American Society of Human Genetics 65th Annual Meeting
October 6–10, 2015 Baltimore, MD

PLATFORM ABSTRACTS

The following is a suggested style for citing ASHG 2015 Meeting abstracts-Example Only:

Simpson R.S., Barnes P., Disruption of the microRNA pathway; (Abstract/Program #XX). Presented at the 65th Annual Meeting of The American Society of Human Genetics, Date, Location (e.g., October 7, 2015 in Baltimore, MD).

The submitting author has copyright to the individual abstract, so permission to use material derived from the abstract (other than citation) must be obtained from that author.

1

Massively parallel experimental analysis of missense mutations in *BRCA1* for interpreting clinical variants of uncertain significance.

L.M. Starita[1], M. Islam[2], J. Gullingsrud[1], S. Fields[1], J.D. Parvin[2,3], J. Shendure[1]. 1) Department of Genome Sciences, University of Washington, Seattle, WA; 2) Department of Biomedical Informatics, The Ohio State University, Columbus, OH; 3) The Ohio State University Comprehensive Cancer Center, The Ohio State University, Columbus, OH.

Women inheriting a pathogenic mutation in the *BRCA* genes are at increased risk for developing breast and ovarian cancer. However, a substantial proportion of women who undergo *BRCA1/2* testing learn that they carry a variant of uncertain significance (VUS), whose consequences for BRCA1/2 activity and therefore cancer risk are unknown. Individuals harboring a VUS must make decisions about cancer prevention without clear information. A path towards addressing this failure in cancer-risk assessment and prevention is to experimentally measure the functional consequences of all possible mutations in *BRCA1/2*, and to make these measurements publicly available as a resource for guiding variant interpretation. To this end, we are applying “deep mutational scanning” (Fowler et al. Nature Methods 2010), a method in which the effects of thousands of mutations in a single gene can be concurrently measured, to BRCA1. For example, we comprehensively evaluated the effects of >1,300 amino acid substitutions on the biochemical functions of the RING domain of BRCA1 (Starita et al. Genetics 2015) as well as the effects of nucleotide substitutions in exon 18 of *BRCA1* on mRNA splicing (Findlay et al. Nature 2014). However, these studies were limited in that they incompletely assessed the function of BRCA1 that is most physiologically relevant to its role in cancer risk. BRCA1 is required for homology-directed repair (HDR) of double strand DNA breaks and this function is required for tumor suppression. We have adapted a cellular assay to test the HDR function of the full-length BRCA1 protein for deep mutational scanning. Preliminary results show that we can distinguish HDR-functional from nonfunctional BRCA1 variants a multiplexed format. We have now integrated thousands of BRCA1 missense variants into an HDR reporter cell line and are presently performing a large-scale experiment that will quantify the impact of each of these mutations on HDR activity. Based on the complexity of this library, we anticipate that this experiment will elucidate the functional consequences of >3,000 BRCA1 protein variants.The sum of our results-to-date show that predictions based on massively parallel experimental analysis markedly outperform commonly used computational tools in predicting BRCA1 function. As such, we anticipate that these measurements will facilitate the prospective interpretation of BRCA1 mutations when they are observed for the first time in a clinical setting.

2

Matrix metallopeptidase 21 (*MMP21*) is mutated in human heterotaxy and is an essential determinant of vertebrate left-right asymmetry.

A. Guimier[1, 2], G.C. Gabriel[3], F. Bajolle[4], M. Tsang[3], H. Liu[5], A. Noll[6, 7], L.D. Smith[6, 7], S. Lyonnet[1, 2, 9], L. de Pontual[1, 2], S.A. Murray[8], D. Bonnet[2, 4], S.F. Kingsmore[6, 7], J. Amiel[1, 2, 9], P. Bouvagnet[5], C.W. Lo[3], C.T. Gordon[1, 2]. 1) Laboratory of embryology and genetics of congenital malformations, INSERM UMR1163, Institut Imagine, Paris, France; 2) Paris Descartes-Sorbonne Paris Cité University, Institut Imagine, Paris, France; 3) Department of developmental biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; 4) Unité Médico-Chirurgicale de Cardiologie Congénitale et Pédiatrique, Centre de référence Malformations Cardiaques Congénitales Complexes - M3C, Hôpital Necker-Enfants Malades, APHP, Paris, France; 5) Laboratoire de cardiogénétique - Hospices Civils de Lyon, Bron, France; EA 4173 Université Lyon 1 et Hôpital Nord Ouest, Lyon, France; 6) Center for Pediatric Genomic Medicine, Departments of Pediatrics and Pathology, Children’s Mercy - Kansas City, Kansas City, MO, USA; 7) University of Missouri - Kansas City School of Medicine, Kansas City, MO, USA; 8) The Jackson Laboratory, Bar Harbor, Maine, USA; 9) Service de Génétique, Hôpital Necker-Enfants Malades, AP-HP, Paris, France.

Heterotaxy (HT) results from a failure to establish normal left-right asymmetry (LRA) early in embryonic development and comprises visceral malformations among which congenital heart defects (CHDs) are the major cause of morbidity and mortality. Mutations in several genes controlling early left–right patterning have been implicated in HT but account for a minority of cases.We performed whole exome or genome sequencing in 2 families with recurrence of complex CHDs associated with laterality defects of abdominal organs. We identified compound heterozygous mutations (stop and missense or frameshift and exonic deletion) in matrix metallopeptidase 21 (*MMP21*) in both families. MMP family members are involved in extra-cellular matrix turnover. Interestingly, mice homozygous for ENU-induced missense mutations in *Mmp21* exhibit CHDs and HT. We then performed next generation sequencing of *MMP21* in a cohort of 264 index cases comprising a group of HT cases (n=154) and a group of cases with CHDs (such as tetralogy of Fallot or truncus arteriosus) but without HT (n=110). From this cohort we identified 7 other families with one or more affected siblings exhibiting biallelic variations in *MMP21*, including a homozygous missense affecting the start codon in one family and a homozygous frameshift in another. All these cases were found in the HT subgroup. Based on these findings, *MMP21* mutations account for 5.9% of non-syndromic HT cases. Also, we knocked down *mmp21* expression in zebrafish using a splice-blocking or a translation-blocking morpholino, which resulted in abnormal cardiac looping, a consequence of disrupted left-right patterning. Whole mount *in situ* hybridization for *mmp21* in zebrafish embryos revealed expression only in the region of Kupffer’s vesicle, a ciliated organ that generates LRA in fish. We then used CRISPR/Cas9 mediated genome editing in mouse zygotes to knock-in a missense mutation identified in one of the affected families, and phenotyping of mutation-positive embryos revealed CHDs and laterality phenotypes. Altogether our results indicate that *MMP21* is a novel disease-causing gene for heterotaxy in humans and that *MMP2*1 is an essential component of the pathway specifying LRA.

3

Meta-analysis of OncoArray, iCOGS and GWAS data for more than 220,000 women identifies more than 50 novel breast cancer susceptibility loci. *K. Michailidou[1], S. Lindstrom[2], J. Dennis[1], D.J. Hunter[2], Z. Wang[3], S. Chanock[3], J. Simard[4,5], P. Kraft[2], D.F. Easton[1,6] on behalf of BCAC, DRIVE and PERSPECTIVE.* 1) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; 2) Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, MA, USA; 3) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA; 4) Centre Hospitalier Universitaire de Québec Research Center, Quebec City, Quebec, Canada; 5) Laval University, Quebec City, Quebec, Canada; 6) Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom.

Genome-wide association studies (GWAS) have identified 94 loci associated with breast cancer susceptibility in the general population. If combined multiplicatively, these loci explain approximately 15% of the familial relative risk of the disease. To identify novel breast cancer susceptibility loci we conducted a GWAS involving 119,000 European Ancestry cases and 101,000 European Ancestry controls, including: 9 GWAS (11,000 breast cancer cases and 12,000 controls); 47,000 cases and 43,000 controls from 41 studies genotyped on a 200K SNP custom array (iCOGS); and 61,000 cases and 46,000 controls genotyped on the OncoArray, a 570K SNP custom array that included a 260K GWAS backbone (http://epi.grants.cancer.gov/oncoarray/), together with SNPs identified through previous GWAS, fine-mapping and sequencing studies in multiple cancer types. The GWAS, iCOGS and OncoArray samples were imputed using the October 2014 release of the 1000 genomes project data as reference. Association results on more than 15M SNPs were combined across platforms using inverse variance fixed effect meta-analysis. Of the 94 previously identified loci, 89 showed evidence for association in the Oncoarray for either overall, ER-positive or ER-negative breast cancer ($P<0.01$). After exclusion of the regions surrounding previously associated variants (± 500 kb of the top hit) we identified more than 50 novel independent variants associated with overall breast cancer at $P<5x10^{-8}$. The new loci combined explain a further ~5% of the familial relative risk of breast cancer. Seven additional loci were specifically associated with ER-negative disease at $P<5x10^{-8}$. Possible candidate genes within close proximity of the newly identified variants include *APOBEC3A/B*, involved in viral immunity and hypermutation, *FAM175A*, encoding *BRCA1* interacting protein *HIVEP3*, a transcription factor regulating kappaB mediated transcription and *MCM8*, involved in genome replication. The heritability due to SNPs in regulatory features was 6-10 fold enriched relative to the genome-wide average; these features included histone methylation marks, DNase I hypersensitive sites, transcription factor binding sites and enhancers. The heritability due to coding SNPs was three-fold enriched. These results provide further insight into the mechanisms driving breast carcinogenesis and will improve the utility of genetic risk scores for targeted prevention and screening.

4

Exome sequencing to identify new genes underlying early-onset breast cancer susceptibility. *D. Koboldt[1], K. Kanchi[1], J. Ivanovich[2], R. Fulton[1], I. Borecki[3], P. Goodfellow[4], R. Wilson[1], E. Mardis[1].* 1) The McDonnell Genome Institute, Washington University, Saint Louis, MO; 2) Department of Surgery, Washington University, Saint Louis, MO; 3) Department of Genetics, Washington University, Saint Louis, MO; 4) College of Medicine, Ohio State University, Columbus, OH.

In 2015, an estimated 230,000 American women will be diagnosed with breast cancer and 40,000 will die from it. Inherited genetic factors play a considerable role in this malignancy, particularly among women diagnosed at a young age. At least 20% of early-onset cases harbor mutations in the best-understood breast cancer susceptibility genes (*BRCA1*, *BRCA2*, and *TP53*). While others may have germline mutations in other established susceptibility genes (*CHEK2*, *PALB2*, *NBS1*, *RAD51*, *ATM*, *BRIP1*, and others), the genetic factors contributing to cancer risk for the majority of early-onset cases remain to be determined.We set out to identify additional early-onset breast cancer predisposition alleles by exome sequencing in an enriched cohort: 375 women diagnosed with breast cancer before the age of 40 who had a positive family history but were negative for BRCA1/2 mutations. As a population control, we obtained exome data for 557 women enrolled in the Women's Health Initiative of the NHLBI Exome Sequencing Project. We identified 1,251 genes enriched for rare deleterious variants in early-onset cases. For completeness, we expanded this set with another 237 genes implicated in breast cancer by other studies.Using a custom capture reagent, we sequenced the exons of these ~1,500 candidate genes in 952 additional early-onset cases, 269 first-degree relatives, and 218 cancer-free controls. We also included exome data from an additional 500 WHI participants.Taken together, our dataset encompasses sequencing data for 1,500 genes in 2,800 samples. Our analysis reveals that 10% of early-onset cases harbor loss-of-function mutations in DNA repair genes including *CHEK2* (2.7%), *ATM* (1.8%) and others. Genes involved in chromatin remodeling (*ARID1A*, *ARID1B*), cellular signaling (*AR*, *PTPN21, PTEN*), and other pathways were enriched for rare damaging mutations and offer new candidates for early-onset breast cancer susceptibility genes.

5

Five independent 6q25 breast cancer risk variants regulate *ESR1* and *RMND1* and display genotype-phenotype correlations. *S. Edwards[1], A. Dunning[2], K. Michailidou[3], K. Kuchenbaecker[3], D. Thompson[3], J. French[1], J. Beesley[1], C. Healy[2], S. Kar[2], K. Pooley[2], E. Dicks[2], D. Barrowdale[3], N. Sinnott-Armstrong[4], R. Cowper-Sallari[4,5], K. Hillman[1], S. Kaufmann[1], H. Sivakumaran[1], M. Moradi Marjaneh[1], E. Lopez-Knowles[6,7], M. Dowsett[6,7], P. Pharoah[2,3], J. Simard[8], P. Hall[9], M. Garcia-Closas[10,11], C. Vachon[12], G. Chenevix-Trench[1], A. Antoniou[2], D. Easton[2,3].* 1) Department of Genetics, QIMR Berghofer Medical Research Institute, Australia; 2) Department of Oncology, University of Cambridge, UK; 3) Department of Public Health and Primary Care, University of Cambridge, UK; 4) The Broad Institute of MIT and Harvard, Cambridge, USA; 5) The Princess Margaret Cancer Centre-University Health Network, Canada; 6) Breast Cancer Research, Breakthrough Breast Cancer Research Centre, UK; 7) Academic Biochemistry, Royal Marsden Hospital, UK; 8) Centre Hospitalier Universitaire de Québec Research Center, Laval University, Canada; 9) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden; 10) Division of Cancer Studies, Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, UK; 11) Division of Genetics and Epidemiology, Institute of Cancer Research, UK; 12) Department of Health Sciences Research, Mayo Clinic, Rochester, USA.

Single nucleotide polymorphisms (SNPs) at 6q25 are reported to be associated with breast cancer susceptibility, breast cancer risk for *BRCA1* mutation carriers and breast density. To date, however, attempts to identify the causal SNPs underlying the associations have been inconclusive. Here, we analysed 3872 SNPs across the 6q25 locus in 118,816 subjects from three international consortia and found evidence for five independent sets of correlated, highly trait associated variants (iCHAVs). The iCHAVs are distributed upstream, within introns, and downstream of *ESR1* - the most obvious target gene in the region. At all five sites, the minor allele of the candidate causal SNPs increased risks of ER- tumors and, with one exception, all are more strongly associated with risk of developing ER- than ER+ tumor-subtypes. We also identified associations with mammographic density, human ERB2 (HER2) tumor status and with high-grade breast cancer. The strongest candidate causal SNPs within each iCHAV lay in *cis*-regulatory elements and chromosome conformation capture assays confirmed they physically interact with the promoters of *ESR1, RMND1, C6orf211* and *CCDC170*. Allele-specific expression analyses identified significant associations between iCHAV1-3 SNPs and the allelic ratio of *ESR1* and *RMND1* transcripts. Furthermore, IHC in 150 normal breast samples showed iCHAV1 risk alleles to be associated with reduced ER levels. Reporter assays demonstrated that *cis*-elements within iCHAVs 1,2,4 and 5 act as transcriptional enhancers and one element within iCHAV3 acts as a silencer on target genes. Consistent with expression analyses, constructs including the risk alleles decreased *ESR1* and *RMND1* promoter activity. Electrophoretic mobility shift assays on representative SNPs displayed allele-specific transcription factor binding. Of these, the iCHAV4 risk allele disrupts a CTCF binding site and displays allele-specific chromatin looping, suggesting this contributes to reduced *ESR1* expression. This study provides definitive evidence for genetic control of both major breast tumour subtypes by genetic variants in the 6q25 locus. We also provide the first evidence of genetic risk factors for developing two rarer classes of tumor: the ER-/PR-/HER2+ subtype (responsive to Herceptin^{R}) and ER+/high-grade tumors. Our functional data implicate *ESR1* as the main target gene driving the associations, and highlights the potential importance of ER in establishing both ER- and ER+ breast cancer.

6

Breast cancer risk at the 5p12 locus is mediated through chromatin looping and regulation of FGF10 and MRPS30. *M. Ghoussaini[1,5], J. French[2,5], K. Michailidou[3], S. Nord[4], J. Beesley[2], J. Dennis[3], K. Hillman[2], S. Kaufmann[2], E. Dicks[3], S. Ahmed[1], M. Maranian[1], C.S. Healey[1], C. Baynes[1], C. Luccarini[1], M. Bolla[3], J. Wang[3], V.N. Kristensen[4], P.D.P. Pharoah[1,3], G. Chenevix-Trench[2], D.F. Easton[1,3], A.M. Dunning[1,5], S.L. Edwards[2,5], Breast Cancer Association Consortium.* 1) Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK; 2) Department of Genetics, Queensland Beghofer Institute of Medical Research, Brisbane, Australia; 3) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 4) Department of Genetics, University of Oslo, Oslo, Norway; 5) These authors contributed equally to the work.

Genome wide association studies (GWAS) have previously identified multiple breast cancer susceptibility loci on 5p12. In an effort to identify the likely causal variant(s), we performed fine-scale mapping of this region by genotyping 545 single nucleotide polymorphisms (SNPs) across a 1 Mb region in 89,050 subjects of European ancestry and 12,893 subjects of Asian ancestry from 50 case-control studies in the Breast Cancer Association Consortium (BCAC). Genotypes for 2,820 SNPs were imputed using the 1000 genomes project as a reference. Using forward stepwise logistic regression, we identified three independent risk signals; In the first signal, a single SNP (rs10941679 A/G) explained the association with breast cancer and the risk was restricted to ER-positive (ER+) breast tumours (per-G allele OR ER+=1.15, 95% CI 1.13-1.18 ; P=x10^{-30}). In signals two and three, 37 and 86 SNPs respectively could not be excluded from causality at odds>1:100 relative to the best SNP in each peak. Using chromatin conformation studies, we identified a long-range physical interaction between the enhancer/flanking sequence encompassing rs10941679 and the promoters of *FGF10* (Fibroblast Growth Factor 10) and *MRPS30* (Mitochondrial Ribosomal Protein S30). We then conducted expression quantitative trait (eQTL) analysis in normal breast tissue samples from 219 women and breast tumors from 1,019 women in the Norwegian Breast Cancer Study and the TCGA project. SNP rs10941679 was associated with a significant increase in the expression levels of *MRPS30* and *FGF10* in both normal and breast cancer tissues. We also measured the endogenous levels of *FGF10* and *MRPS30* mRNA in four ER+ breast cancer cell lines either homozygous for the protective A allele of rs10941679 (AA) or heterozygous (A/G). *FGF10* and *MRPS30* mRNA levels were significantly increased in heterozygous compared to homozygous cell lines. FGF10 is a member of the fibroblast growth factors and specifically binds to FGFR2 to control cell differentiation, proliferation, apoptosis and migration. FGF10 acts as an oncogene and is over-expressed in ~10% of human breast cancers. MRPS30 plays a key role in apoptosis. Taken together, these data suggest that breast cancer susceptibility at the 5p12 locus is likely to be mediated by over-expression of *FGF10* and *MRPS30*, two candidate genes for cancer pathogenesis.

