>A or IVS4+919G>A) variant of α-Gal A gene (*GLA*) is known to create a 44-nucleotide pseudoexon within the 4th intron, serving as a causative mutation of late-onset cardiac FD, which is highly prevalent in Taiwanese population (1 in 1,000 of male newborns, 1:500 in female newborns). We have been investigating splice-switching small molecules that could potentially be used to treat genetic disorders, especially focusing on SRSF phosphorylation. We found that our splicing modulator RECTAS successfully suppressed the pseudoexon inclusion by promoting SRSF6 binding to downstream exon 5 of *GLA*, recovered α-Gal A activity in patient B-lymphoblastoid cells, and ameliorated abnormal Gb3 accumulation in patient iPSC-derived cardiomyocytes. To improve and optimize the efficacy and pharmacokinetics of the original RECTAS, we obtained 431 RECTAS derivatives and successfully discovered RECTAS2.0, which is orally available and exhibits 10-fold higher potency with sufficient heart distribution. Therapeutic efficacy of RECTAS2.0 was further verified in patient iPSC-derived cardiomyocytes and transgenic mice harboring the c.604-801G>A human *GLA*. Safety profile of RECTAS2.0 was confirmed in pre-clinical studies. RECTAS2.0 has the possibility to pave the way for splice-switching therapy to treat cardiac FD with the c.604-801G>A variant.

Title: Adjunct treatment with glycogen synthase (GYS1) antisense oligonucleotides and enzyme replacement therapy (ERT) reduces glycogen in the Pompe disease mouse model.

Authors:

A. Martin Rios[1], L. Weiss[1], M. Carrer[2], A. Shmara[1], V. Boock[1], M. Ibrahim[1], L. Hettrick[2], A. Watt[2], P. Jafar-Nejad[2], V. Kimonis[1]; [1]Univ. of California, Irvine, Irvine, CA, [2]Ionis Pharmaceuticals, Inc., Carlsbad, CA

Abstract:

Pompe disease is a progressive myopathy resulting from the deficiency of acid alfa-glucosidase (GAA). Enzyme replacement therapy (ERT) with recombinant human (rh) GAA works well in alleviating the cardiomyopathy; however, many patients continue to have progressive muscle weakness from muscle glycogen accumulation produced by muscle glycogen synthase (GYS1). Previous studies have provided proof of principle that knockdown of GYS1 mRNA by antisense oligonucleotide reduced glycogen. To impart specificity for the muscle variant of the enzyme, antisense oligonucleotides (ASOs) targeting mouse Gys1 designed and previously screened in vitro and in vivo by Ionis Pharmaceuticals were assessed. In this report, we focused on the three ASOs to identify the most efficacious *Gys1* ASO in Gaa-/- mice. The results from treatment with the three *Gys1* ASOs we identified *Gys1* ASO#3 as the most effective with 84% and 98% downregulation of Gys1 mRNA and GYS1 protein respectively, and 47% efficiency in clearing glycogen accumulation in quadriceps muscle of 1-month-old Gaa-/- Pompe mice. Combination therapy with *Gys1* ASO#3 and ERT further reduced glycogen accumulation and effectively alleviated the massive autophagic buildup in Gaa-/- mice quadriceps to wild type mice level. The reversal of lysosomal and autophagic pathologies led to improved muscle function in treated Gaa-/- mice. Studies are currently ongoing with early treatment with ASOs in 4-day-old pups as potential monotherapy. These results will provide proof of principle that ASOs inhibiting GYS1 can reduce glycogen in Pompe disease.