7

Cross-cancer genome-wide pleiotropy analysis based on GAME-ON and GECCO across five common cancers: lung, ovary, breast, prostate and colon cancer. *R.J. Hung[1], G. Fehringer[1], P. Kraft[2], C.A. Hiaman[3], P. Pharoah on behalf of OCAC[4, 17], R. Eeles on behalf of PRACTICAL[5, 18], N. Chatterjee[6], D. Seminara[6], S. Chanock[6], F. Schumacher[3], S. Lindström[2], K. Stefansson[7], D.C. Christiani[2], H. Shen[8], K. Shiraishi[9], A. Takahashi[10], AABC. AAPC, JAPC, LABC, LAPC[19], Y. Bosse[11], M. Obeidat[12], M.L. Freedman[13], U. Peters on behalf of GECCO[16, 20], S. Gruber on behalf of CORET[3, 21], C.I. Amos on behalf of TRICL/ILCCO[14, 22], T.A. Sellers on behalf of FOCI[15, 23], D. Easton on behalf of BCAC[4, 24], D.J. Hunter on behalf of DRIVE[2, 25], B.E. Henderson on behalf of ELLIPSE[3, 26], Genetic Associations and Mechanisms in Oncology (GAME-ON) Network.* 1) Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada; 2) Harvard T.H. Chan School of Public Health, Boston MA; 3) University of South California, Los Angeles, USA; 4) University of Cambridge, Cambridge, UK; 5) Institute of Cancer Research, London, UK; 6) National Cancer Instititute, Bethesda, USA; 7) deCODE genetics, Amgen, Reykjavik, Iceland; 8) Nanjing Medical University School of Public Health, Nanjing, China; 9) Division of Genome Biology, National Cancer Center Research Institute. Tokyo, Japan; 10) Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; 11) Institut universitaire de cardiologie et de pneumologie de Québec, Department of Molecular Medicine, Laval University, Québec, Canada; 12) University of British Columbia Centre for Heart Lung Innovation, St. Paul's Hospital. Vancouver, Canada; 13) Dana-Farber Cancer Institute, Boston, USA; 14) Geisel School of Medicine, Dartmouth College, Lebanon; 15) Moffitt Cancer Center, Tampa, USA; 16) Fred Hutchinson Cancer Research Center, Seattle, USA; 17) OCAC, Ovarian Cancer Association Consortium; 18) PRACTICAL, Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome consortium; 19) AABC, American Breast Cancer Consortium; AAPC, African Ancestry Prostate Cancer Consortium; JAPC, Japanese American Prostate Cancer Consortium ; LABC, Latino American Breast Cancer Consortium; LAPC, Latino American Prostate Cancer Consortium; 20) GECCO, Genetic and Epidemiology of Colorectal Cancer Consortium; 21) CORET, Colorectal Transdisciplinary study; 22) TRICL/ILCCO, Transdisciplinary Research for Cancer of Lung and International Lung Cancer Consortium; 23) FOCI, Transdisciplinary Cancer Genetic Association and Interacting Studies; 24) BCAC, Breast Cancer Association Consortium; 25) DRIVE, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; 26) ELLIPSE, Elucidating Loci in Prostate Cancer Susceptibility.

Background. Identifying genetic variants with pleiotropic associations can uncover common pathways influencing multiple cancers and enable further understanding of cancer susceptibility. **Method.** We conducted a genome-wide cross-cancer pleiotropy analysis across five common cancers: lung, ovary, breast, prostate and colon cancer from the GAME-ON/GECCO Network with a total of 61,851 cases and 61,820 controls of European ancestry using the association analysis based on subsets (ASSET) method. Results were validated in additional independent studies from Harvard, deCODE and Collaborative Oncological Gene-Environment Study (iCOGS) with a total of 55,789 cases and 330,490 controls of European ancestry. We have also evaluated the generalizability in Chinese, Japanese, Latinos and African Americans. Expression quantitative trait loci (eQTL) analyses were conducted for validated loci based on data derived from lung, ovary and prostate tissues.**Results:** We identified a novel pleiotropic association at 1q22 with a variant associated with breast and lung squamous cell carcinoma (overall P-value for both cancers combined=8.9×10^{-8}) in European descendants and the results were validated in the replication set. The eQTL analysis of this region showed a consistent association with *ADAM15*/*THBS3* gene expression in lung tissues in three independent studies. New pleiotropic associations were also found at previously known cancer loci: variants at a known *BRCA2* locus for lung and breast cancer were associated with serous ovarian cancer (overall p-value=4.0×10^{-8}); a known breast cancer locus, *CASP8*/*ALS2CR12*, with a variant associated with prostate cancer (overall P-value=1.9×10^{-8}), and a known breast cancer locus, *CDKN2B-AS1*, where one variant was associated with lung adenocarcinoma (overall P-value=1.0×10^{-5}) and a second was associated with prostate cancer (overall P-value=9.5×10^{-7}).**Conclusions:** Our results provide important insights into common carcinogenesis across multiple major cancers and highlight the value of pleiotropy analysis.

8

Common genetic variants modify breast and prostate cancer risks for male *BRCA1* and *BRCA2* mutation carriers: implications for risk prediction. *L. Ottini[1], J. Lecarpentier[2], K. Kuchenbaecker[2], K. Offit[3], F. Couch[4], J. Simard[5], M. Thomassen[6], R. Schmutzler[7], G. Chenevix-Trench[8], D. Easton[2], A.C. Antoniou[2] on behalf of CIMBA.* 1) Department of Molecular Medicine, Sapienza University of Rome, Rome, Italy; 2) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 3) Clinical Genetics Research Laboratory, Department of Medicine, Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA; 4) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA; 5) Centre Hospitalier Universitaire de Québec Research Center and Laval University, Quebec City, Quebec, Canada; 6) Department of Clinical Genetics, Odense University Hospital, Odense C, Denmark; 7) Center for Hereditary Breast and Ovarian Cancer, Medical Faculty, University Hospital Cologne, Germany; 8) Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, QLD , Australia.

BRCA1 and *BRCA2* mutations confer increased risks of developing male breast and prostate cancer. More than 100 common variants are now known to be associated with breast cancer risk in women and more than 80 common variants are associated prostate cancer risk. No study to date has investigated the associations of common genetic variants and breast or prostate cancer risk for men with mutations in *BRCA1* and *BRCA2*. In our study, we used data on 1802 male *BRCA1* and *BRCA2* mutation carriers (288 diagnosed with breast cancer, 243 diagnosed with prostate cancer) from the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA) who were genotyped using the custom Illumina Oncoarray to perform a GWAS in male *BRCA1/2* mutation carriers. We also investigated the combined effects of known breast and prostate cancer susceptibility variants on cancer risk for male mutation carriers by constructing polygenic risk scores (PRS). No variant was associated with risk of prostate or male breast cancer at genome-wide significance ($P<5 \times 10^{-8}$). However, a PRS constructed on the basis of 77 known female breast cancer susceptibility variants was associated with male breast cancer risk for both *BRCA1* and *BRCA2* mutation carriers (Hazard Ratio (HR) per standard deviation (SD) of PRS: 1.32 (95%CI: 1.15-1.52, $p=7 \times 10^{-5}$). A stronger association was observed with a PRS based on SNPs that are associated with female estrogen receptor (ER) positive breast cancer (HR per SD of ER-positive PRS: 1.33, $p= 4 \times 10^{-5}$ VS HR per SD of ER-negative PRS: 1.13, p= 0.05). Similarly, the PRS based on 81 known prostate cancer susceptibility variants was strongly associated with prostate cancer risk for male *BRCA1/2* mutation carriers (HR per SD of PRS: 1.48 (95%CI: 1.28-1.69, $p= 7 \times 10^{-8}$). These analyses demonstrate that the breast and prostate cancer risks for male *BRCA1/2* mutation carriers are modified by common genetic variants, and could have implications for the cancer risk management and screening in male mutation carriers. For example, the predicted prostate cancer risk by age 80 for the 5% of male *BRCA1* mutation carriers with the lowest prostate cancer PRS is <11%, but it is greater than 40% for the 5% of men with the highest PRS.

9

Whole exome sequencing in 75 high-risk families identifies eight previously unknown prostate cancer susceptibility genes. *D.M. Karyadi[1], M.S. Geybels[2], E. Karlins[1], B. Decker[1], L. McIntosh[2], S. Kolb[2], S.K. McDonnell[3], S. Middha[3], L.M. FitzGerald[4], M.S. DeRycke[5], D.J. Schaid[3], S.N. Thibodeau[5], J.L. Stanford[2,6], E.A. Ostrander[1].* 1) Cancer Genetics and Comparative Genomics Branch, NHGRI/NIH, Bethesda, MD; 2) Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA; 3) Department of Health Sciences Research, Mayo Clinic, Rochester, MN; 4) Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Victoria, Australia; 5) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; 6) Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA.

Prostate cancer (PCa) is the most common non-cutaneous tumor in men from the United States with 220,800 estimated new cases and 27,540 expected deaths in 2015. Epidemiological studies suggest that PCa has a strong genetic component with approximately 42% to 58% of risk attributed to genetic factors. The disease is genetically heterogeneous and predicted to be caused by a continuum from common, low-penetrant to rare, high-penetrant variants. While genome-wide association studies have been successful in confirming at least 100 loci associated with PCa risk in Europeans, these loci likely represent common, low-penetrant variants and, to date, only a few moderately to highly penetrant susceptibility variants have been consistently implicated in PCa. In an effort to identify low-frequency, moderately penetrant variants, we leveraged whole exome sequence (WES) data with array-based SNP haplotypes from 75 hereditary PCa families. WES data was available for 160 affected men in the 75 families with one to six affected sequenced per family and 31 families with multiple WES men. Array-based SNP haplotypes were generated for 508 individuals including all 373 affected men with DNA available. Integrating the WES and haplotype data allowed us to predict the affected carrier frequency for each variant, which is key in identifying moderately penetrant variants. We then performed rigorous disease model-based variant filtering taking into account the likely genetic heterogeneity and incomplete penetrance of PCa susceptibility variants. Analysis of 341 of the most compelling candidate risk variants in a population-based, case-control study of 2,495 individuals confirmed nine variants significantly associated with PCa risk, which have predicted protein consequences in *SWSAP1*, *HOXB13*, *D2HGDH*, *CHAD*, *EPHA8*, *TANGO2*, *BRD2*, *PPP6R2* and *OR5H14*. Eight of the variants are in genes not previously implicated in PCa susceptibility with *BRD2* A605P having the strongest association (OR = 4.99, 95% confidence interval, 1.09 – 22.86). Inheriting at least one of the nine susceptibility alleles increased risk 2.5-fold (95% confidence interval, 1.81 – 3.36) and can explain 13% of familial risk for PCa in the population. The results of this study support that exome sequencing of a limited number of individuals within hereditary cancer families combined with haplotype data offers a robust strategy for identifying low-frequency, moderate-penetrance risk variants contributing to PCa susceptibility.

10

Population and evolutionary genomics of prostate cancer-associated variants: implications for health disparities in men of African descent. *J. Lachance[1], C. Hanson[1], M.E.B. Hansen[2], S.A. Tishkoff[2], T.R. Rebbeck[3].* 1) School of Biology, Georgia Institute of Technology, Atlanta, GA; 2) Departments of Biology and Genetics, University of Pennsylvania, Philadelphia, PA; 3) Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA.

Prostate cancer is a highly heritable disease that disproportionately affects African and African-American men. To determine the genetic architecture of differences in prostate cancer risk across populations, we integrated GWAS results and scans of selection with whole genome sequencing and genotype array data from 45 African and 19 non-African populations. Risk alleles with elevated frequencies in African populations were identified, as were GWAS hits in genomic regions that are highly divergent across continents (e.g. rs2660753, rs17023900, and rs9284813). Although the majority of prostate cancer-associated loci are in neutrally evolving genomic regions, we found multiple instances where alleles at prostate cancer-associated loci have hitchhiked with linked alleles that are under selection. For example, a protective allele at 2q37.3 appears to have risen to high frequency in Europe due to selection for lighter pigmentation, and a risk allele at 19q13.2 may have risen to high frequency in Africa due to selection for resistance to tropical diseases. Genetic risk scores correctly predict elevated prostate cancer risk in African-Americans over Europeans, and genetic risk scores suggest that prostate cancer risk is highest in West African populations and lowest in non-African populations. Despite the polygenic nature of prostate cancer (over 100 loci are known to contribute to disease risk), we find that a small number of loci of major effect drive the difference in predicted risk across populations (e.g. rs17631542, rs10505483, rs4242382, rs1447295, and rs10505477).

11

Integrative approaches for large-scale transcriptome-wide association studies. *A. Gusev[1,2], A. Ko[3,4], H. Shi[5], G. Bhatia[1,2], W. Chung[1], B.W.J.H. Penninx[15], R. Jansen[15], E.J.C. de Geus[16], D.I. Boomsma[16], F.A. Wright[17], P.F. Sullivan[13,14], E. Nikkola[3], M. Alvarez[3], M. Civelek[6], A.J. Lusis[3,6], T. Lehtimaki[7], M. Kahonen[8], I. Seppala[7], O.T. Raitakari[9,10], J. Kuusisto[11], M. Laakso[11], A.L. Price[1,2], P. Pajukanta[3,4], B. Pasaniuc[3,5,12].* 1) Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; 2) Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA; 3) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California, USA; 4) Molecular Biology Institute at UCLA, Los Angeles, California, USA; 5) Bioinformatics Interdepartmental Program, UCLA, Los Angeles, California, USA; 6) Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California, USA; 7) Department of Clinical Chemistry, Fimlab Laboratories and University of Tampere School of Medicine, Tampere, Finland; 8) Department of Clinical Physiology, Pirkanmaa Hospital District and University of Tampere School of Medicine, Tampere, Finland; 9) Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland; 10) Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku and Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland; 11) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 12) Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California, USA; 13) Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, 27599, USA; 14) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, 171 77, Sweden; 15) Department of Psychiatry, VU Medical Center, Amsterdam, The Netherlands; 16) Department of Biological Psychology, VU University, Amsterdam, The Netherlands; 17) Bioinformatics Research Center, Department of Statistics, Department of Biological Seicnes; North Carolina State University, Raleigh, North Carolina, USA.

Many genetic variants influence complex traits by modulating gene expression, but the mechanistic steps between genetic variation, expression, and trait are generally not well understood. Studies of this relationship using direct measurements of expression and disease have been limited in scope due to specimen availability and cost. Here we introduce a powerful strategy that integrates gene expression measurements with large-scale genome-wide association studies (GWAS) to discover novel genes whose cis-regulated expression is associated with trait. We use a reference panel of individuals with both genotype and gene expression data to quantify the cis-genetic component of expression and impute it into a large cohort of individuals with genotype data but no measured expression, empowering us to test for expression-trait association and conduct a transcriptome-wide association study (TWAS). By using a reference panel of LD, this method can also impute the TWAS association from publically available GWAS association statistics. Simulations show that the TWAS is substantially more powerful at identifying expression-trait relationships than either traditional GWAS or association informed by individually significant expression associations (eQTLs), especially in the presence of multiple variants effecting expression (increasing power from 18% for GWAS to 92% for TWAS). In contrast to previous methods, TWAS does not require the presence of a significant association at the locus but instead aggregates all of the cis effects. We applied this approach to SNP and expression data from blood and adipose tissue measured in ~3,000 individuals imputed into summary statistics from multiple large GWAS totaling >900,000 individual phenotypes, performing the first TWAS of this size. At previously known loci, we find that the TWAS identified genes that explain significantly more variance than the best eQTL. By imputing into two lipid GWAS with increasing sample size, we show that the novel genes identified by our method are highly predictive of genome-wide significant SNP associations in larger studies (hypergeomtric $P=1x10^{-24}$). Overall, we identified 676 significant gene-trait associations, of which 70 were novel and did not overlap a genome-wide significant SNP. We replicated many of the genes using two external expression studies. Our results showcase the power of integrating genotype, gene expression and phenotype to gain insights into the genetic basis of complex traits.

12

Improved identification of the genetic basis of disease by integrated analysis of RNA-seq together with whole-genome and exome-based sequencing data. *D.W. Craig, S. Szelinger, A.M. Claasen, I. Schrauwen, R.F. Richholt, M. De Both, B.E. Hjelm, S. Rangasamy, A.L. Siniard, A.L. Courtright, M.J. Huentelman, V. Narayanan.* Center for Rare Childhood Disorders, Translational Genomics Research Institute, Phoenix AZ, USA.

Recent advancements in genomic technologies have markedly increased our ability to genetically diagnose of rare childhood disorders through whole-exome and whole-genome sequencing of genomic DNA. Many groups now report that whole-exome sequencing can lead to a genetic diagnosis in approximately 25-35% of undiagnosed cases. Advancing these approaches further, generation and integrative analysis of RNA-seq data with genomic exome sequencing data may be able to improve our ability to diagnose and understand the genetic basis of rare childhood disorders. With a few key exceptions, reduction to practice of integrative analysis of RNA and DNA for diagnosis of rare childhood disorders has largely not been reported. Within this study, we report on the development of integrated methods for joint RNA/DNA analysis and their utility within our clinical research protocols focused on genetic diagnosis of rare childhood disorders. Within our study, RNA-seq and exome/whole-genome sequencing was conducted on whole-blood lymphocytes from 75 individuals encompassing 25 family trios where a rare childhood condition was presented. We describe a framework for non-parametric multivariate outlier analysis, such that multiple analyses approaches for of RNA-seq data are considered together, and can be used quantitatively to better prioritize candidate genetic variants from whole-exome and genome sequencing. We show how our approach allows joint integration of gene abundance, differential gene expression, exon abundance, and alternative exon usage data in the context of rare, private or de novo candidate variants. We demonstrate how re-prioritized variants utilizing integrated RNA/DNA analysis both corroborated previously identified causal variants where the genetic basis of disease was known. We also highlight where new candidate variants, clearly functional in nature, were found in cases where the genetic diagnosis was unknown. In these cases, we identified novel functional elements (e.g. cryptic splice-site variants in genes relevant to disease), inferred false negative variants impacting transcription, and observed altered transcription due epigenetic modifiers (e.g. methylation) that were relevant in the manifestation of disease. Overall, our results show how integrative genomic analysis of RNA and DNA aid in the diagnosis of disease and help provide novel insight into molecular underpinnings of un-characterized conditions.

13

Comprehensive transcriptome analysis using synthetic long read sequencing reveals molecular co-association and conservation of distant splicing events. *H. Tilgner, F. Jahanbani, M. Rasmussen, M. Snyder.* Genetics, Stanford University, Stanford, CA.

RNA molecules may contain several variable sites (TSS, splice sites, polyA-sites, RNA-editing), which allow for complex combinatorial patterns. Genome annotations contain many of these isoforms but do not exhaustively describe all isoforms and the dependencies between distant variable sites. To overcome this and the limitations of short-read RNA sequencing, we have invested considerable energy into long-read approaches, employing the 454[1], the Pacific Biosciences platform (PacBio)[2,3], and most recently the Illumina based SLR-RNA-Seq[4]. Using variable length and (quasi-) single molecule approaches like PacBio or SLR-RNA-Seq, we find between 14 and 19.5% of all spliced molecules to be inconsistent with all annotated isoforms. We then aimed to test for dependencies between distant variable sites on RNA-molecules. Using PacBio we could in a limited number of cases of allele-specific splicing, irrespectively of the distance between the SNP and the exon. Using SLR-RNA-Seq[4] in a sample of human brain RNA, we constructed ~5 million long reads of ~1.9kb average length. Gene expression measurements in molecules per million (MPM) correlate highly with short read FPKMs, apart from an underrepresentation of very short genes and Percent-Isoform (PI)-values for major isoforms correlate highly between two mouse brains. We discovered 70-160 (depending on datasets and FDR) distant molecularly associated pairs (dMAPs) of alternative exons that are separated by one or more constitutive exon. Many of these exon-pairs were entirely coding, which strongly suggests the existence of a phased proteome, in which distant peptides are included into protein molecules in a coordinated manner. With shallower sequencing in two mice brains, we find 16 such dMAPs, 9 of which are conserved between human and mouse – suggesting their importance for cellular function. Deep long-read sequencing such as the approach presented here will be very useful in completely describing transcriptomes and the proteomes they encode. Furthermore the existence of long-distant coordinated variable pairs raises many questions about the molecular mechanisms causing coordination.**references**1. Tilgner H et al. G3 (Bethesda). 2013 doi: 10.1534/g3.112.004812.2. Sharon D*, Tilgner H*, Grubert F, Snyder M. Nat Biotechnol. 20133. Tilgner H*, Grubert F*, Sharon D*, Snyder MP. PNAS. 20144. Tilgner H*, Jahanbani F* et al. Nat Biotechnol. 2015; doi: 10.1038/nbt.3242.

14

Comprehensive genome and transcriptome structural analysis of a breast cancer cell line using PacBio long read sequencing. *M. Nattestad[1], K. Ng[2], S. Goodwin[1], T. Baslan[1], J. Gurtowski[1], F. Sedlazeck[1], E. Hutton[1], E. Tseng[3], J. Chin[3], T. Beck[2], Y. Sundaravadanam[2], M. Kramer[1], E. Antoniou[1], J. Hicks[1], M. Schatz[1], W.R. McCombie[1].* 1) Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; 2) Ontario Institute for Cancer Research, Toronto, Ontario Canada; 3) Pacific Biosciences, Menlo Park, CA.

Genomic instability is one of the hallmarks of cancer, leading to widespread copy number variations, chromosomal fusions, and other structural variations in many cancers. The breast cancer cell line SK-BR-3 is an important model for HER2+ breast cancers, which are among the most aggressive forms of the disease and affect one in five cases. Through short read sequencing, copy number arrays, and other technologies, the genome of SK-BR-3 is known to be highly rearranged with many copy number variations, including an approximately twenty-fold amplification of the HER2 oncogene, along with numerous other amplifications and deletions. However, these technologies cannot precisely characterize the nature and context of the identified genomic events and other important mutations may be missed altogether because of repeats, multi-mapping reads, and the failure to reliably anchor alignments to both sides of a variation. To address these challenges, we have sequenced SK-BR-3 using PacBio long read technology. Using the new P6-C4 chemistry, we generated more than 70X coverage of the genome with average read lengths of 9-13kb (max: 71kb). Using Lumpy as well as our novel assembly-based algorithms for analyzing split-read alignments, we have developed a detailed map of structural variations in this cell line. Taking advantage of the newly identified breakpoints and combining these with copy number assignments, we have developed an algorithm to reconstruct the mutational history of this cancer genome. From this we have characterized the amplifications of the HER2 region, discovering a complex series of nested duplications and translocations between chr17 and chr8, two of the most frequent translocation partners in primary breast cancers. We have also carried out full-length transcriptome sequencing using PacBio's Iso-Seq technology, which has revealed a number of previously unrecognized gene fusions and isoforms. Combining long-read genome and transcriptome sequencing technologies enables an in-depth analysis of how changes in the genome affect the transcriptome, including how gene fusions are created across multiple chromosomes. This analysis has established the most complete cancer reference genome available to date, and is already opening the door to applying long-read sequencing to patient samples with complex genome structures.

15

Tissue-specific transcriptome-wide networks reflect joint regulation of alternative splicing and gene expression. *A. Saha[1], Y. Kim[1], D. Knowles[2], S. Mostafavi[3], A. Battle[1], The GTEx Consortium.* 1) Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA; 2) Department of Radiology, Stanford University, Stanford, CA, USA; 3) Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada.

Alternative splicing plays an important role determining cellular function in eukaryotes; its causal role in human diseases including cancer is well established. However, the regulation of alternative splicing is not fully understood, including the factors that lead to tissue-specific control of alternative splicing. Unraveling the regulation of alternative splicing in the form of a network has the potential to elucidate complex biological mechanisms. Until recently, well-studied regulatory network inference methods predominantly used total gene expression data. However, with the advent of RNA-sequencing, we are now able to simultaneously quantify a diverse range of transcription phenotypes, including alternative splicing, non-coding transcripts, and allele-specific expression. We can discover new regulatory relationships by using these phenotypes during network construction. Here we build transcriptome-wide networks (TWN) over both gene expression and alternative splicing from RNA-seq data using sparse Gaussian Markov random fields. We identify multiple types of candidate regulatory relationships where i) gene expression regulates alternative splicing, ii) alternative splicing regulates gene expression, iii) gene expression of a gene regulates that of other genes, and iv) alternative splicing of a gene regulates that of other genes. We learn TWNs for 23 different human tissues using RNA-seq data from the GTEx project. Our networks are enriched with edges between known transcription factors and their target genes, demonstrating our model's effectiveness to find true regulatory relationships. Further, candidate splicing regulator genes identified by learned TWNs include many known RNA-binding proteins (RBPs), reflecting the role of RBPs in splicing. We integrate expression quantitative trait loci and allele-specific expressions to validate potentially causal relationships in TWNs, and demonstrate improved power to detect trans expression quantitative trait loci. Numerous tissue-specific regulatory relationships are identified by the TWNs, and are informative in characterizing context-specific behavior of regulatory elements. Our transcriptome-wide networks provide an opportunity to understand tissue-specific regulatory mechanisms for both alternative splicing and gene expression.

16

Massively parallel quantification of the regulatory effects of non-coding variation reveals functional variants associated with fetal adiposity. *C. Guo[1,2], C.M. Vockley[2,3], W.H. Majoros[2,4], M. Nodzenski[5], D.M. Scholtens[5], M.G. Hayes[6], W.L. Lowe[6], T.E. Reddy[4,7].* 1) University Program in Genetics & Genomics, Duke University, Durham, NC; 2) Center for Genomic & Computational Biology, Duke University Medical School, Durham, NC; 3) Department of Cell Biology, Duke University Medical School, Durham, NC; 4) Program in Computational Biology & Bioinformatics, Duke University, Durham, NC; 5) Department of Preventive Medicine, Division of Biostatistics, Northwestern University Feinberg School of Medicine, Chicago, IL; 6) Division of Endocrinology, Metabolism & Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL; 7) Department of Biostatistics & Bioinformatics, Duke University Medical School, Durham, NC.

Recent evidence strongly supports the hypothesis that non-coding mutations that alter gene expression are a major contributor to the heterogeneity of heritable, complex human traits. To test this hypothesis, we develop a novel high-throughput approach to measure the effects of non-coding variants on gene expression across human populations. We take advantage of the previously developed STARR-seq assay and adapt it at the population scale by assaying candidate regulatory elements captured from the genomes of a GWAS cohort. By assaying DNA directly from donor genomes, we measure regulatory effects of rare and private variants that are not present in existing databases of human genetic variation; and are also able to determine the cumulative effects of regulatory variants empirically. We apply our assay to 3q25.31, a locus we previously found to be associated with adiposity at birth in a GWAS of 4281 newborns. Because none of the associated variants reside in protein coding exons, we hypothesize that they instead alter expression of nearby genes. We identify ~100 candidate regulatory elements across the associated locus using genome-wide measurements of open chromatin across ~50 different cell types including pre-adipocytes, liver, and islet β cells. The elements span 250 kb of the genome surrounding the lead GWAS SNP rs900400. The region encompass all variants in LD (r^2 >0.5) with rs900400. We capture the candidate regulatory elements from 760 newborns at the extreme tails (90^{th} and 10^{th} percentiles) of birth weight; and use high-throughput reporter assays to identify ~300 regulatory variants, most of which were rare (MAF<0.01). Associating variation in regulatory activity with newborn adiposity reveals regulatory regions that independently associated with sum of skinfolds, suggesting an underlying regulatory mechanism that contributes to phenotype. We also identify features of non-coding variants that are enriched in those with regulatory element activity. Because the approach developed relies only on donor DNA and a relevant cell model, it can be broadly applied to uncover regulatory mechanisms contributing to various human traits and disease.

17

Detection of trans and cis splicing QTLs through large scale cancer genome analysis. *K. Lehmann[1], A. Kahles[1], C. Kandoth[1], N. Schultz[1], O. Stegle[2], G. Rätsch[1].* 1) Computational Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA; 2) European Bioinformatics Institute, Cambridge, UK.

The comprehensive survey of molecular characteristics provided by The Cancer Genome Atlas (TCGA) enables large scale analyses across multiple cancers. However, sophisticated tools for the joint analysis of the thousands of samples that tackle the cancer specific challenges are needed. In an effort to enable joint analysis, we have re-aligned and re-analyzed RNA and whole exome sequencing data of ~4,000 individuals across 11 cancer types in a uniform manner. We used the newly developed open source SplAdder pipeline to count gene expression as well as annotate and quantify a comprehensive set of alternative splicing events. We identified threefold more high confidence alternative splicing events than annotated in the GENCODE annotation which reflect cancer-specific and tissue-specific splicing variation. Comparisons to matching tissue normal samples confirm a ~20% increase of splicing complexity in tumor samples. We have identified sets of genes with splicing changes that recurrently occur in tumor samples (>10%) but are virtually never observed in normal samples or ENCODE cell lines (<0.5%) and could be possible targets for new drugs. While population structure is one of the most severe confounding factors in the analysis of quantitative trait loci (QTL), tumor samples open up many new additional challenges. Tumor-specific somatic mutations and recurrence patterns as well as sample heterogeneity can lead to spurious associations. Thus, we have developed a new strategy to perform a common variant association study using linear mixed models on tumor samples enabling us to account for tumor specific genotypic and phenotypic heterogeneity in addition to population structure. Due to sample size constraints, many previous QTL studies have been limited to the analysis of cis-associated variants. The large sample size available from TCGA enables us to overcome this limitation and discover trans-associated variants as well. We can demonstrate that we find cis-associations for ~10% of the analyzed genes, of which a large fraction replicates across tissue and cancer types. We also confirm recently reported trans-associations in the splice factors U2AF1 as well as SF3B1.

18

The landscape of X inactivation across human tissues: from single cells to population sequencing. *T. Tukiainen[1,2], A. Villani[2,3], M. Rivas[2], A. Kirby[1,2], D. DeLuca[2], R. Satija[2,4], A. Byrnes[1,2], J. Maller[1,2], T. Lappalainen[4,5], A. Regev[2], N. Hacohen[2,3], K. Ardlie[2], D. MacArthur[1,2], The GTEx Project Consortium.* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Broad Institute of Harvard and MIT, Cambridge, MA; 3) Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, MA; 4) New York Genome Center, New York, NY; 5) Columbia University, New York, NY.

Incompleteness and skewing of X chromosome inactivation (XCI) can result in biases in disease susceptibility and presentation between sexes and across individuals, but the full extent and heterogeneity of XCI remains unclear. We have deployed several complementary approaches based on high-throughput RNA sequencing to comprehensively profile the landscape of XCI across multiple human tissues.Using gene expression data from the GTEx consortium, including more than 30 tissue types and over 350 individuals, we show that a large majority of previously reported escape genes demonstrate male/female expression differences detectable at population-level. For many of these genes sex-biased expression is present and directionally similar across the various tissues studied, a pattern distinct from autosomal sex-biased expression, suggesting XCI is tightly and uniformly regulated across human tissues. Notably, however, escape genes close to a boundary of an escape domain (e.g. *KAL1*) show more tissue heterogeneity and subtle sex-bias.By assessing the degree of allelic imbalance across the X chromosome from deep sequencing of 16 tissues from a female presenting with completely skewed XCI we further confirm the observation of largely consistent gene inactivation status across tissues, with *KAL1* being a notable exception showing tissue-specific escape. Additionally, such data allows for the interrogation of the inactivation state of multiple genes, therefore for instance replicating candidates from the population-level analysis (e.g. *ZRSR2*).To complement these observations we have analyzed single-cell RNA-seq data from a total of 384 dendritic and lymphoblastoid cells from four deeply sequenced females, allowing us to directly assess the inactivation state of 150 X-chromosomal genes. These analyses highlight well-known escape genes (e.g. *USP9X*), suggest novel candidates and confirm variable escape genes (e.g. *TIMP1*) and elaborate the underlying dynamics. In line with the known incomplete and variable nature of XCI, we find in total that approximately 20% of the assessed genes appear to fully or partially escape from inactivation.Together these analyses provide a comprehensive view of the landscape of escape from XCI in adult tissues, essential for understanding the impact of this process on sex differences, sex chromosome aneuploidies and inter-female variability.

19

The contribution of the cysteinyl leukotriene 1 (CysLT1) gene and other genetic loci to atopic asthma in the Tristan da Cunha population. *M.D. Thompson[1], J. Stankova[2], M. Clunes[3], G.E. Rovati[4], D.E. Cole[1], M.C. Maj[5], V. Capra[6], D.L. Duffy[7].* 1) Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; 2) Division of Immunology and Allergy, Department of Pediatrics, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, QC, Canada; 3) Department of Physiology/Neuroscience, School of Medicine, Saint George's University, P.O. Box 7, St. George's, Grenada; 4) Dipartimento di Scienze Farmacologiche e Biomolecolari, Università degli Studi di Milano, Milano, Italy; 5) Department of Biochemistry, School of Medicine, Saint George's University, P.O. Box 7, St. George's, Grenada; 6) Department of Health Sciences, University of Milan, San Paolo Hospital, Italy; 7) QIMR Berghofer Medical Research Institute, 300 Herston Road, Herston, Queensland 4006, Australia.

We present an analysis identifying at least three genes that contribute to the atopic asthma phenotype found in the Tristan da Cunha population. This remote island in the South Atlantic has an approximately 45% prevalence of atopy and a 36% prevalence of asthma. The population represents a unique opportunity for genetic study since it derives from only seven founders. We previously described a rare coding cysteinyl leukotriene 1 receptor gene (CysLT1, *CYSLTR1*, Xq13-21.1) variant that is associated with disease on the island. The CysLT1 mutation encodes a Gly300Ser variant not seen in any sequences from the Exome Aggregation Consortium collection, and may be the first example of an X-linked private mutation that confers risk for asthma. We no estimate that at least one-quarter of asthma and atopy in female Tristanians could be attributed to the CysLT1 Gly300Ser variant. In addition, a 601A>G variant in the cysteinyl leukotriene 2 receptor gene (*CYSLTR2*, 13q14.2), 601A>G that encodes a Met201Val change was also seen at high frequency in the population (14.2% in Tristan da Cunha, versus 2.6% in persons of European descent) and could account for over 10% of the risk for asthma in this population. The *CYSLTR2* variant may therefore be necessary but not sufficient for the development of asthma. Our *in vitro* work showed that the disruptions we identified in both receptors are likely to be physiologically relevant. While the CysLT1 variant was found to be activating with respect to ligand binding, Ca^{2+} flux and inositol phosphate (IP) generation, the CysLT2 variant was found to be inactivating: suggesting that atopy pathogenesis may be exacerbated by the loss of CysLT2 negative regulation of CysLT1 signalling. Although risk for atopic asthma was heightened when variants of both receptors were inherited, these variants do not account for all of the risk – suggesting that at least one more gene, in addition to environmental factors such as smoking, may contribute to the phenotype. Results are interpreted with respect to their relevance to other island and mainland populations. It is interesting to note that the exacerbation of CysLT1 signalling by the Gly300Ser Tristan da Cunha mutation may be blocked selectively by antagonists such as Montelukast. The advances in asthma research discussed will lead to improved diagnosis and treatment of patients in the context of global health.

20

Utilizing an African specific genotyping array for a large-scale GWAS for Asthma in African Americans. *H.R. Johnston[1], N. Rafaels[2], D. Hu[3], D. Torgerson[3], S. Chavan[2], J. Gao[1], G. Abecasis[4], M. Hansen[5], R. Mathias[2], Z.S. Qin[1], K. Barnes[2], Y.J. Hu[1], CAAPA Consortium.* 1) Department of Biostatistics and Bioinformatics, Emory University Rollins School of Public Health, Atlanta, GA; 2) Department of Medicine, Johns Hopkins University, Baltimore, MD; 3) Department of Medicine, University of California at San Francisco, San Francisco, CA; 4) University of Michigan School of Public Health, Ann Arbor, MI; 5) Illumina, Inc. San Diego, CA.