Title: Correction of metabolic abnormalities in murine glycogen storage disease type Ia using a CRISPR/Cas9-based double-stranded DNA oligonucleotide insertional strategy

Authors:

A. Samanta[1], G. Nelson[1], I. Arnaoutova[1], B. Mansfield[1], C. Hart[2], T. Carlo[2], J. Chou[1]; [1]NICHD/NIH, Bethesda, MD, [2]Prime Med. Inc, Cambridge, MA

Abstract:

Type Ia glycogen storage disease (GSD-Ia), characterized by impaired blood glucose homeostasis, is caused by a deficiency in glucose-6-phosphatase-α (G6Pase-α or G6PC) that catalyzes the hydrolysis of G6P to glucose in the terminal step of gluconeogenesis and glycogenolysis. We have developed a rAAV-G6PC-mediated gene augmentation therapy for GSD-Ia, which, licensed to Ultragenyx Pharmaceutical, is now in a phase III clinical trial (NCT05139316). Currently there is insufficient clinical data to understand if the episomal rAAV-*G6PC*-mediated transgene expression can be maintained at a therapeutic level for multiple decades in the human liver. We therefore explored genome editing, a more permanent alternative genetic technologies for GSD-Ia therapy. We have previously generated a *G6pc*-R83C mouse strain carrying the prevalent pathogenic G6PC-p.R83C variant and showed that the mice exhibit the pathophysiology mimicking human GSD-Ia. Using the *G6pc*-R83C mice, we now explored a CRISPR/Cas9-based double-strand DNA oligonucleotide (dsODN) insertional strategy to correct the p.R83C variant. This strategy uses the non-homologous end joining repair mechanism, which is effective in both dividing and non-dividing cells and the major pathway for DNA double-strand break repair. To reduce any risk associated with AAV delivery and the resulting extended expression of the editing reagents, we delivered the reagents transiently, formulated as lipid nanoparticles (LNP).The LNP formulated with the editing reagents consisting of Cas9 mRNA, guide RNA, and dsODN (LNP-dsODN) was infused into the *G6pc*-R83C mice either with one dose at age 4 weeks or two doses at age 2 and 4 weeks, and phenotypic correction of the treated mice was evaluated at age 8 and 16 weeks, respectively. We have previously shown that restoring hepatic G6Pase-α activity to ≥3% of normal hepatic G6Pase-α activity is sufficient to normalize metabolism. We now show that the dsODN insertional strategy restored hepatic G6Pase-α activity to ~4% of normal hepatic activity in the *G6pc*-R83C mice and increased their 8-week survival rate from 0% for untreated mice to 100% for the treated mice. Furthermore, the LNP-dsODN-treated *G6pc*-R83C mice maintained glucose homeostasis and displayed normalized metabolic phenotype similar to that observed in the preclinical studies of rAAV-G6PC-mediated gene augmentation therapy, that has entered phase III clinical trials. This dsODN insertional gene editing strategy may offer the basis for a therapeutic approach with an earlier clinical intervention than gene augmentation, with the additional benefit of a potentially permanent correction of the GSD-Ia phenotype.

Title: RGX-111: An Investigational Gene Therapy For The Treatment of Severe Mucopolysaccharidosis Type I (MPS I): Interim Analysis of Data From The First In Human Study

Authors:

L. Pisani[1], R. Wang[2], C. Ficicioglu[3], R. Giugliani[4], M. Fiscella[1], L. Yang[1], H. Wang[1], M. Gilmor[1], Y. Cho[1], C. mulatya[1], D. Phillips[1], P. Falabella[1], J. burke[1]; [1]RegenXBio, Rockville, MD, [2]CHOC Children's Specialists, Orange, CA, [3]Children S Hosp. of Philadelphia, Philadelphia, PA, [4]HCPA/UFRGS, Porto Alegre, Brazil

Abstract:

Background: MPS I is a rare, autosomal recessive disease caused by deficiency of alpha-L-iduronidase (IDUA), an enzyme required for the breakdown of lysosomal glycosaminoglycans (GAGs). RGX-111, a recombinant adeno-associated virus serotype 9 capsid containing a human IDUA expression cassette (AAV9.CB7.hIDUA) administered to the CNS, may provide a permanent source of secreted IDUA that corrects neurologic and systemic disease manifestations. Methods: This phase I/II, first-in-human, 104-week, open-label trial enrolled participants $\geq$ 4 months of age with severe MPS I or documented CNS involvement (NCT03580083) who received one image-guided RGX-111 injection to the CNS. Assessments include safety and tolerability; CSF, plasma, and urine biomarkers; cognition, language, and motor neurodevelopmental scales; and imaging. Participants are encouraged to enroll in a long-term follow-up study. Results: As of January 17, 2023, 8 participants were enrolled in 2 dose cohorts ($1.0x10^{10}$ and $5.0x10^{10}$ genome copies/gram brain mass). RGX-111 was reported to be well-tolerated with no drug-related serious adverse events. Post-administration follow-up ranged up to 103 weeks. CSF GAGs were decreased through last time point available. Interim neurodevelopmental assessments demonstrated continued neurodevelopmental skill acquisition on age and developmentally appropriate scales. Emerging evidence of systemic enzyme expression and biomarker activity was present after CNS RGX-111 administration. Additionally, follow-up will be reported for a severe MPS I child treated at age 21 months utilizing a single-patient, investigator-initiated Investigational New Drug application. Conclusions: RGX-111 has the potential to provide sustained CNS clinical outcomes and additional systemic effects in MPS I patients.

Session 137: Rare germline variants and cancer susceptibility

Location: Conv Ctr/Room 202A/Level 2

Session Time: Sunday, November 5, 2023, 10:00 am - 11:00 am

Title: Leveraging ~937K exomes to estimate cancer risk conferred by rare deleterious germline variants in hereditary cancer genes

Authors:

K. Burch, M. D. Kessler, J. Tang, A. Marcketta, DiscovEHR, Mayo Clinic Project Generation, Penn Medicine Biobank, Regeneron Genetics Center, G. Abecasis, E. Jorgenson; Regeneron Genetics Ctr., Tarrytown, NY

Abstract:

Identifying genes and variants associated with multiple cancers has important implications for prioritizing potential drug targets, repurposing existing drugs, and improving cancer screening and genetic testing. More than 100 genes containing highly penetrant cancer risk-increasing germline variants have been found through linkage analyses and sequencing in familial clusters, but efforts to systematically test for pleiotropy and estimate risk conferred upon carriers have been limited by sample size. Here, we meta-analyze the exomes of ~937K individuals (African American, East Asian, European, and South Asian ancestries) from 8 studies with electronic health record data to characterize the pleiotropic landscape of 121 hereditary cancer risk genes. We first generate gene burden "masks" for each of the 121 genes, combining multiple allele frequency (AF) cutoffs (<1%, <0.1%, <0.01%, singletons) with predictions of mutational impact (for example, missense, loss-of-function), to identify variants within each gene that are most likely to be causally associated with cancer risk. We then use these masks to estimate cancer type-specific risk (for 23 solid cancer/tumor types) conferred upon carriers of different sets of these variants. We find support for associations with 2+ cancer types for 47 genes, of which 27 remain significant after multiple testing correction (FDR<0.05). *ATM* is the most pleiotropic (16 associated cancer types), followed by *BRCA1* (12); *BRCA2*, *MSH6*, and *CHEK2* (10 each); *MSH2* (8); *TP53* and *PTEN* (7 each); *APC* (6); and *MLH1* (5). We identify multiple statistically significant associations that have not been previously reported or replicated in any population-based studies, including *BRCA1* (singletons) with lung cancer (OR 95%-CI=[3.3,10.8], $P=3x10^{-9}$), colon cancer (OR=[1.9,4.0], $P=1x10^{-7}$), and non-melanoma skin cancer (NMSC, OR=[1.7,3.3], $P=1x10^{-7}$); *BRCA2* (AF<0.01%) with NMSC (OR=[1.4,1.9], $P=8x10^{-9}$); and *ATM* (singletons) with NMSC (OR=[1.3,2.0], $P=2x10^{-6}$). Further, the magnitude of cancer-type specific risk conferred varies widely across cancers for some genes (e.g., *BRCA1* carriers (AF<0.01%) have 27-49x higher risk for ovarian cancer versus 1.7-2.8x higher risk for NMSC) and less so for other genes (e.g., *CHEK2* carriers (AF<0.01%) have 1.3-1.8x higher risk for colon cancer and up to 2.1x-3.8x higher risk for thyroid cancer). By leveraging computational predictions of variant deleteriousness to refine association signals in the world's largest repository of exome-sequencing data, we provide comprehensive estimates of risk that can inform cancer screening and decision-support efforts.