The **C**onsortium on **A**sthma among **A**frican-ancestry **P**opulations in the **A**mericas (CAAPA) includes high coverage whole genome sequence data (~30x depth) on ~1,000 subjects of African ancestry and extends the patterns of variation catalogued in the Thousand Genomes Project and Exome Sequencing Project to a spectrum of populations representing a wide range of African ancestry in the Americas. An interim data freeze (N=643) of CAAPA includes: 329 African Americans; 125 African Caribbeans; 164 African ancestry samples with a notable Latino component; and 25 samples from Nigeria. One of the primary goals of CAAPA is to develop an 'African Diaspora Power Chip' to address the concern that current commercially available GWAS chips have made a limited effort to tag African specific variation. The African Diaspora Power Chip utilizes sequence variants from within CAAPA to tag as much African specific variation as possible. The successful design of design of the ADPC, with projected imputation coverage of greater than 85% down to a minor allele frequency of 0.5% in 1000 Genome Project African populations, gives us the best available coverage of African variants available on the market. This level of coverage is accomplished by significantly skewing the MAF spectrum of the ADPC toward low-frequency variants. Pairing the ADPC with a standard GWAS array, such as OmniExpress, allows for comprehensive coverage of the entire allele frequency spectrum. The CAAPA Asthma GWAS study includes 9230 individuals from three populations: African Americans, Puerto Ricans and Barbadians. Utilizing genotypes from the ADPC alone, we have identified eight suggestive association results. These results have been found on chromosomes 1, 8, 9, 11, 16, 20 and X. The result at chr9:6144332 is in close proximity to previously identified associations near the IL33 gene. Given that this initial analysis includes only the SNPs on the ADPC but no imputed SNPs, this is a solid start from which to move forward, but was never intended to deliver comprehensive results. The next step will involve the combination of standard GWAS array data with the ADPC array data for each individual. This combined data set will then be used as the scaffold for imputation with both the 1000 Genomes Project African population and CAAPA sequencing results as reference panels. This will enable much finer scale analysis of potential associations with Asthma within the CAAPA study.

21

Integration of genome-wide association data and human protein interaction networks identifies a gene sub-network underlying childhood-onset asthma. *Y. Liu[1, 2], M. Brossard[1, 3], C. Sarnowski[1, 3], P. Margaritte-Jeannin[1, 2], F. Llinares[4], A. Vaysse[1, 2], M.H. Dizier[1, 2], E. Bouzigon[1, 2], F. Demenais[1, 2], GABRIEL asthma consortium.* 1) UMR-946, INSERM, Paris, France; 2) Université Paris Diderot, Paris, France; 3) Université Paris Sud, Paris, France; 4) ETH, Basel, Switzerland.

Genome-wide association studies (GWASs) have identified 21 loci associated with asthma. However, these loci account for a small part of asthma susceptibility. These GWASs, which focused on single-SNP analysis, are underpowered to detect SNPs with small effect. Alternative approaches, such as network-based analysis that uses information from the Human Protein Interaction Network (HPIN) to search for groups of genes which may jointly contribute to disease risk, have been proposed. To identify new asthma genes, we performed an integrated analysis of HPIN and GWAS data of childhood-onset asthma.We used two datasets from the GABRIEL Asthma Consortium that consisted of the outcomes of two meta-analyses of 9 childhood asthma GWASs each (including 3,031 cases/2,893 controls and 2,679 cases/3,364 controls, respectively). GWAS signals were overlaid to HPIN by assigning SNPs to genes and using gene-wise P-values obtained through circular genomic permutations (CGP). Modules enriched with childhood asthma-associated genes were generated by a dense module search (DMS) strategy. We selected the gene modules that showed the highest pairwise similarity between the two datasets. These modules were further evaluated for their association with asthma using CGP and for their biological relevance through pathway analysis using DAVID.We identified 10 gene-module pairs that had high similarity (from 0.4 to 0.6) between the two datasets. By merging the selected modules within each dataset and intersecting the two gene lists, we identified a sub-network consisting of 91 genes and 106 connections among them. Among these genes, 14 were reported associated with asthma by previous GWASs and 22 with nominally significant gene-wise P-values were novel candidates. The identified sub-network was significantly associated with childhood asthma ($P<10^{-4}$ using 10,000 CGP). Moreover, the number of connections (14) among known and novel candidate genes was significantly higher than expected by chance ($P=3\times10^{-4}$). Three KEGG pathways were found significantly enriched in genes from the identified network: cytokine-cytokine receptor interaction (Bonferroni-corrected $P=3\times10^{-8}$), chemokine signaling pathway (Bonferroni-corrected $P=5\times10^{-8}$), natural killer cell mediated cytotoxicity (Bonferroni-corrected $P=3\times10^{-6}$). This study shows the benefit of integrating GWAS data and HPIN to identify novel functionally related genes underlying childhood asthma. Funding: FP7-316861, ANR-11-BSV1-027, ANR-USPC-2013.

22

The Utility of Real World Data for Performing Genetic Target Validation: TRPV4 and Lung Edema. *D. Waterworth[1], L. Warren[2], M. Hurle[3], D. Behm[4], J. Pulley[5], E. Bowton[6], J. Denny[7,8], D. Sprecher[9], M. Ehm[10].* 1) Genetics, GlaxoSmithKline, King of Prussia, PA; 2) PAREXEL International (previously employed by GSK; work performed on this publication done while employed by GSK); 3) Computational Biology, GlaxoSmithKline, King of Prussia, PA; 4) Heart Failure Discovery Performance Unit, Metabolic Pathways & Cardiovascular Therapy Area Unit, GlaxoSmithKline, King of Prussia, PA; 5) Department of Medical Administration, School of Medicine, Vanderbilt University, Nashville, TN; 6) Institute for Clinical and Translational Research, School of Medicine, Vanderbilt University, Nashville, TN; 7) Department of Medicine, School of Medicine, Vanderbilt University, Nashville, TN; 8) Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN; 9) Metabolic Pathways and Cardiovascular Therapy Area Unit, GlaxoSmithKline, King of Prussia, PA; 10) Genetics, GlaxoSmithKline, Research Triangle Park, Durham, NC.

Transient receptor potential vanilloid 4 (TRPV4) is a Ca2+ permeable, nonselective cation channel that is thought to be involved in the regulation of systemic osmotic pressure. A TRPV4 blocker (GSK2798745) is in early clinical development within GlaxoSmithKline for pulmonary edema. The majority of mechanistic insight is derived from preclinical data; revealing influence on lung edema resulting from poor cardiac function. Therefore we set out to see if we could use human genetic data to validate this indication in humans and also provide insight into potential alternative indications. Rare mutations in TRPV4 result in a range of neuromuscular disorders and skeletal dysplasias (OMIM 605427), but more frequent variants have not thus far been robustly associated with any trait or disease within the GWAS literature. A search of the 1000 Genomes Project catalog identified two coding variants in the 0.5 to 2% frequency range, also present on the exome chip (V562I and E840K), that could potentially yield insights into drug effects. The E840K variant was also predicted to be functional by SIFT.A PheWAS study was performed using over 29,000 patients from BioVU, a hospital electronic health record (EHR) and biobank at Vanderbilt University. Over 1500 traits were defined using ICD9 codes for association analysis (multiple testing threshold was $p \leq 1.6e\text{-}5$). No associations with either variant were statistically significant, though there were twice as many $p<0.05$ associations for the E840K than the V562I. However, within the top 20 results for the E840K, the second most significant association was pulmonary edema and hypostasis (OR 1.97, p=4e-4) as well as five lung infection traits (OR range 1.7-3.1, $p<0.01$). The same variant was also nominally associated with renal failure as well as a cluster of menstruation-linked phenotypes. These results align well with TRPV4 expression, which is highest in the kidney and bronchial epithelia and has been reported within the endometrium. Replication of these results is planned and characterization of the variant is ongoing. Should these results be confirmed, they suggest that the E840K is a good tool variant for investigating TRPV4 in human disease and that renal failure and endometriosis have potential as alternative indications for the TRPV4 blocker. They also illustrate the value of real world data as the majority of these phenotypes are not available within cohort studies that constitute the majority of the GWAS literature.

23

Quantifying heritability explained in inflammatory bowel disease using 18,000 GWAS and 9,000 next generation sequencing data. *Y. Luo, K. de Lange, UK. IBD Genetics Consortium, C.A. Anderson, J.C. Barrett.* Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom.

Crohn's disease (CD) and ulcerative colitis (UC) are two main forms of inflammatory bowel disease (IBD). They both have been found to be highly heritable in twin studies (~60%) and have had substantial success in GWAS (201 loci in the most recent meta-analysis). Nonetheless, the total fraction of risk explained by common variants is only a fraction of the total heritability. In the recent study using genetic-relationship-matrices (GRM) estimated 26% of CD and 19% of UC risk to be captured by HapMap3 imputed data. However, whether or not rare variants with MAF <1% and of relatively high penetrance explain a large fraction of unexplained variance in IBD remains unanswered. Here, we try to address these questions by creating an imputation panel using one of the largest low-coverage whole genome sequencing projects of complex disease to date.We genotyped 8336 IBD cases and 9495 controls not part of any previous GWAS on the Human Core Exome platform (290,510 SNPs) and imputed them using a reference panel that consists of 4915 IBD samples at whole genome sequenced at 3x, 3910 control samples at 6x, and 2504 1000Genomes Phase 3 samples at 5x. 7 out of 19 million variants with MAF≥0.1% remain after stringent quality control (r2≥0.6 and missing rate<1%) post imputation to avoid the spurious estimation of genetic co-variance from genotype errors. We then applied joint variance component models with and without LD-adjusted GRM to dissect the genetic contribution to risk of CD and UC across various MAF spectrum. In total, we report 27% (SE 0.013) and 21% (SE 0.012) of variation in liability can be explained for CD and UC respectively. The total SNP-heritabilities estimated based on univariate analysis, MAF-bin partitioning analyses, with and without LD-adjusted approaches were consistent and similar to those from previously published studies, suggesting that our estimates are robust and reliable.Overall, the total amount of heritability explained did not substantially change (~1% increase) after introducing four million extra rare variants. This suggests that while common variants of individually small effect explain a significant proportion of heritabilty en masse, SNPs in the 0.1%-1% frequency range have less of an impact. However, from this study we cannot rule out the existence of truly rare variants (with moderate effect sizes) that are not imputed well even using the larger imputation panel.

24

The X-factor of complex disease: Methods, software, and extensive application for studying the X chromosome in association studies. *A. Keinan on behalf of the XWAS Consortium.* Biological Statistics & Computational Biology, Cornell University, Ithaca, NY.

The X chromosome plays an important role in human disease, especially those with sexually dimorphic characteristics. Analysis of X requires special attention due to its unique inheritance pattern leading to analytical complications that have resulted in the majority of GWAS either not considering or mishandling it with tools designed for non-sex chromosomes. We overcame many of the analytical complications by developing an array of X-specific methods that span all stages of GWAS, from genotype calling, through imputation and extensive QC, and to statistical association testing. Specifically, we developed four types of association tests for X-linked variants: (1) the standard test between a SNP and disease risk or quantitative trait, including after first stratifying individuals by sex, (2) a test for a differential effect of a SNP between males and females, (3) motivated by X-inactivation, a test for higher variance of a trait in heterozygous females as compared to homozygous females, and (4) for all tests, a version that allows combining evidence from all SNPs across a gene. We implemented the analysis pipeline and all methods as part of a publicly available software, XWAS (chromosome X-Wide Analysis toolSet). We applied these to conduct X-wide association studies in ~45 GWAS, with focus on autoimmune diseases and risk factors of coronary artery disease. We discovered and replicated many novel significant X-linked associations, e.g. (i) variants in *CENPI* as contributing, with different effect sizes in males and females, to the risk of three different autoimmune diseases, the risk of all of which is highly different between sexes. Other, autosomal genes in the same family as *CENPI* have previously been associated to other autoimmune diseases; (ii) *ARHGEF6* to Crohn's disease, and replicated in ulcerative colitis, another inflammatory bowel disorder. ARHGEF6 has been shown to interact with a gastric bacterium that has been associated to IBD. (iii) Significantly increased variance of systolic blood pressure in females that are heterozygous for a variant that might regulate *ATRX*, a gene that has been previously associated with alpha-thalassemia. We also showed that several previously reported associations are false positives due to ignoring the unique nature of X. In conclusion, XWAS will provide the tools for many to incorporate the X chromosome into GWAS, enabling discoveries of novel loci implicated in many diseases and in their sexual dimorphism.

25

Imputation of low-frequency variants identifies novel Alzheimer's disease loci in the IGAP Consortium. *K. Hamilton-Nelson[1], B. Grenier-Boley[2,3,4], B.W. Kunkle[1], M. Vronskaya[5], V. Chouraki[6], M. Butkiewics[7], S.V. Van der Lee[8], R. Sims[5], A.M. Toglehofer[9], J. Jakobsdottir[10], B. Dombroski[11], O. Valladeres[11], J. Bis[12], E.R. Martin[1], R. Mayeux[13,14], L.A. Farrer[15,16,17,18,19], C. Duijn[8], J.L. Haines[7], P. Holmans[20], J.C. Lambert[2,3,4], J. Williams[5], S. Seshadri[19,21], P. Amouyel[2,3,4], G.D. Schellenberg[11], M. Pericak-Vance[1] For The IGAP Consortium.* 1) Human Genomics, University of Miami Miller School of Medicine, Miami, Fl., USA; 2) Institut Pasteur de Lille, Lille, France; 3) INSERM, U744, Lille, France; 4) Universite Lille 2, Lille, France; 5) Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University, Cardiff, UK; 6) Boston University School of Medicine, Boston, MA, USA; 7) Department of Epidemiology & Biostatistics, Case Western Reserve University, Cleveland, OH, USA; 8) Deparment of Epidemiolgy, Erasmus Medical Centre, Rotterdam, NL; 9) Institute of Molecular Biology and Biochemistry, Centre for Molecular Medicine, Medical University of Graz, Graz, AT; 10) Icelandic Heart Association, Kopavogur 201, IS; 11) Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; 12) Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA; 13) Taub Institute on Alzheimer's Disease and the Aging Brain, Department of Neurology, Columbia University New York, NY, uSA; 14) Gertrude H. Sergievsky Center, Department of Neurology, Columbia University, New York, NY, USA; 15) Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; 16) Department of Medicine (Biomedical Genetics), Boston University School of Medicine, Boston, MA, USA; 17) Department of Ophthalmology, Boston University School of Medicine, Boston, MA, USA; 18) Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA; 19) Department of Neurology, Boston University School of Medicine, Boston, MA, USA; 20) Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University, Cardiff, UK; 21) The Framingham Heart Study, Framingham, MA, USA.

In 2013 the International Genomics of Alzheimer's Project (IGAP) published the largest Alzheimer disease (AD) Genome-wide Association Studies (GWAS) to date. This analysis identified 19 susceptibility loci, in addition to APOE, for Late-onset Alzheimer disease (LOAD). Following up these analyses, IGAP conducted a GWAS using updated 1000 genomes imputation of 38 datasets (including 17 new datasets), increasing our discovery sample to 21,433 cases and 44,340 controls. All datasets were imputed to a 1000 Genomes reference panel (Phase 1 v3, March 2012) of over 37 million variants, many of which are low-frequency (MAF<=5%) single nucleotide variants (SNV) and indels. Single-variant-based association analysis was conducted adjusting for age, sex and population substructure. Individual datasets were analyzed with the score test for case-control datasets and general estimating equations (with generalized linear mixed model for rare variants) for family-based analyses. Within-study results were meta-analyzed in METAL. Gene-based testing was conducted on summary statistics using VEGAS. Pathway analyses were performed with ALIGATOR and PARIS. Twenty-two loci were genome-wide significant at P≤5×10-8, including previously reported rare and low-frequency variants in *TREM2* and *SORL1* and two novel associations of common intergenic variants between the genes *USP6NL* and *ECHDC3* at Chr10: 10:11720308 (P=2.91x10^{-9}) and the genes *CYYR1* and *ADAMTS1* at Chr21: 28,156,856 (P=1.44x10^{-8}). Low-frequency SNVs in the common loci BIN1 (MAF=0.026) and CLU (MAF=0.029) show suggestive significance (P≤5×10-7), while twelve additional loci produced signals with suggestive significance, seven driven by low-frequency or rare variants and five driven by common variants. Genotyping to confirm imputation quality, and replication genotyping using the Sequenom MassArray are underway. Gene-based analyses identified 13 significantly associated genes (Bonferroni P≤2.83x10-6), four of which are novel loci driven by nominally significant low-frequency variants. Pathway analyses confirmed previously associated enrichment of immune-related, endocytocytic, and lipid pathways. Using updated imputation and an increased sample size we identified novel candidate loci for LOAD, including several low-frequency variant associations.

26

A new locus of genetic resistance to severe malaria is associated with a locus of ancient balancing selection. *G. Band on behalf of the MalariaGEN consortium.* Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN UK.

We describe a genome-wide association study of severe malaria susceptibility using DNA from over 10,000 individuals from across sub-Saharan Africa with replication in a further 15,000. We identify a new locus of association near the glycophorin gene cluster on chromosome 4, which encodes red cell surface proteins previously shown to interact with malaria parasite surface receptors during invasion, and determines the MNS blood group. A single haplotype at this locus, common in parts of East Africa, confers 33% protection against severe malaria, and is linked to variation displaying signatures of ancient balancing selection. We describe attempts to elucidate the possible causal mutations, including imputation into an African-enriched reference panel and the refinement and imputation of large structural variants in the region. This association brings the number of loci confirmed by GWAS to be associated with severe malaria to four, all of which are involved in red blood cell function or morphology, and at least three of which display unambiguous signals of balancing selection. These analyses bring important new insights into malaria biology and may have implications for genome-wide association studies of infectious diseases more generally.

27

Long read single-molecule real-time (SMRT) full gene sequencing of cytochrome P450 2D6 (*CYP2D6*). *Y. Yang[1], W. Qiao[1], R. Sebra[1,2], G. Mendiratta[1], A. Gaedigk[3,4], R. Desnick[1], S. Scott[1].* 1) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 3) Division of Clinical Pharmacology, Toxicology & Therapeutic Innovation, University of Missouri-Kansas City, Kansas City, MO 64108, USA; 4) School of Medicine, University of Missouri-Kansas City, Kansas City, MO, USA.

The CYP2D6 enzyme metabolizes ~25% of common medications, yet homologous pseudogenes and copy number variation make interrogating the polymorphic *CYP2D6* gene with short-read sequencing challenging. Moreover, accurate prediction of *CYP2D6* metabolizer status necessitates direct analysis of the duplicated copy when an increased copy number is detected, particularly when identified concurrently with normal activity and loss-of-function alleles in compound heterozygosity. Given the importance and polymorphic nature of *CYP2D6* and the paucity of available *CYP2D6* next-generation sequencing assays, we developed a novel long-read, full gene *CYP2D6* single-molecule real-time (SMRT) sequencing method using the Pacific Biosciences platform. Long-range PCR and *CYP2D6* SMRT sequencing of 10 previously genotyped positive control samples identified expected star (*) alleles, but also enabled suballele resolution, diplotype refinement, and discovery of the novel *CYP2D6*107* allele. Coupled with an optimized variant calling pipeline and a novel long-read sequencing error correction script, *CYP2D6* SMRT sequencing was highly reproducible as triplicate intra- and inter-run testing resulted in completely concordant non-reference genotype calls. Importantly, qPCR coupled with targeted SMRT sequencing of upstream and downstream *CYP2D6* copies using specific long-range PCR primers characterized the duplicated gene copy in 15 *CYP2D6* copy number variant controls. The utility of *CYP2D6* SMRT sequencing was further underscored by testing 14 samples with discordant or unclear *CYP2D6* configurations from previous targeted genotyping, which resulted in suballele resolution, genotype refinement, duplicated allele characterization, and discovery of a novel tandem arrangement, *CYP2D6*36-*41*. Taken together, these data indicate that long-read *CYP2D6* SMRT sequencing is an innovative, reproducible, and validated method for full gene characterization, duplication-specific analysis and novel allele discovery, which will likely improve *CYP2D6* metabolizer phenotype prediction for both research and clinical testing applications.

28

An NGS-based Carrier Screen for Congenital Adrenal Hyperplasia with 95% Detection Rate. *D. Muzzey, M.R. Theilmann, K.M. D'Auria, H.H. Lai, C.S. Chu, I.S. Haque, E.A. Evans, H.P. Kang, J.R. Maguire.* Counsyl, South San Francisco, CA.

INTRODUCTION: There are two strong arguments for including congenital adrenal hyperplasia (CAH) on expanded carrier screen (ECS) panels: (1) with a carrier rate of 1 in 60, it is one of the ten most common recessive diseases, and (2) screening only ten variants in *CYP21A2* affords a 95% detection rate. However, CAH is absent from most ECS panels—or at best detected at a rate <20%—because 98% homology between the coding sequences of *CYP21A2* and its pseudogene *CYP21A1P* complicates variant identification. Indeed, seven of the ten common variants are pseudogene-derived; thus, expensive, low-throughput assays (e.g., long-range PCR + Sanger sequencing) are typically used to assess CAH carrier status. Here we report the development of an NGS-based carrier screen for CAH that detects all ten common deleterious variants in *CYP21A2*.

METHODS: Hybrid-capture probes were designed to anneal adjacent to bases that differ between *CYP21A2* and *CYP21A1P* ("diff bases"). Paired-end NGS of captured fragments allows designation of reads as being either gene- or pseudogene-derived based on diff bases. CAH variants were identified using two strategies: SNP-based calling and copy-number analysis. SNP-based calling searched for deleterious bases in a pileup composed of reads with gene-derived diff bases distal from the position of interest. By contrast, copy-number analysis used read depth of diff bases to calculate the relative abundance of each variant, and deleterious variants were identified as those with excess copy number of pseudogene-derived sequence (and, conversely, depleted copy number of gene-derived sequence). Long-range PCR and Sanger sequencing were used to confirm variants in a validation study.

RESULTS: We have run the validated CAH test on nearly 150,000 clinical samples, with variant frequencies comparable to the literature. Gene and pseudogene copy number varies greatly: 38% of patients have at least one haplotype that does not simply have one copy of each. Recombination is widespread, with at least 83% having a *CYP21A2* haplotype containing pseudogene-derived bases. Finally, the test identifies compound variants consistent with specific rare haplotypes, e.g., (1) three copies of *CYP21A2* where one has the Q319X mutation, and (2) *CYP21A2* with a V282L mutation in cis with two copies of *CYP21A1P*, enriched in Ashkenazi Jewish patients.

CONCLUSIONS: With the appropriate probe design and analysis suite, CAH can be assessed with high detection rate on NGS-based ECS panels.

29

Supplemental CNV analysis in NGS genepanel data in a diagnostic setting. *J.J. Saris, R. van Minkelen, M.A. van Slegtenhorst, H.T. Brüggenwirth, L.H. Hoefsloot.* Clinical Genetics, Erasmus MC, Rotterdam, Netherlands.

Standard diagnostic laboratory flows based on PCR and Sanger sequencing often include MLPA analysis when available and relevant. Introducing NGS genepanels in a clinical setting promises a higher diagnostic yield due to scanning more genes, while reducing costs but is usually designed only for sequence variant detection. Normalized Depth-of-Coverage (DOC) calculation using NGS data can indicate exon deletion or duplication events, i.e. copy number variation (CNV) and might replace MLPA. Our lab has established a NGS sequencing and reporting flow using the SeqNext software (www.jsi-medisys.de). Recently we validated the effectiveness and usability of a DOC-CNV submodule using data from genes located on chromosome X. Type and number of reference files (1-on-1 and n=12), signal-to-noise ratio and detection cut-off were evaluated in SeqNext and, separately, in MS Excel via Z-score analysis.Aim: Retrospective analysis of twelve runs of genepanel experiments (144 samples, mainly cardiomyopathy, CM), comparison to validation results and prospective analysis of future experiments (±6) for CNV on exon-ROI level and wetlab confirmation of calls ≥ 2 exons and <5% frequency. Creating an internal database of frequent occurring CNV events and "noisy" ROI. Results: In 120 samples (10 experiments) DOC based gender determination using chromosome X located genes, calling was concordant with the recorded clinical gender. No CNV ≥ 2 exon were detected, except in a region within TTN, which appears a noise-region due to mapping or capturing artefacts. Graphical interpretation of single exon CNV did indicate the presence of intronic polymorphic InDels influencing DOC results on capturing and/or mapping level of analysis, e.g. 5' ANKRD1-exon4 (MAF ±30%).Conclusion: CNV detection using DOC-analysis module in SeqNext promises to be an extension and replacement of MLPA in diagnostics.Validation, update on current results and lessons learned will be presented.

30

The application of the CNVSeq method for whole genome copy number variant detection. *A. Tzika[1], P. Roberts[1], S. Hewitt[1], C. Watson[2], L. Crinnion[2], D. Bonthron[2].* 1) Cytogenetics Laboratory , St James's Hospital, Leeds, United Kingdom; 2) Leeds Institute of Molecular Medicine, St James's Hospital, Leeds, United Kingdom.

We present a Next Generation Sequencing (NGS) based approach to detect copy number variation referred to as the CNVSeq methodology. The CNVSeq method has been incorporated successfully in the diagnostic repertoire of the laboratory since January 2014. The Leeds Genetics Department at St James's Hospital has been the first and so far the only provider of such a service in the UK. The technique uses the Illumina HiSeq 2500 Sequencing platform for whole genome copy number variant detection and relies on custom-designed Python scripts for copy number variant calling. Since the introduction of this service we have sequenced over 125 diagnostic samples from difference sources, i.e. blood, tissue from products of conception, amniotic fluid, saliva. As a current practice we use CNVSeq for samples that are of too poor/low quality to be tested by array-CGH. Among this cohort of patients, we have managed to provide a diagnosis by CNVSeq using precious, non-repeatable samples such as samples from deceased patients or tissue samples from miscarriages.The average resolution we achieve per patient is 40Kb and the analytical sensitivity for imbalances ≥40Kb has been estimated to be 98%. Due to the nature of the test a uniform and not targeted resolution across the genome is obtained. To deliver this resolution 20-25 million reads are required and this is achieved by running 10 samples in the multiplex with 100 cycles of single read sequencing. Higher resolution can be achieved by sequencing the libraries at higher coverage with relatively minimal cost implications.In summary the key drivers for implementation of this technique have been to utilise the laboratory's NGS facilities in order to improve success rates on low volume/poor quality DNA samples, and to assess the potential benefit of a more digital approach to evaluating DNA dosage changes. We believe that our method is cost effective, reliable and accurate and that it potentially overcomes the technical challenges of calling dosage changes by whole exome sequencing. Our experience with the technique so far has provided a wide variety of example cases that highlight the benefits of using an NGS based method for copy number variant detection.

31

Detection of relevant pharmacogenomic variants and CYP2D6 copy number using a highly multiplexed next generation sequencing assay. *F.C.L. Hyland[1], M. Manivannan[1], S.C. Chen[1], T. Chen[2], T. Hartshorne[1], M. Anderson[2], G. Liu[1].* 1) Thermo Fisher Scientific, South San Francisco, CA; 2) Thermo Fisher Scientific, Carlsbad, CA.

PurposeCytochrome P450 enzymes metabolize about 75% of drugs, with UGT enzymes metabolizing about another 15%. Variations in gene sequence or in copy number may result in an inactive, defective, unstable, mis-spliced, low expressed, or absent enzyme, an increase in enzyme activity, or an altered affinity for substrates. Pharmacogenomic genes may predict whether an individual is a poor or rapid metabolizer, facilitating dose optimization. Failure to adjust dosage of drugs metabolized by the relevant enzyme can lead to adverse drug reaction, or conversely to too rapid drug metabolism and no drug response. **Materials and Methods**We designed a highly multiplexed pharmacogenomics (PGx) research panel to profile a set of known 137 markers and CYP2D6 copy number variation in a single amplification reaction using Ion Torrent semiconductor sequencing. The panel also includes specially designed sample ID primer pairs for sample discrimination and gender determination. High quality sequencing libraries were produced from as little as 10 ng of input DNA from archived buccal swab and cell lines. **Results and Conclusions**Technical verification of the Ion AmpliSeq pharmacogenomics research panel was established by sequencing 91 well characterized and annotated cell lines from Coriell. Following optimization on cell lines, we sequenced hundreds of buccal swab samples from 5 different labs, using both manual and highly automated protocols. We multiplexed between 8 and 96 samples per chip. We were able to achieve high uniformity of sequencing depth of the targets. To measure genotyping accuracy, we compared the AmpliSeq panel genotypes to annotated genotypes of the samples, and to genotypes generated from TaqMan OpenArray PGx assays. The study showed genotype concordance >99.8%; genotype reproducibility > 99.8%, and no-call genotype rate < 0.2%. We measured the accuracy of the CYP2D6 gene copy number calls. Externally annotated copy number and PCR copy number were compared with our calls. The accuracy of the gene-level copy number calls was > 98%, with a no-call rate below 2%. The research panel facilitates accurate genotyping and copy number determination of key pharmacogenomic variants. This method is customizable and custom panels containing additional pharmacogenomics markers may be designed so that additional markers can be profiled in the same multiplex reaction. *For Research Use Only. Not for use in diagnostic procedures.*.

32

Next-generation sequencing-based genetic testing of autosomal dominant polycystic kidney disease. *P.C. Wu[1], Y.H Yang[2], T.W. Kao[3], J.W. Huang[3], T.S. Chu[3], P.L. Chen[2,4,5,6].* 1) Graduate Institute of Molecular Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan; 2) Department of Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan; 3) Division of Nephrology, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan; 4) Graduate Institute of Medical Genomics and Proteomics, College of Medicine, National Taiwan University, Taipei, Taiwan; 5) Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan; 6) Division of Endocrinology and Metabolism, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan.

Autosomal dominant polycystic kidney disease (ADPKD) is the most common inherited kidney disorder affecting 1 in 400 to 1 in 1000 live birth worldwide. The disease is late onset and often diagnosed through renal imaging of cysts formation in the kidney in the mid to late adulthood. Approximately 50% of the patients develops end-stage renal disease and eventually require dialysis or renal transplantation. Molecular diagnosis can be a great help to assess potential related donor and also identify at risk family members at young. The only two known causative genes of ADPKD are *PKD1* (16p13.3) and *PKD2* (4q21). One major obstacle for genetic diagnosis of ADPKD is that *PKD1* has six pseudogenes located on the same chromosome that shares 97% sequence identity. Long-range PCR (LRPCR) has been used to deal with the pseudogene issue with only partial success. Another major obstacle is the large number of exons (46 and 15 exons for *PKD1* and *PKD2*, respectively), which makes the traditional Sanger sequencing-based method quite expensive and labor intensive. We consider next-generation sequencing (NGS) to be an appropriate solution thanks to its high throughput, low cost and single-strand sequencing. In this study, we incorporated a capture-based targeted enrichment method coupled with supplementary LRPCR of *PKD1* and *PKD2* genes. The samples were then sequenced with Illumina Miseq. Bioinformatics pipeline includes BWA, SAMtools, Picard, GATK and ANNOVAR. All variants were then checked with PKD Database (http://pkdb.mayo.edu) for its significance. Novel variants were assessed with SIFT and PolyPhen2 for pathogenicity. All possible pathogenic variants were confirmed by Sanger sequencing. Among the 50 families tested, we found 23 probands with definitely pathogenic variants (46%) due to premature stop codon (n=18), frameshift indel (n=3) or splicing site variant (n=2). Five probands with likely or highly likely pathogenic variants (10%) listed on PKDB. Noted that three probands with definitely pathogenic variants also carries a likely pathogenic variant. A founder effect was observed in 11 probands with the same definitely pathogenic mutation (*PKD2* c.2407C>T, p.R803X, NM_000297). Our method provides a reliable, sensitive, fast and inexpensive genetic diagnostic platform for ADPKD. .

33

A method for detecting intragenic copy number changes of *BRCA1* and *BRCA2* using next-generation sequencing data. *P.J.B. Sabatini, K. Chun.* North York General Hospital, Toronto, ON, Canada.

Comparing the number of next-generation sequencing fragments can detect copy number changes, although the accuracy of this analysis is not well reported. Using a large set of normal controls and positive samples we have developed a screening algorithm to detect copy number changes as low as 150 bp in the *BRCA1* and *BRCA2* genes. The TruSight cancer panel was used to capture the *BRCA1* and *BRCA2* coding regions and then sequenced with the NextSeq 500 instrument (Illumina, CA). NextGene analytical software (SoftGenetics) was used for the alignment, variant calling and the copy number analysis. The reads per kilobase per million mapped reads (RPKM) counting method and a Hidden Markov model was used to detect both duplications and deletions. A tiled bed file of 100 bp fragments overlapping by 50 bp across the coding regions of the *BRCA1* and *BRCA2* genes was set for the RPKM analysis. A total of 98 normal, 4 deletions and 6 duplications were analyzed. To limit false positives, a dispersion value of 0.003 for deletions and 0.005 for duplications were used. With these values, all deletions and duplications were detected with a false positive rate of 13%. We are now in the process of prospectively analyzing up to 700 clinical samples using these parameters to detect copy number changes. These results describe a screening method to detect CNVs within the *BRCA1* and *BRCA2* genes with 100% sensitivity.

34

DNA Combing as a First-Tier Genetic Testing for Facioscapulohumeral Dystrophy Type 1: A Cohort of 155 Patients. *J. Wang, F.Z. Boyar, X.J. Yang, B.H. Nguyen, V. Sulcova, P. Chan, Y. Liu, A. Anguiano, C.M. Strom.* Cytogenetics Laboratory, Nichols Institute, Quest Diagnostics, San Juan Capistrano, CA.

Facioscapulohumeral dystrophy (FSHD) is the third most common muscular dystrophy. Type-1 FSHD is due to a contraction of the D4Z4 macrosatellite repeat motif on the 4qA allele. Shortening of the 4qB, 10qA, or 10qB allele is not associated with FSHD. The southern blot (SB) assay for FSHD has limitations for detecting mosaicism, borderline contractions, and rearrangements. DNA combing analysis hybridizes multi-color DNA probes onto uniformly stretched DNA fiber and accurately identifies the 4qA and other alleles. Direct visualization of the DNA fibers provides information on non-permissive 4qB, 10qA, and 10qB alleles, which is not available by SB; it also allows detection of mosaicism, sub-classification of cases by the contraction size, and detection of complex rearrangements. Here, we report our experience using DNA combing as a genetic test for type-1 FSHD. In this cohort of 155 patients, 147 (95%) had enough DNA fibers for analysis (including 2 cases with mosaicism). The average number of DNA fibers obtained for each patient was 54 (~13 for each allele). A total of 62 cases (positive rate = 42%) had either abnormal (51 cases, ≤8 repeats at 4qA) or borderline results (11 cases, 9 to 11 repeats). Our analysis also revealed a possible association between the number of repeats and age at referral. For abnormal cases with 2 to 4 repeats at 4qA, the average age was 29 years old when referred to us for analysis. In contrast, for cases with 5 to 8 repeats at 4qA, the average age was 53 years of age ($p < 0.0001$, Student's t-Test). Among the 85 cases with normal results, 35 (41%) had either a contracted 4qB (7 cases) or shortened 10qA allele (28 cases). Based on the results of this cohort, we recommend that DNA combing be offered as first-tier genetic testing for FSHD because of its ability to identify the repeat number of the 4qA allele and detect type-1 FSHD rearrangements and mosaicism. Furthermore, the precise measurement of the D4Z4 repeat motif by DNA combing may help correlate the size of the contracted 4qA allele with the timing of type-1 FSHD onset.

35

Pitfalls in development of statistical methods for rare variant association studies. *GT. Wang[1,2], D. Zhang[1], Z. He[1], D. Hang[1], B. Li[1], SM. Leal[1].* 1) Center for Statistical Genetics, Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX; 2) Department of Human Genetics, The University of Chicago, Chicago, IL.

Statistical analyses for rare variant association are typically conducted by grouping all rare variants in given genomic region, usually a gene, into a unit and test for difference in features of the unit between individuals. Such group-based strategy increase statistical power compared to analyzing individual rare variants. To date, numerous variations on group-based rare variant association methods have been published. Rigorous assessment of these methods in terms of type I and type II errors on high quality empirical benchmark and real world data is crucial before implementation in practice; ignorance of this point in many method papers have resulted in biased conclusions. Here we caution that development of rare variant association methods is subject to several potential pitfalls, including misuse of empirical genotype data source for power analysis, underrepresentation of genetic architecture in simulation models at whole genome level, and inappropriate modelling of protective variants. We demonstrate that sample size estimation for rare variant association study designs can differ drastically due to difference in rare variant minor allele frequency (MAF) spectrum resulted from simulations of genomic sequences under various demographic models and from resampling of real world sequence data, and the use of MAF from exome variant databases for genotype simulation is flawed. We show that relative power profile of association methods is heavily influenced by genetic context assessed other than merely phenotypic models, a crucial point underappreciated in most method papers. We also demonstrate that the impact of protective variants, if properly modelled, is minor compare to the impact of deleterious variants and thus burden tests should be preferred over variant component tests in practice. Our results highlights the importance to perform sample size estimation and evaluation of rare variant methods only using simulated data with sufficiently large samples whose properties are validated to ensure a close resemblance of real world data, and such analysis should be performed and truthfully reported at whole genome scale rather than genomic regions of choice. Our experiments have produced simulated sequence data freely available to public which will facilitate future study design and method development for rare variant association studies.