Title: A pathway-based, ultra-rare variant association study identifies phenotype specific germline risk of breast cancer.

Authors:

J. Asmussen[1], K. Lee[2], P. Katsonis[1], L. Donehower[1], Y. Li[1], O. Lichtarge[1]; [1]Baylor Coll. of Med., Houston, TX, [2]Baylor Coll. Med., Houston, TX

Abstract:

Breast cancer (BRCA) is a complex disease with well-characterized genetic and lifestyle risk factors. However, current polygenic risk scores are only clinically actionable at the extremes, leaving most individuals without clear guidance. To address this, we sought to identify germline genes and pathways associated with phenotypic subsets of BRCA. Our method, EA-Pathways, uses evolutionary information to find pathways in which ultra-rare germline variants show altered energy profiles in fitness landscapes relative to simulated controls, allowing for the independent identification of predisposition and protective genes and pathways in affected and disease-free individuals, respectively.

Here, we used 16,000 BRCA women and 16,000 controls with whole exome sequencing from the UK Biobank. Lifestyle and female specific factors (parity, age at menarche, height, body mass index (BMI)) were used to define six phenotypes across all 32,000 women with unsupervised k-means clustering. Clusters effectively separated women based on combinations of extreme phenotypes (ANOVA p-val < 1e-16 for each phenotype). As expected, five of the six phenotypes were significantly biased to cases or controls based on known risk factor combinations (Fisher's Exact p-val: 9.8e-3, 1.0e-13, 5.8e-20, 0.2, 3.1e-13, 4.5e-5).

EA-Pathways then identified pathways with abnormal ultra-rare variant mutational energy profiles in the germlines of each phenotype. These genes overlapped and were connected to known familial BRCA risk genes (Hypergeometric p-val: 4.3e-9, 5.5e-7, 8.3e-11, 1.3e-11, 2.1e-14, 2.6e-8; First neighbor connectivity z-score: 14.5, 9.8, 10.9, 11.1, 10.4, 15.8). Odds ratio (OR) analyses revealed that each phenotype was associated with a different profile of risk pathways. Shared pathways included "Diseases of DNA repair" (OR: 1.30, 1.42; FDR: 0.01, 1e-3), "Homologous DNA Pairing and Strand Exchange" (OR: 1.39, 1.43; FDR: 5e-3, 0.03), and "Impaired BRCA2 binding to PALB2" (OR: 1.63, 1.41, 1.54, 1.40; FDR: 8e-3, 5e-3, 0.02, 0.01). Unique pathways included "Methylation" (OR: 0.41; FDR: 0.06) and "Eicosanoids" (OR: 0.65; FDR: 0.07), both of which were associated with disease protection. Notably, the phenotype with the highest BMI lacked DNA repair OR associations, suggesting that the genetic risk associated with elevated BMI differs from other phenotypes. Critically, these trends persist in an independent subset of held-out women.

These findings support EA-Pathways as a general, novel, ultra-rare germline variant pathway association method to identify risk genes and biology across BRCA phenotypes.