36

Ignoring pleiotropy bias in Mendelian randomization with polygene scores can easily lead to incorrect inferences about causality. *C.M. Astley[1, 2, 3], M. Kals[4,5], T. Esko[4,2], J.N. Hirshhorn[1,2,3], K. Fischer[4].* 1) Boston Children's Hospital, Boston, MA; 2) Broad Institute, Cambridge, MD; 3) Harvard Medical School, Boston, MA; 4) Estonian Genome Center, University of Tartu, Estonia; 5) Institute of Mathematical Statistics, University of Tartu, Estonia.

Establishing the causes of disease is essential for designing effective therapies. Mendelian randomization is a popular analytical tool to estimate causal effects, even in the presence of unmeasured confounding that can preclude causal inferences. In Mendelian randomization, genetic variants are used as instruments to assess whether an observed exposure (e.g. LDL-cholesterol) causes an outcome (e.g. myocardial infarction). The validity rests on key assumptions: the genetic variant must be a strong instrument, explaining an appreciable fraction of the variance in the exposure, and must not have pleiotropic effects on the outcome independent of the exposure. Violating the first assumption, by using weak variant as an instrument, can lead to inflated effect estimates. Novel methods for combining multiple weak alleles into a stronger polygenic instrument may circumvent weak instrument bias, but may also increase the chance of introducing pleiotropy bias if alleles have pleiotropic effects. As genome wide association studies enable stronger polygenic instruments by combining numerous alleles, many with unknown biological function, quantifying pleiotropy bias becomes crucial. To quantify the scope of the problem, we built a robust simulation model to examine the parameter space in which pleiotropy bias is substantial for polygenic allele score-based Mendelian randomization. The data-generating model included two polygenic traits (exposure and outcome) confounded by an unmeasured variable. We adjusted the proportion of pleiotropic alleles, the relative magnitude of an allele's effect on one trait compared to another, as well as other biologically relevant parameters. We found that pleiotropy bias can severely inflate effect estimates, even when the proportion of pleiotropic alleles is low (e.g. 10%), when the relative magnitude of pleiotropy is low (e.g. 1%), or when there is no true causal effect of the exposure on the outcome. We show analytically that pleiotropy bias is the ratio of instrument-outcome and instrument-exposure causal effects. Realistic biological scenarios demonstrate that pleiotropy bias is likely to be in the same direction as the confounded crude association measure, further obscuring interpretation of Mendelian randomization results. We present methods that attempt to distinguish pleiotropic from suitable instruments or that attempt to correct for pleiotropy bias, but these methods require additional assumptions about the underlying biology.

37

Mendelian randomization provides evidence for a causal effect of low vitamin D on multiple sclerosis risk: Results from the Kaiser Permanente MS Research Program. *B. Rhead[1], M. Gianfrancesco[1], A. Mok[1], X. Shao[1], H. Quach[1], L. Shen[2], A. Bernstein[3], C. Schaefer[2,4], L.F. Barcellos[1,2].* 1) Genetic Epidemiology and Genomics Laboratory, UC Berkeley, Berkeley, CA; 2) Kaiser Permanente Division of Research, Oakland, CA; 3) Palm Drive Hospital, Sebastopol, CA; 4) Research Program on Genes, Environment and Health, Kaiser Permanente, Oakland, CA.

Low serum levels of 25-hydroxyvitamin D (25[OH]D) are associated with a higher risk of multiple sclerosis (MS [MIM 126200]) and with greater MS activity and disease progression. However, a causal relationship between 25[OH]D and MS has not been firmly established: it is unclear whether low 25[OH]D levels are a cause or a consequence of MS. We conducted a Mendelian randomization analysis using three single nucleotide polymorphisms (SNPs) found to be associated with serum 25[OH]D level in a genome-wide association study (Ahn et al., 2010) to estimate the causal effect of low 25[OH]D on MS susceptibility in White, non-Hispanic members of Kaiser Permanente Northern California (1,200 MS cases, 10,000 controls). Participants were genotyped on Affymetrix or Illumina arrays, and additional variants were imputed using IMPUTE2 and the 1000 Genomes reference panel. We constructed the instrumental variable (IV) by computing a weighted genetic risk score for low 25[OH]D level using the estimated effect of the three published risk variants: rs2282679-C, in an intron of *GC*; rs2060793-G, upstream of *CYP2R1*; and rs3829251-A, in an intron of *NADSYN1*. We analyzed the effect of the IV on MS susceptibility in a logistic regression model that controlled for sex, year of birth, smoking, education, ancestry, self-reported body mass index at age 18-20, a weighted genetic risk score for 110 known MS-associated variants, and the presence of one or more *HLA-DRB1*15:01* alleles, the strongest genetic risk factor for MS. Results showed that a higher 25[OH]D genetic risk score is associated with an increased risk of MS, with a causal odds ratio of 1.29 (p=0.02, 95% CI: 1.04-1.61). A stronger effect was observed when analysis was restricted to those who did not carry an *HLA-DRB1*15:01* allele, with a causal odds ratio of 1.49 (p=0.01, 95% CI: 1.09-2.03). This multivariable analysis provides evidence that low serum 25[OH]D concentrations are a cause, rather than a result, of MS, independent of established risk factors. Furthermore, these odds ratios provide unconfounded estimates of the causal effect of low serum 25[OH]D on MS susceptibility.

38

Contrasting regional architectures of schizophrenia and other complex diseases using fast variance components analysis. *P. Loh[1,2], G. Bhatia[1,2], A. Gusev[1,2], H.K. Finucane[3], B.K. Bulik-Sullivan[2,4], S.J. Pollack[1,2,5], T.R. de Candia[6], S.H. Lee[7], N.R. Wray[7], K.S. Kendler[8], M.C. O'Donovan[9], B.M. Neale[2,4], N. Patterson[2], A.L. Price[1,2,5], Schizophrenia Working Group of the Psychiatric Genomics Consortium.* 1) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 3) Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA; 4) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 5) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; 6) Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO; 7) The Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia; 8) Department of Psychiatry and Human Genetics, Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA; 9) MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK.

Heritability analyses of GWAS cohorts have yielded important insights into complex disease architecture, and increasing sample sizes hold the promise of further discoveries. Here, we analyze the genetic architecture of schizophrenia in 49,806 samples from the Psychiatric Genomics Consortium, and nine complex diseases in 54,734 samples from the GERA cohort. For schizophrenia, we infer an overwhelmingly polygenic disease architecture in which ≥71% of 1Mb genomic regions harbor at least one variant influencing schizophrenia risk. We also observe significant enrichment of heritability in GC-rich regions and in higher-frequency SNPs for both schizophrenia and GERA diseases. In bivariate analyses, we observe significant genetic correlations (ranging from 0.18 to 0.85) among several pairs of GERA diseases; genetic correlations were on average 1.3x stronger than correlations of overall disease liabilities. To accomplish these analyses, we developed a fast algorithm, BOLT-REML, for multi-component, multi-trait variance components analysis that overcomes prior computational barriers that made such analyses intractable at this scale. The overall framework of the BOLT-REML algorithm is Monte Carlo AI REML, a Newton-type iterative optimization of the (restricted) log likelihood with respect to the variance parameters sought. BOLT-REML begins a multi-variance component analysis by computing an initial estimate of each parameter using the single variance component estimation procedure of BOLT-LMM (Loh et al. 2015 Nat Genet), which is the only analysis possible with BOLT-LMM. Then, in each iteration, BOLT-REML rapidly approximates the gradient of the log likelihood using pseudorandom Monte Carlo sampling and the Hessian of the log likelihood using the average information matrix. The approximate gradient and Hessian produce a local quadratic model of the likelihood surface, which we optimize within an adaptive trust region radius—key to achieving robust convergence—to update the variance parameter estimates. These procedures allow BOLT-REML to consistently achieve convergence in $\approx O(MN1.5)$ time; in contrast, existing multi-component REML algorithms are less robust and/or require $O(MN2+N3)$ time (e.g., GCTA). For example, a six-variance component analysis of N=50K GERA samples typed at M=600K SNPs that would have required ≈150 CPU hours and ≈200 GB RAM using GCTA required only 16 CPU hours and 7 GB RAM using BOLT-REML.

39

Multivariate analysis of whole exome sequence data identifies rare variants with pleiotropic effects on obesity-related metabolic traits in 31,000 participants of the Regeneron Genetics Center – Geisinger MyCode collaborative project – DiscovEHR. *S. Mukherjee[1], C. O'Dushlaine[1], C.V. Hout[1], S. Bruse[1], J.B. Leader[2], D.N. Hartzel[2], J. Staples[1], S. Hamon[1], J. Overton[1], J.G. Reid[1], A. Baras[1], D.J. Carey[2], H.L. Kirchner[2], M.D. Ritchie[2], S.A Pendergrass[2], M. Murray[2], D.H. Ledbetter[2], O. Gottesman[1], F. Dewey[1], A.R Shuldiner[1].* 1) Regeneron Genetics Center, Regeneron Pharmaceuticals, Inc., Tarrytown, NY; 2) Geisinger Health System, Danville, PA.

Identification of genes linking clinical phenotypes that occur together in pathophysiological states has great potential to unveil novel mechanistic insights into human disease. However, systematic detection of these genes is challenging. To identify genes exerting pleiotropic effects on obesity-related metabolic traits, we implemented a unified framework for multivariate test statistics using whole exome sequence data from 31,000 participants in the Regeneron Genetics Center (RGC) – Geisinger MyCode collaborative project – DiscovEHR. We used canonical correlation analysis, a multivariate generalization of the Pearson product-moment correlation, to jointly measure the association between genotypes and metabolic traits. The seven traits used in joint testing, extracted from the electronic health record(EHR), were median lifetime low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglyceride (TRIG), body mass index, fasting glucose, systolic blood pressure (SBP) and diastolic blood pressure. Whole exome sequencing (Illumina HiSeq) was performed at the RGC; correlations among ~350,000 exonic variants passing QC and all seven traits were computed in plink multivariate package (MQFAM). A total of 45 exonic variants showing evidence of pleioropic effects reached exome-wide significance ($P < 1.4E-7$) in joint testing of the seven traits. Top signal at *TOMM40* (P=1.92E-54) showed opposite effects for LDL and TRIG. *PCSK9* (rs11591147, P=1.96E-14) was associated with LDL and SBP in opposite directions. Twenty-five other variants within *CETP, MLXIPL, GCKR, APOB, APOE, FADS1, BCAM, TM6SF2, LPL* showed pleiotropic effects linking glycemic and lipid traits. An additional trivariate analysis focused on lipid traits only was performed, and 86 exonic variants reached exome-wide significance, 23 of which demonstrated pleiotropic effect for all three lipid traits with reversed allelic effect in HDL and TRIG. We also found improved statistical power in the detection of rare variants in *APOE* (rs5742904, P=2.954E-18, MAF=0.06%), *ANGPTL3* (chr1: 63063592, P=8.54E-08, MAF=0.1%) and *APOC3* (rs138326449, P= 1.399E-46, MAF=2% and rs76353203, P=1.14E-12, MAF=0.06%). Overall, we demonstrated that the multivariate approach provides additional information for the identification of key variants with pleiotropic effects and pathways that drive the correlated architecture of phenotypes comprising metabolic syndrome, with potential insights into points of intervention.

40

Quantifying penetrance in a dominant disease gene using large population control cohorts. *E.V. Minikel[1,2,3,4], S.M. Vallabh[1,2,4], M. Lek[2,3], K.O. Estrada[2,3], K.E. Samocha[2,3,4], J.F. Sathirapongsasuti[5], C.Y. McLean[5], J.Y. Tung[5], L.P.C. Yu[5], M.J. Daly[2,3], D.G. MacArthur[2,3], Exome Aggregation Consortium.* 1) Prion Alliance, Cambridge, MA 02139, United States; 2) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, United States; 3) Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, United States; 4) Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, United States; 5) Research, 23andMe Inc., Mountain View, CA 94041, United States.

More than 100,000 genetic variants are reported to cause Mendelian disease in humans, but the penetrance - the probability that a carrier of the purported disease-causing genotype will indeed develop the disease - of the vast majority of these variants is unknown. The development of large-scale genotyping and sequencing methods has recently made it tractable to perform unbiased assessments of penetrance in population controls. In several instances such studies have suggested that previously reported Mendelian variants, as a class, are substantially less penetrant than had been believed. To date, however, all of these studies have been limited to fairly prevalent (>0.1%) diseases, and point estimates of the penetrance of individual variants have been limited to large copy number variations. Here, we examine the penetrance of variants previously reported as pathogenic in a dominant, monogenic disease gene, the prion protein gene (*PRNP*). By analyzing 16,025 prion disease cases, 60,706 population control exomes, and 531,575 individuals genotyped by 23andMe, Inc., we show that missense variants in *PRNP* previously reported to be pathogenic are ~100x more common in the population than expected based on genetic prion disease prevalence. Some of this excess can be attributed to at least three likely completely benign variants falsely assigned as pathogenic. However, we show that other variants have genuine effects on disease susceptibility, conferring lifetime risks ranging from <0.1% to ~100%. We also show that truncating variants in *PRNP* have position-dependent effects, with true loss-of-function alleles found in healthy older individuals, supporting the safety of therapeutic suppression of prion protein expression. Our results provide the first quantitative estimates of lifetime risk for hundreds of asymptomatic individuals who have inherited incompletely penetrant *PRNP* variants, and demonstrate the value of large reference datasets of human genetic variation for informing both genetic counseling and therapeutic strategy.

41

Evaluation of the regional variability of missense constraint in 60,000 exomes. *K.E. Samocha[1,2,3], M. Lek[1,2], D.G. MacArthur[1,2], B.M. Neale[1,2], M.J. Daly[1,2,3], Exome Aggregation Consortium.* 1) Massachusetts General Hospital, Boston, MA; 2) Broad Institute of Harvard and MIT, Cambridge, MA; 3) Harvard Medical School, Boston, MA.

Just as the resolution of sequence conservation improves when using 100 mammalian species instead of 5, exploring variation within tens of thousands of individuals, instead of hundreds, improves our ability to detect genetic sequences intolerant of mutations. We have previously used exome sequences from a large reference population to identify genes that are significantly depleted for missense and/or loss-of-function variation, indicating selective constraint against those types of mutations. While loss-of-function variation is usually considered to be a property of the gene, it has been well established that missense variants can have dramatically different effects depending on their locations in the gene. We therefore expect that, for a subset of genes, only regions of them will be truly missense constrained. We searched for these patterns of missense intolerance within genes by leveraging the genetic variation data from the 60,706 individuals in the Exome Aggregation Consortium (ExAC).In order to identify the regions of genes that show significant missense constraint, we extracted the rare (minor allele frequency < 0.001) missense variants from ExAC. The expected number of such variants was determined by using a sequence-context based model of mutation (Samocha et al 2014). We then employed a likelihood ratio test that used the number of observed and expected variants per exon to search for evidence of varying levels of missense tolerance between regions of a gene. Of the 5,760 genes with overall missense depletion ($\chi^2 \geq 10$), roughly a quarter of them (n=1,597) show regional variability in missense constraint.To explore the functional relevance of the regions with the most severe missense constraint, we overlaid *de novo* missense variants from autism cases (n=3,982) as well as those from controls (n=2,078). Autism exome studies have shown only a modest excess (~1.14) of *de novo* missense variants; we find that the regions of genes under the greatest constraint have an OR of 3, while those regions under no constraint show no difference between case and control *de novo* rates. We evaluated additional *de novo* and disease-associated variants sets to confirm the importance of these missense-constrained regions. The identification of these intolerant regions of genes, in conjunction with variant and amino acid level annotation, will be critical in the interpretation of variants found within a human exome.

42

MARV: A novel method and software tool for genome-wide multi-phenotype analysis of rare variants. *M. Kaakinen[1], R. Mägi[2], K. Fischer[2], M-R. Järvelin[3,4], AP. Morris[5], I. Prokopenko[1].* 1) Department of Genomics of Common Disease, Imperial College London, London, United Kingdom; 2) Estonian Genome Center, University of Tartu, Tartu, Estonia; 3) Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, Imperial College London, London, United Kingdom; 4) Institute of Health Sciences and Biocenter Oulu, University of Oulu, Oulu, Finland; 5) Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom.

Recently, genome-wide association studies (GWAS) have been expanded to analysis of low-frequency and rare variants (MAF≤5%, both denoted by RVs). Power for variant detection could also be increased by jointly analysing multiple correlated phenotypes. We have developed a method and software for genome-wide Multi-phenotype Analysis of RVs (MARV), combining features from both RV burden tests and multi-phenotype analyses. Specifically, the proportion of rare variants at which an individual carries minor alleles within a gene region is modelled on linear combinations of phenotypes in a regression framework. MARV also implements model selection via the Bayesian Information Criterion (BIC). Our preliminary simulation studies on 1,000 individuals with 10,000 replicates and two continuous phenotypes with a correlation ranging between -0.9 and 0.9 show good control of type I error rate. Power is increased when the genetic effect is in the opposite direction than the correlation between the phenotypes. We have applied MARV also on empirical data with three correlated phenotypes: fasting insulin (FI), triglycerides (TG) and waist-to-hip ratio (WHR), using data from 4,788 individuals from the Northern Finland Birth Cohort 1966. Individuals were genotyped on the Illumina370CNV array and imputed to the 1,000 Genomes Project all ancestries reference panel (March 2012). FI/TG/WHR were adjusted for sex, body mass index and three principal components to control for population structure, and the resulting residuals were used in the analysis. The following transformations were applied: natural logarithm for FI and inverse normal for the residuals of TG and WHR. We identified RV associations, at genome-wide significance (P<1.7x10^{-6}, Bonferroni correction for 30,000 genes) in *APOA5*, which is known to harbour both common and rare variants for TG and other lipids, and in *ZNF259*, which maps to a common variant GWAS locus for TG, several other lipids and coronary heart disease. For *APOA5* the model with TG only had the best fit (P_{TG}=2.2x10^{-7}), whereas for *ZNF259*, the model with TG and FI provided the best fit (P_{model}=3.1x10^{-9}), and stronger associations than in univariate analyses (P_{TG}=6.7x10^{-8}; P_{FI}=0.13). Using MARV, we demonstrate its ability to identify RV multi-phenotype associations with greater statistical significance than in univariate analyses, and for the first time show a role of *ZNF259* RVs in T2D/CHD-related trait variability, suggesting shared pathophysiology.

43

Association of copy number variations with decreased cognitive phenotypes and fitness in unselected populations. *A. Reymond[1], R. Magi[2], A. Mace[3,4], B. Cole[5], A. Guyatt[6], H. Shihab[6,7], A. Maillard[3], H. Alavere[2], A. Kolk[2,8], A. Reigo[2], E. Mihailov[2], L. Leitsalu[2,9], A.M. Ferreira[1,4], M. Nıukas[2,9], A. Teumer[10], E. Salvi[11], D. Cusi[11,12], M. McGue[13], W.G. Iacono[13], T.R. Gaunt[6,7], J.S. Beckmann[4], S. Jacquemont[3], Z. Kutalik[3,4,14], N. Pankratz[5], N. Timpson[6,7], A. Metspalu[2,9], K. Mannik[1,2].* 1) Ctr Integrative Genomics, Univ Lausanne, Lausanne, Switzerland; 2) Estonian Genome Center, University of Tartu, Tartu, Estonia; 3) Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland; 4) Swiss Institute of Bioinformatics, Lausanne, Switzerland; 5) University of Minnesota Medical School, Department of Laboratory Medicine & Pathology, Minneapolis, MN, USA; 6) Bristol Genetic Epidemiology Laboratories, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom; 7) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom; 8) Department of Neurology and Neurorehabilitation, Children's Clinic, Tartu University Hospital, Tartu, Estonia; 9) Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; 10) Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany; 11) Deparment of Health Sciences, University of Milan, Italy; 12) Institute of Biomedical Technologies, Italian National Research Council, Milan, Italy; 13) University of Minnesota Department of Psychology, Minneapolis, MN, USA; 14) Institute of Social and Preventive Medicine, Lausanne University Hospital, Switzerland.

The association of rare copy number variants (CNVs) with complex disorders was almost exclusively evaluated using clinically ascertained cohorts, thus the contribution of these variants to complex phenotypes in the general population remains unclear. We assessed the genome-wide burden of rare autosomal and X-linked CNVs on carriers' cognitive traits and fertility in the general population, as well as investigated clinical features of adults carrying a CNV, either syndromic or polymorphic, in genome intervals associated with known genomic disorders.For CNV analysis and genotype-phenotype associations with education and disease traits, we used a random sample of 12,000 individuals from the population biobank of Estonia (EGCUT). We identified altogether 3378 CNV calls in the intervals of DECIPHER-listed genomic disorders. Among these are 56 carriers of autosomal and 12 of X-linked syndromic CNVs. Their phenotypes are reminiscent of those described for carriers of identical rearrangements ascertained in clinical cohorts, thus our results challenge the assumption that carriers of known syndromic CNVs identified in population cohorts are asymptomatic.We also generated a genome-wide map of rare (frequency ≤0.05%) CNVs and identified 10.5% of the screened general population as carriers of CNVs ≥250kb. Carriers of deletions ≥250kb or duplications ≥1Mb show, compared to the Estonian population, a greater prevalence of intellectual disability (P=0.0015, OR=3.16; P=0.0083, OR=3.67, respectively), reduced mean education attainment (a proxy for intelligence; P=1.06e-04; P=5.024e-05, respectively), an increased fraction of individuals not graduating from secondary school (P=0.005, OR=1.48; P=0.0016, OR=1.89, respectively) and a decreased number of offsprings of females. These deletions show evidence of enrichment for genes with a role in neurogenesis, development, cognition, learning, memory, behavior and fertilization. Evidence for an association between rare CNVs and decreased educational attainment was confirmed by analyses in adult cohorts of Italian (HYPERGENES) and European American (Minnesota Center for Twin and Family Research) individuals, as well as in adolescents from the Avon Longitudinal Study of Parents and Children birth cohort.These results indicate that individually rare but collectively common intermediate-size CNVs contribute to the variance in educational attainment and other complex traits such as fitness.

44

The role of transcription in mammalian cell copy number variant formation. *S. Park[1], M.F. Arlt[1], S. Rajendran[1], M.T. Paulsen[2], R. Beroukhim[3,4,5], M. Ljungman[2], T.E. Wilson[1,6], T.W. Glover[1,6].* 1) Human Genetics, University of Michigan, Ann Arbor, MI; 2) Radiation Oncology and Translational Oncology, University of Michigan, Ann Arbor, MI; 3) Cancer Biology, Dana-Farber Cancer Institute, Boston, MA; 4) Medical Oncology, Dana-Farber Cancer Institute, Boston, MA; 5) Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA; 6) Pathology, University of Michigan, Ann Arbor, MI.

Copy number variants (CNVs) play a major role in human genomic variation, genetic disease, and cancer. Non-recurrent CNVs, characterized by variable breakpoint junctions often with microhomologies, are associated with many developmental disorders and are the predominant CNVs in cancers. Non-recurrent CNVs are thought to arise from replication errors, but the exact mechanism is unknown. We have demonstrated that partial inhibition of DNA replication (replication stress) induces CNVs in cultured human and mouse cells that mimic non-recurrent CNVs found in patients. While they occur throughout the genome, there are hotspots that are more prone to CNVs. To better understand the molecular basis of the hotspots, we compared large sets of *de novo* CNVs, both spontaneous and induced by aphidicolin, hydroxyurea, or ionizing radiation, in normal human and mouse cell lines. We found a high correlation among the genomic locations of CNV hotspots, active transcription units larger than 1 Mb, and common fragile sites, suggesting that large active transcription contributes to CNV formation. Unlike most transcribed genes, the large transcription units replicate late in the cell cycle. However, the majority of late-replicating regions were not CNV hotspots. Our observations suggest that late replication alone is insufficient for CNV formation, and large active transcription units drive CNV formation. This observation allows us to predict locations of CNV hotspots in any cell type with known transcription profiles. The hotspots are prone to CNVs only when they are actively transcribed. The genes associated with CNV hotspots are implicated in many human disorders. *De novo* germline CNVs in our CNV hotspot genes are associated with neurodevelopmental and other disorders. Our CNV hotspots are also CNV hotspots in human cancers. In addition, many of these large genes are predominantly transcribed in the brain, suggesting a risk for somatic CNVs in replicating neural stem and progenitor cells. We hypothesize that large active transcription units interfere with replication by removing DNA-bound proteins necessary for replication initiation, so origins that fire late to ameliorate the effect of replication stress are no longer available. We are testing this model by building a profile of key replication proteins at different stages of the cell cycle in cell lines with different transcription profiles.

45

Quantitative functional studies using *Drosophila melanogaster* identify dosage sensitive and sex-specific effects of neurodevelopmental genes. *J.S. Iyer, P. Patel, L. Pizzo, K. Vadodaria, Q. Wang, A. Kubina, S. Yennawar, R. Pandya, S. Girirajan.* Department of Biochemistry and Molecular Biology and Anthropology, The Pennsylvania State University, University Park, PA.

While recent studies have associated several genes and genomic regions for human diseases, systematic functional analysis of these candidates is limited due to a lack of high throughput and quantitative assays. We developed a battery of quantitative assays for high throughput functional evaluation of *Drosophila melanogaster* orthologs of human neurodevelopmental genes, including those mapping within the rare copy-number variant regions and recently identified single nucleotide variants from exome sequencing studies. We took advantage of the tissue-specific expression system conferred by the UAS-Gal4 system and used RNA interference to achieve eye-specific (*GMR*-Gal4), neuron-specific (*Elav*-Gal4), and ubiquitous (*Da*-Gal4) knockdown of neurodevelopmental genes in flies. Combining data from gene expression with quantitative assessment of neuronal phenotypes using multiple fly RNAi lines allowed us to correlate the effect of dosage alterations to severity. We performed a series of proof-of-concept experiments. We first tested >75 fly lines representing six human CNV regions including 1q21.1, 15q11.2, 15q13.3, 16p11.2, distal 16p11.2, and 16p12.1, for dosage sensitivity of fly orthologs of human genes. We find differential effects of gene dosage for several genes within CNV regions. For example, 8 out of 11 fly orthologs within 16p11.2 showed dosage-dependent change in severity including *C16ORF53*, *DOC2A*, *CDIPT*, *KCTD13*, *FAM57B*, *ALDOA*, *PP419C*, and *MAPK3*. Decreased head size was observed with knockdown of *KCTD13* (20 SD), *MAPK3* (14 SD), CDIPT (18 SD) and *C16ORF53* (6 SD) within 16p11.2 and *UQCRC* (9 SD), *POLR3E* (12 SD) and *CDR2* (14 SD) within 16p12.1 recapitulating the autism features observed in individuals with these deletions. We also tested 21 fly lines with disruption of 12-neurodevelopmental genes (such as *CTNNB1*, *SHANK3*, and *NRNX1)* and found sex-dependent severity of phenotypes for *CHD8*, *MCPH1* and *SCN1A*. Interaction studies using two-locus models and expression analysis identified key modifiers enhancing and suppressing the CNV phenotypes. For example, reduced expression of *CHD8* rescued phenotypes due to *KCTD13* knockdown and reduced expression of *SCN1A* rescued *DOC2A* knockdown phenotypes in flies. Our results suggest that neurodevelopmental genes are involved in differential roles in a dosage-sensitive and sex-specific manner, and a complex interplay between genes *in cis* and *in trans* contribute to the observed phenotypic variability in human patients with CNVs.

46

Single-cell analysis reveals that endogenous retrotransposons generate somatic mosaicism in neuronal and non-neuronal cells. *J.A. Erwin[1], A.C. Paquola[1], T. Singer[1], C. Quayle[1], T. Bedrosian[1], A. Muotri[2], R. Lasken[3], F.G. Gage[1].* 1) Salk Institute, La Jolla, CA; 2) University of California San Diego School of Medicine, La Jolla, CA; 3) J. Craig Venter Institute, La Jolla, CA.

It has long been thought that neuronal genomes are invariable; however, recent studies have demonstrated that mobile elements actively retrotranspose during neurogenesis, thereby creating genomic diversity between neurons. In addition, mounting data demonstrate that mobile elements are misregulated in certain neurological disorders, including Rett syndrome and schizophrenia. The unique composition of genetic mosaicism present in the brain may contribute to disease and also the behavior differences observed between genetically identical organisms. Many questions remain regarding the regulation of retrotransposition, the full characterization of other mobile elements and the functional significance of retrotransposition during neurogenesis. Because each individual neuron has the potential to have a unique genome, single-cell approaches are essential to measure and observe this genomic diversity, which is obscured in bulk samples. I will present data using single-cell genome sequencing to characterize the nature of genome mosaicism within the non-diseased soma. In order to address the question whether somatic retrotransposition occurs in other tissues and, more generally, how it impacts human development and function, we developed a targeted sequencing approach to identify Alu and L1 retrotransposition events in single cells and bulk tissues. We applied this method to cortex, hippocampus, heart and liver postmortem samples from four non-diseased young adults. We confirm that somatic L1 retrotransposition occurs in hippocampal neurons, and we also found evidence of somatic Alu retrotransposition in the liver as well as somatic L1 retrotransposition in non-neuronal cells in the cortex and liver. We observe similar rates of retrotransposition in neuronal and non-neurons cells. Therefore, that somatic retrotransposition is not restricted to neurons but occurs as part of the normal condition of human somatic cells.

47

Evolution and structural diversity of the complement factor H related gene cluster. *S. Cantsilieris[1], L. Harshman[1], N. Janke[1], E.E. Eichler[1,2].* 1) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; 2) Howard Hughes Medical Institute, Seattle, WA, USA.

The complement factor H related (CFHR) gene family maps to a complex ~420 kbp genomic region on chromosome 1q32 and shows extensive diversity in human populations. Both structural variants and SNVs have been associated with complex human genetic diseases including age-related macular degeneration (AMD), systemic lupus erythematosus (SLE) and atypical haemolytic uraemic syndrome (aHUS). Using massively parallel and PacBio SMRT sequencing of large insert-clones, we generated high quality finished sequence (~3Mb) over the 1q32 CFHR locus in multiple human haplotypes, great apes (orangutan, chimpanzee and gorilla) and macaque in an effort to reconstruct its evolutionary history. Comparative and phylogenetic analysis reveals that ~60 kb of sequence was added via segmental duplication to the homininae lineage in two separate events. Our initial timing estimates indicate that these events occurred 6.1 and 7.5 million years ago, leading to the creation of two genes (*CFHR3* and *CFHR1)*. Orangutans show the simplest genomic organization, lacking almost all duplications identified in other great apes and most modern humans. There is clear evidence that this gene cluster has been restructured multiple times during primate evolution; we have identified lineage specific structural events affecting genes, including a complete duplication of *CFHR1* to the distal end of *CFHR4* in the chimpanzee and a ~20kb duplication distal to the *CFH* gene in the macaque that may represent a novel *CFHR*-like gene. We find evidence of positive selection in great apes and humans, specifically in exon 22 of the *CFH* gene which forms part of a larger ~30kb segment, duplicated to two additional locations in the chimpanzee. Overall our analysis has revealed several structural changes affecting genes that we are currently genotyping in a diversity panel of >2500 human and >100 ape genomes. In addition, we are currently using molecular inversion probes (MIPs) combined with massively parallel sequencing to associate structural changes with coding variation, identify breakpoints, and detect signals of interlocus gene conversion in patients with AMD and SLE. These data reveal a remarkably dynamic region with recurrent CFHR gene gain and gene loss over the last 25 million years of primate evolution. Our set of high-quality alternate reference sequences provides an evolutionary and population genetic framework necessary to investigate the association of this locus with immune associated diseases. .

48

Repetitive DNA at CNV breakpoints is susceptible to gross chromosomal rearrangement. *K. Rudd[1], K.E. Hermetz[1], Y. Nishida[2], Y. Zhang[2], N. Saini[2], K.S. Lobachev[2].* 1) Human Genetics, Emory University School of Medicine, Atlanta, GA; 2) School of Biology and Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA.

Genomic deletions, duplications, and translocations are a major cause of neurodevelopmental disorders. These chromosome rearrangements arise via diverse mutational mechanisms and lead to recurrent or non-recurrent forms of copy number variation (CNV). Although most human CNVs are non-recurrent, sequence analysis of CNV breakpoints reveals an enrichment of certain types of repetitive DNA. These data suggest that double-strand breaks do not arise randomly and implicate particular types of repetitive DNA in chromosome rearrangements. However, we know little about the risk factors for chromosome breakage and CNV formation. Here we assess the fragility of 13 human CNV breakpoint motifs in a yeast gross chromosomal rearrangement (GCR) assay. Motifs are 500-6015 bp long and were derived from breakpoints of terminal deletions, interstitial deletions, inverted duplication-terminal deletions, and translocations on 11 different chromosomes. Eight motifs exhibited GCR 13-420-fold above background levels, and six of these had an orientation bias in fragility potential. Our analyses indicate this bias results from the ability of sequence motifs to form secondary structures during lagging strand synthesis. (TG)n dinucleotide, (GAA)n trinucleotide, *Alu*, and some classes of tandem repeats exhibit elevated GCR. Notably, the most fragile motif contains a combination of inverted *Alu*s and tandem repeats. This sequence is derived from human chromosome 17p13.3, recognized recently as a hotspot for deletions, duplications, and complex chromosome rearrangements. Southern blot and array CGH analyses of individual colonies reveals terminal deletions, translocations, and inverted duplications resulting from GCR in the yeast genome. Like fragile sites, some CNV breakpoints are made up of repetitive DNA that is susceptible to genomic instability. Our functional annotation of specific repeats points to new rearrangement-prone loci and reveals mechanisms of CNV formation.

49

A potential role for the linker for activation of T-cells (*LAT*) in the neuroanatomical phenotype of the 16p11.2 BP2-BP3 CNVs. *M.N. Loviglio[1], M. Leleu[2,3], T. Arbogast[4], K. Mannik[1,5], G. Giannuzzi[1], J. Beckmann[2,6], J. Rougemont[2,3], S. Jacquemont[6], N. Katsanis[4], C. Golzio[4], A. Reymond[1], 16p11.2 Consortium.* 1) Center for Integrative Genomics (CIG), University of Lausanne, Lausanne, Vaud, Switzerland; 2) Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland; 3) School of Life Sciences, EPFL (Ecole Polytechnique Fédérale de Lausanne), Lausanne, Switzerland; 4) Center for Human Disease Modeling, Duke University Medical Center, Durham, North Carolina; 5) Estonian Genome Center, University of Tartu, Tartu, Estonia; 6) Service of Medical Genetics, Lausanne University Hospital (CHUV), Lausanne, Switzerland.

Copy number variants (CNVs) are major contributors to genomic imbalances disorders. Phenotyping of 137 unrelated carriers of the distal 16p11.2 220 kb BP2-BP3 deletion and duplication region showed that these rearrangements are associated with mirror phenotypes of obesity/underweight and macro-/microcephaly, and autism spectrum disorders (ASD); such phenotypes, with the same direction of effect, have been previously reported for the proximal 16p11.2 600 kb BP4-BP5 deletion and reciprocal duplication,. These two CNVs-prone regions at 16p11.2 are also reciprocally engaged in complex chromatin looping, as successfully confirmed by 4C, FISH, Hi-C, concomitant expression changes and quantitative co-variation of transcription factor binding in one interval and gene expression in the other.Using the zebrafish embryo as an *in vivo* model, we dissected the 220kb BP2-BP3 region at 16p11.2, encompassing 9 genes: *CD19*, *NFATC2IP, ATXN2L, TUFM, ATP2A1, RABEP2, SPNS1*, *LAT* and *SH2B1*, known for its critical role in the control of human food intake and body weight, and candidate gene for the obesity/underweight phenotype displayed by the carriers of these CNVs.We modeled the duplication by overexpressing each individual human transcript in zebrafish embryos and we determined the level of cell proliferation in the brain by phospho-histone H3 antibody staining. We showed that zebrafish embryos injected with the linker for activation of T-cells (*LAT*) message showed a decreased number of proliferating cells in the brain at 2 days post-fertilization. Such phenotype is often associated with microcephaly at later developmental stages (Our studies on *KCTD13*, *AUTS2*, *BTG2* are few exemplars). When used as a 4C "viewpoint" in human control LCLs, *LAT* was found to interact strongly with both *MVP* and *KCTD13*, two of the three major players for the head size phenotypes associated with the 16p11.2 600kb BP4-BP5 CNVs, and its 4C cis- and trans-interacting partners were enriched for SFARI ASD genes (Fisher's exact test, p=5.58E-03, OR=1.9). We propose a new role for *LAT* in 16p11.2 (BP2-BP3) 220kb CNVs neurodevelopmental phenotypes, besides its well-recognized function in T-cells development.

50

Paired-Duplications Mark Cryptic Inversions and are a Common Signature of Complex Structural Variation that is Misclassified by Chromosomal Microarray. *H. Brand[1,2], R.L. Collins[1], C. Hanscom[1], J.A. Rosenfeld[3], V. Pillalamarri[1], M.R. Stone[1], F. Kelley[4], T. Mason[4], L. Margolin[4], S. Eggert[1], E. Mitchell[5], J.C. Hodge[5,6], J.F. Gusella[1,4,7], S.J. Sanders[8], M.E. Talkowski[1,2,4].* 1) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 2) Department of Neurology, Harvard Medical School, Boston, MA; 3) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 4) Program in Medical and Population Genetics and Genomics Platform, Broad Institute, Cambridge, MA; 5) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; 6) Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, CA; 7) Department of Genetics, Harvard Medical School, Boston, MA; 8) Department of Psychiatry, University of California San Francisco, San Francisco, CA.