Title: Underlying germline genetic architecture of pediatric sarcomas: evaluating the role of common and rare variants in 4,160 patients

Authors:

L. Mirabello[1], L. E. Egolf[1,2], B. Zhu[1], D. Gianferante[1], K. Wang[1,2], S. Li[1,2], M. J. Machiela[1], L. G. Spector[3], J. D. Schiffman[4,5], A. Sabo[6], A. Renwick[6], B. Martin-Giacalone[7], M. E. Scheurer[6,8], S. Plon[6,8], D. Hawkins[9], R. Venkatramani[6,8], D. Stewart[1], L. M. Morton[1], M. M. Hudson[10], G. T. Armstrong[11], S. Bhatia[12], M. Dean[1], K. A. Janeway[13], A. Patiño-Garcia[14], F. Lecanda[15], M. Serra[16,17], C. Hattinger[16,17], K. Scotlandi[16], A. M. Flanagan[18,19], F. Amary[19], I. L. Andrulis[20], J. S. Wunder[20], M. L. Ballinger[21,22], D. M. Thomas[21,22], O. Delattre[23], A. K. Hubbard[1], J. Liu[1,2], W. Luo[1,2], B. D. Hicks[1,2], M. Yeager[1,2,24], N. D. Freedman[1], W-Y. Huang[1], M. T. Landi[1], A. Lori[25], R. Diver[25], S. A. Savage[1], S. J. Chanock[1], P. J. Lupo[6,8]; [1]Div. of Cancer Epidemiology and Genetics, NCI, NIH, Bethesda, MD, [2]Cancer Genomics Res. Lab., Frederick Natl. Lab. for Cancer Res., Frederick, MD, [3]Dept. of Pediatrics, Univ. of Minnesota, Minneapolis, MN, [4]Dept. of Pediatrics (Hematology/Oncology) and Oncological Sci., Univ. of Utah Sch. of Med., Salt Lake City, UT, [5]Huntsman Cancer Inst., Univ. of Utah Sch. of Med., Salt Lake City, UT, [6]Human Genome Sequencing Ctr., Baylor Coll. of Med., Houston, TX, [7]Div. of Publ. Hlth.Sci., Dept. of Surgery, Washington Univ. in St. Louis Sch. of Med., Saint Louis, MO, [8]Texas Children's Cancer Ctr., Houston, TX, [9]Dept. of Pediatrics, Seattle Children's Hosp., Seattle, WA, [10]Dept. of Oncology, St. Jude Children's Res. Hosp., Memphis, TN, [11]Dept. of Epidemiology and Cancer Control, St. Jude Children's Res. Hosp., Memphis, TN, [12]Inst. for Cancer Outcomes and Survivorship, Univ. of Alabama at Birmingham, Birmingham, AL, [13]Dana-Farber/Boston Children's Cancer and Blood Disorders Ctr., Harvard Med. Sch., Boston, MA, [14]Dept. of Pediatrics/Med. Genomics Unit and Program in Solid Tumors, Cima-Univ. de Navarra, Cancer Ctr. Clínica Univ. de Navarra (CCUN), Pamplona, Spain, [15]Ctr. for Applied Med. Res. (CIMA)-Univ. of Navarra, IdiSNA, and CIBERONC, Pamplona, Spain, [16]Lab. of Experimental Oncology, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy, [17]IRCCS Istituto Ortopedico Rizzoli, Osteoncology, Bone and Soft Tissue Sarcomas and Innovative Therapies, Pharmacogenomics and Pharmacogenetics Res. Unit, Bologna, Italy, [18]Univ. Coll. London Cancer Inst., London, United Kingdom, [19]Royal Natl. Orthopaedic Hosp. NHS Trust, Stanmore, United Kingdom, [20]Lunenfeld-Tanenbaum Res. Inst., Sinai Hlth.System, Univ. of Toronto, Toronto, ON, Canada, [21]Garvan Inst. of Med. Res., Darlinghurst, Australia, [22]St Vincent's Clinical Sch., Faculty of Med., Univ. of New South Wales, Sydney, Australia, [23]PSL Université, and SIREDO Oncology Ctr., Inst. Curie, Paris, France, [24]Dept. of Biology, Hood Coll., Frederick, MD, [25]Dept. of Population Sci., American Cancer Society, Kennesaw, GA

Abstract:

Some evidence suggests that pediatric sarcomas have both shared and distinct genetic profiles; however, large-scale efforts to characterize germline genetic susceptibility across these malignancies are limited by their rarity. We evaluated the role of common and rare variants in the genetic etiology of the more frequent pediatric sarcomas: osteosarcoma (OS); Ewing sarcoma (ES); and rhabdomyosarcoma (RMS), subcategorized into embryonal (ERMS) and alveolar (ARMS). **METHODS:** We evaluated 4,161 European-ancestry cases with genotype data (1,843 OS, 733 ES, 1,585 RMS, and ~61,000 in-house cancer-free adult controls) and 2,474 cases with exome or genome sequencing (1,002 OS, 579 ES, 893 RMS, and 1,057 controls; jointly called with the same QC). Analyses included: 1) estimating disease heritability for both common SNPs (MAF>3%; genome-wide) and rare loss-of-function (LOF) variants (MAF<1%; exome-wide); and 2) determining the frequency of rare predicted pathogenic (ACMG-AMP) variants in cancer susceptibility genes (CSG). **RESULTS:** For OS, common variants explained 2.7% (SE 1.1%) of disease heritability, while rare LOF variants explained 12.7% (SE 1.5%; compared to 0.4% for synonymous variants). A similar pattern was observed for RMS, particularly ERMS, where rare LOF variants explained a greater proportion of disease heritability (RMS 8.6%, SE 1.2%; ERMS 12.6%, SE 1.7%; ARMS n/a due to small sample size) compared to common variants (0.9%, SE 1.6%). Conversely, common variants explained a greater proportion of ES heritability (5.4%, SE 0.5%). For 127 autosomal dominant (AD) CSG, 21.1% of OS, 18.1% of ERMS, 12.6% of ARMS and 10.5% of ES cases had a predicted pathogenic rare variant. Similarly, for the 60 moderate-to-high penetrant AD CSG, OS and ERMS had significantly more pathogenic variants than controls (P_{exact}<0.01); whereas ARMS and ES had similar or fewer than controls. ERMS was enriched for *NF1* (3.2%), *TP53* (2.7%), *HRAS* (1.1%), and *DICER1* (0.8%) relative to controls (P_{exact}<0.05), and OS was enriched for *TP53* (2.6%) and *CDKN2A* (0.5%). ARMS and ES had no AD genes significantly enriched for pathogenic variants. For 49 autosomal recessive CSG, pathogenic variant carrier frequencies were similar among these sarcomas (10-13%); *FANCD2* was significantly enriched across all cases (0.74%) compared to controls (0.09%). **CONCLUSION:** In the largest set of pediatric sarcoma cases assembled to date, genetic susceptibility was largely driven by rare pathogenic AD gene variants in tumor types not characterized by canonical somatic fusions (OS and ERMS). In contrast, fusion-driven tumor types (ES and suggestively for ARMS) were driven more by common variants.

Title: Population-level transcriptomics links rare alternative polyadenylation variations to complex disease

Authors:

L. Li, X. Zou, R. Ding, Q. Yangmei, T. Ni; Shenzhen Bay Lab., Shenzhen, China

Abstract:

Rare variants are abundant in the human genome and frequently contribute to rare and common human diseases. However, identifying and functionally interpreting of these rare variants remains a significant challenge. Alternative polyadenylation (APA) is a key post-transcriptional modification for most human genes that substantially impacts upon cell behavior. In our previous study, we described the atlas of genetic variants associated with APA (3′aQTL), which are colocalized with approximately 16.1% of 15 human diseases and traits loci. Here, we describe the pan-tissue atlas of alternative polyadenylation (APA) outliers (aOutliers) in 830 genes across 49 human tissues from 17,382 samples. A distinct class of rare variants is significantly enriched among aoutliers. Mechanically, these aOutlier-associated RVs frequently alter poly(A) signal and splicing sites. Additionally, they also have a synergistic effect with common APA variants, such as in the DDX18 gene. We further developed a Bayesian-based rare APA variants prediction model and revealed many new functional rare variants with large impacts on complex traits and diseases. Our study highlights the utility of using APA in discovering new functional rare variations in human diseases and provides a new framework for interpreting individual genomes.