Copy number variants (CNVs) have been the predominant focus of genetic studies of structural variation (SV), and chromosomal microarray (CMA) for CNV detection is the recommended first-tier screen for neurodevelopmental anomalies. However, CMA utility is limited to detecting large dosage imbalances and is blind to balanced structural variation (SV). We performed whole-genome sequencing (WGS) using large-insert jumping libraries in 259 individuals diagnosed with autism spectrum disorder from the Simons Simplex Collection who had previously undergone CMA and exome sequencing. Libraries had a median insert of 3,736 bp and were sequenced to 96.8x physical coverage on average. Comparing our SV classifier pipeline to high quality CMA variants in the same individuals, we found 95.9% sensitivity to detect deletions and 89.7% for duplications. Analyses of variants detected by WGS uncovered a myriad of complex SVs that were cryptic to, or misclassified by, CMA. The most abundant of these was a remarkably common yet previously uncharacterized class of SV that we termed dupINVdup, involving two duplications in close proximity ('paired-duplications') that flank the breakpoints of a large, cryptic inversion. We observed dupINVdups in 8.1% of all subjects, and yet they had not been characterized in previous population-based SV studies, emphasizing the strength of deep coverage from large-inserts for SV detection. Collectively, dupINVdup and other duplication-mediated complex SVs were observed in 15.8% of subjects, and breakpoint analysis suggested microhomology-mediated repair as the predominant mechanism of formation. Based on the striking prevalence of complex SVs, we scrutinized the landscape of all the identified duplications and inversions. Overall, we found that complex rearrangements are the norm among inversion variation detectable at jumping library resolution; 60.7% of all inverted segments were associated with additional complexity. Further, 7.3% of all rare duplications detected by CMA were misclassified and actually represented complex SVs. Collectively, these findings indicate that dupINVdup, as well as other complex duplication-associated SVs, represent relatively common sources of genomic variation that have not been captured by population-based CMA or low-depth WGS analyses. They also suggest that 'paired-duplication' signatures detected by CMA warrant further scrutiny in diagnostic testing as they may mark complex SV of potential clinical relevance.

51

RNA sequencing of a mouse-model of Spinal Muscular Atrophy reveals tissue-wide changes in splicing of U12-dependent introns in genes involved in cell cycle, intracellular trafficking and neuronal function. *T.K. Doktor[1], Y. Hua[2], H.S. Andersen[1], S. Brøner[1], A.R. Krainer[2], B.S. Andresen[1].* 1) Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, Denmark; 2) Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA.

Spinal Muscular Atrophy (SMA) is a severe neuromuscular disorder resulting in the progressive loss of motor neuron function and subsequent loss of voluntary muscle control and in many cases leading to death during infancy. Mutations that delete or disrupt the function of the *SMN1* gene cause SMA, but all patients retain a secondary gene copy – *SMN2* – that provides a minimal level of the essential SMN protein, but fails to fully compensate for the loss of *SMN1*.The SMN protein has multiple reported functions, but it is most well characterized as being part of the SMN complex, which plays a role in the snRNP maturation pathway. Decreased SMN levels leads to perturbation of the snRNP levels and restoration of snRNP levels has previously been shown to alleviate symptoms in animal models. Additionally, aberrant splicing has been demonstrated in several animal models as well as patient cells, and has been investigated in a few tissues using exon arrays, and in isolated cell populations using RNA-seq.However, evidence suggests that SMA pathology is not restricted to motor neurons, but rather that systemic pathologies may contribute to motor neuron loss. To date, a comprehensive multi-tissue study on aberrant splicing has not been published and for this purpose we used RNA-seq to study the transcriptional landscape in multiple tissues in an SMA mouse model. Briefly, we isolated RNA from brain, spinal cord, liver, and muscle from SMA mice and their heterozygous littermates on post-natal day 1 (PND1) and post-natal day 5 (PND5) and performed RNA-sequencing using Illumina paired-end protocols.Here, we present data demonstrating that hundreds of U12-dependent introns are retained in SMA mice and this pattern of aberrant splicing may be an important molecular mechanism in the pathogenesis of SMA. In particular, we use qRT-PCR to confirm missplicing of U12-dependent introns in the cell cycle regulator *Rasgrp3*, in *Myh9* and in the neuronal development genes *Myo10*, *Zdhhc13* and *Cdk5*. Analysis of the RNA-seq data furthermore indicates that retention of U2-dependent introns may also play a role in SMA pathogenesis, but that this aberrant splicing was very heterogenous across tissues. In conclusion, this study identifies several aberrant splicing events associated with SMA that provide important clues to the exact molecular mechanisms behind motor neuron loss and peripheral symptoms observed in SMA patients.

52

BAC transgenic model of C9ORF72 ALS/FTD. *Y. Liu[1,2,3], A. Pattamatta[1,2,3], T. Zu[1,2,3], T. Reid[1,2,3], L.P.W. Ranum[1,2,3].* 1) Center for NeuroGenetics, University of Florida, Gainesville, FL; 2) Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL; 3) Genetics Institute, University of Florida, Gainesville, FL.

Amyotrophic lateral sclerosis (ALS) is a devastating disease which leads to progressive paralysis and death, usually within 2-5 years after disease onset. A GGGGCC hexanucleotide intronic repeat expansion mutation within the *C9orf72* gene was recently shown to be the major genetic cause of familial and sporadic forms of both ALS and frontotemporal dementia (FTD). However, the mechanisms by which this hexanucleotide repeat expansion causes the disease are not clear. We, and others, have previously shown that the *C9orf72* GGGGCC•GGCCCC expansion mutation is bidirectionally transcribed. Also, the sense and antisense RNA foci and **r**epeat **a**ssociated **n**on-ATG (RAN) proteins accumulate in human *C9orf72* ALS/FTD autopsy brains. To gain insight into the molecular mechanisms of the disease we developed a BAC transgenic mouse model. We generated and screened a bacterial artificial chromosome (BAC) library from a patient-derived *C9orf72* ALS/FTD lymphoblastoid cell line and identified a BAC containing an expansion with a full-length *C9orf72* gene for pronuclear injections. We established and characterized multiple transgenic lines including two lines with expansions containing ~500 GGGGCC repeats. These C9 BAC mice express the human transgene at approximately endogenous levels and recapitulate hallmark features of *C9orf72* ALS/FTD. First, these mice develop paralysis and hyperactivity phenotypes that mirror phenotypes observed in ALS patients. Second, these animals show decreased survival that correlates with transgene expression and repeat length in independent transgenic lines. Third, these mice show sense and antisense RNA foci and RAN protein accumulation. Fourth, these mice develop marked neuronal loss in multiple regions of the CNS including the frontal cortex and spinal cord. Finally, these animals, which express both the sense and antisense transcripts using the endogenous human promoter and regulatory regions, provide a novel model for the understanding the molecular mechanisms of ALS/FTD and for the development of therapeutic strategies.

53

Altered RNA processing in ALS4. *C. Grunseich[1], I.X. Wang[2], J. Watts[2,3], T. Lanman[1], G. Ramrattan[2,3], Z. Zhu[2], D. Bakar[1], A.B. Schindler[1], E. Hartnett[1], K.H. Fischbeck[1], V.G. Cheung[2,3,4].* 1) Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA; 2) Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA; 3) Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA; 4) Department of Pediatrics and Genetics, University of Michigan, Ann Arbor, MI 48109, USA.

ALS4 is a degenerative disease caused by autosomal dominant mutations in the RNA-DNA helicase senataxin. The disease is characterized by slowly progressive weakness, with signs of both upper and lower motor neuron involvement. Senataxin has been found to promote the resolution of RNA/DNA hybrids (R-loops) through its helicase activity, however the mechanism by which alteration in senataxin function results in neurodegeneration is unknown. To address this question we have taken a two-pronged approach combining clinical and basic studies to investigate R-loop biology in the disease. We collected detailed phenotypical information from 12 individuals with ALS4, including MRI volumetric imaging of the thigh and brain. The average age of disease onset in our cohort is 19 years, ranging from ages 6 to 40 yrs. Two individuals with the mutation at ages 27 and 31 have no detectable weakness. The ratio of thigh muscle to total cross sectional area was found to have a significant negative correlation with disease duration. Five subjects had evidence of cerebellar dysfunction with dysmetria or dysdiadochokinesis on exam. DNA and RNA profiles from skin fibroblasts, lymphoblastoid cell lines, and white blood cells were analyzed from 12 patients and 10 controls. Similar profiles were also derived from induced pluripotent stem cell (iPSC) lines and differentiated FACS-sorted motor neurons from 5 patients and 3 controls. Senataxin is expressed not only in iPSC-derived motor neurons and human nervous tissue, but is also found in many other tissues.Consistent with senataxin's ability to resolve R-loops and likely gain of function with the ALS4 mutation, a reduction in the abundance of R-loops was detected in patient cells by immunoprecipitation using S9.6 antibody for R-loop identification. These observations suggest that alterations in RNA processing, specifically R-loop resolution, and consequential effects on gene expression, may adversely affect motor neurons in ALS4. We are also evaluating how variation in R-loops could account for the wide phenotypic spectrum in our cohort. Our investigations offer new insight by providing a connection between R-loop and the biology of the disease. Identification of R-loop disruption in the disease may provide new targets for therapeutic development.

54

A recurrent mutation in *KCNA2* in complicated autosomal dominant spastic paraplegia: an expansion of the channelopathy spectrum and a novel disease mechanism. *K.L. Helbig[1], U.B.S. Hedrich[2], A.C. Teichmann[3], J. Hentschel[3], D.N. Shinde[1], W.A. Alcaraz[1], S. Tang[1], C. Jungbluth[4], S.L. Dugan[4,5], R. Schüle[6], H. Lerche[2], J.R. Lemke[3].* 1) Division of Clinical Genomics, Ambry Genetics, Aliso Viejo, CA., USA; 2) Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany; 3) Institute of Human Genetics, University Hospital Leipzig, Leipzig, Germany; 4) Department of Medical Genetics, Children's Hospitals and Clinics of Minnesota, Minneapolis, MN, USA; 5) Division of Medical Genetics, University of Utah, Salt Lake City, UT, USA; 6) Department of Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany.

The hereditary spastic paraplegias (HSPs) are a genetically and clinically heterogeneous group of neurodegenerative disorders characterized by spasticity and weakness in the lower extremities. Over 50 genes have been identified for HSPs, involved in a variety of cellular processes. However approximately 20% of familial HSPs remain unsolved. To date ion channels have not been implicated in HSPs. Diagnostic exome sequencing was performed on DNA from the peripheral lymphoblasts of a three-generation family with three affected individuals with HSP (Family 1). Family 2 with two affected individuals across two generations underwent exome sequencing as part of an ongoing research study on the genetic basis of HSP. Both families were found to have the identical c.881G>A (p.R294H) mutation within the voltage sensor of *KCNA2*, encoding the voltage-gated potassium channel $K_v1.2$, a member of the shaker potassium channel family. This mutation segregated with childhood onset spasticity and intellectual disability in five affected individuals from two unrelated families in an autosomal dominant fashion. Onset of spasticity was as early as two years. Cognitive outcomes were variable, with all three affected individuals in Family 1 displaying mild intellectual disability; one individual additionally had a diagnosis of autism spectrum disorder. In Family 2 the proband had mild intellectual disability but the affected mother had normal intellect. The p.R294H mutation is absent from all population databases (ExAC, 1000 Genomes, and EVS), and is predicted to be deleterious by *in silico* prediction models. The R294 amino acid is the first of seven gating charges in the $K_v1.2$ potassium channel S4 transmembrane segment, which forms the voltage sensor domain. Two-electrode voltage-clamp recordings of *Xenopus laevis* oocytes expressing mutant channels showed a loss of the $K_v1.2$ channel's function with a dominant-negative effect causing a decrease in current amplitude and a small depolarizing shift of the activation curve in comparison to wildtype channels. In addition, it has been shown previously for the shaker potassium channel that replacement of the first arginine within the S4 voltage sensor with a histidine causes the formation of a proton pore at hyperpolarized potentials. This finding expands the channelopathy spectrum to include HSP and represents a novel HSP disease mechanism.

55

Mouse Resources for Comparative Mendelian Genomics. *L. Reinholdt[1], H. Fairfield[1], A. Srivastava[1], R. Liu[2], A. Lakshminarayana[2], B. Harris[1], S. Karst[1], M. Berry[1], P. Ward-Bailey[1], C. Byers[1], A. Czechanski[1], W. Martin[1], K. Cheng[1], L. Goodwin[1], J. Morgan[1], D. Bergstrom[1].* 1) The Jackson Laboratory, Bar Harbor, ME; 2) The Jackson Laboratory for Genomic Medicine, Farmington, CT.

Spontaneously arising mouse mutations have served as the foundation for understanding gene function for over 100 years. Discovery of Mendelian disease genes in the mouse genome is powered by the availability of large consanguineous pedigrees and genetically defined inbred strain backgrounds that minimize genetic heterogeneity. Moreover, causation can be readily supported through bulk segregation analysis and ultimately proven through genetic engineering — a field that is now experiencing a paradigm shift of its own with the advent of CRISPR/Cas9 technology. At The Jackson Laboratory, we are taking advantage of the world's largest collection of mouse strains with Mendelian disease phenotypes and are using whole exome sequencing (WES) to discover the underlying disease genes. To date, we have successfully identified putative pathogenic mutations for 91 strains. However, nearly 50% of our cases remain unsolved using our standard exome sequencing analytics pipeline. Using a combination of approaches, we have sought to understand the nature of exome recalcitrant mutations and here we provide evidence that a large fraction of unsolved exome cases involve structural mutations. This result directly informs efforts to investigate the similar proportion of apparently Mendelian human phenotypes that are recalcitrant to exome sequencing. To complement our forward genetics approach and to advance precision modeling of human Mendelian diseases, we are now working closely with the Baylor-Hopkins Center for Mendelian Genomics (BHCMG) and NHLBI's Bench to Bassinet Program to identify common candidate genes and to provide a core to engineer orthologous mouse models using CRISPR/cas9 with targeted phenotyping.

56

Activation of the DNA damage response in an induced model of spinal muscular atrophy. *M. Jangi[1], H. Li[2], X. Yang[2], P. Cullen[1], A. Thai[1], M. Liu[1], C. Fleet[1], C.F. Bennett[3], F. Rigo[3], A.R. Krainer[4], C. Roberts[2], N. Allaire[1], C. Sun[1], J.P. Carulli[1], J.F. Staropoli[1].* 1) Division of Genetics and Genomics, Biogen, Cambridge, MA; 2) Division of Computational Biology, Biogen, Cambridge, MA; 3) Neuroscience Drug Discovery, Isis Pharmaceuticals, Inc., Carlsbad, CA; 4) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Spinal muscular atrophy is an autosomal recessive neuromuscular disease that is the leading genetic cause of infant mortality. Homozygous loss of the gene *survival of motor neuron 1* (*SMN1*) causes selective degeneration of lower motor neurons and ultimately leads to atrophy of proximal skeletal muscles. Disease severity is modified by variable copies of the paralogous *SMN2* gene, from which ~20% of transcripts encode a stable protein, SMN, that is identical to the *SMN1* protein product. SMN is ubiquitously expressed and is a key factor in the assembly of the core splicing machinery. SMN also plays cytoplasmic roles in stress granule assembly and axonal mRNA transport and translation. It remains unclear how a reduction in SMN levels causes degeneration of one neuronal population with such remarkable specificity. We have developed an antisense oligonucleotide (ASO)-based inducible mouse model of SMA that allows separation of early postnatal developmental changes from SMN-specific signatures. To identify transcriptome changes most proximal to SMN loss, we performed deep sequencing of poly(A)+ RNA from spinal cords of adult mice at 10, 20, and 30 days following SMN depletion with an SMN exon 7-skipping ASO. Reads were mapped to the transcriptome using STAR, and gene-level and isoform-level expression was quantified using RSEM and MISO, respectively. Despite the well-studied role of SMN in spliceosome biogenesis, we found little evidence for widespread splicing defects at any time point; most of the significant splicing changes, including 53 retained introns and 84 alternative exons at day 30, appeared to be downstream of activation of other cellular programs. At the gene level, later time points showed strong induction of the p53 pathway, DNA damage response, and mediators of apoptosis. This was accompanied by expression of cell cycle checkpoint proteins, suggestive of aberrant cell cycle reentry during degeneration of post-mitotic neurons. While glia may be contributing to these expression patterns, the lack of a glial activation signature suggests that the predominant contribution is neuronal. These observations are consistent with reports in other neurodegenerative diseases in which cell cycle reactivation in neurons precedes or may even be required for apoptosis. We propose that decreased SMN expression sensitizes cells to DNA damage through a mechanism distinct from its role in pre-mRNA splicing and activates cell cycle signaling to mediate apoptosis.

57

Recessive mutations in the UGO1-like protein SLC25A46 cause an optic atrophy. *T. Huang[1], S. Zuchner[2,3], J. Dallman[4], V. Carelli[5,6], A. Abrams[2,3,4], R. Hufnagel[1], A. Rebelo[2,3], C. Zanna[5,6], N. Patel[4], M. Gonzalez[2,3], I. Campeanu[4], L. Griffin[7,8], S. Groenewald[4], A. Strickland[2,3], F. Tao[2,3], F. Speziani[2,3], L. Abreu[2,3], R. Schüle[2,3], L. Caporali[5], C. Morgia[5,6], A. Maresca[5,6], R. Liguori[5,6], R. Lodi[9], Z. Ahmed[10], K. Sund[10], X. Wang[1], L. Krueger[1], Y. Peng[1], C. Prada[1], C. Prows[1], Kevin Bove[11], Elizabeth K. Schorry[4], Anthony Antonellis[7,8], Holly H. Zimmerman[12], Omar A. Abdulrahma.* 1) Prof, Pediatrics/Div Human Gen, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 2) John P. Hussman Institute for Human Genomics, Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami, Miami, FL; 3) Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami, Miami, FL; 4) Department of Biology, University of Miami, Coral Gables, FL; 5) IRCCS Institute of Neurological Sciences of Bologna, Bellaria Hospital, Bologna, Italy; 6) Neurology Unit, Department of Biomedical and NeuroMotor Sciences (DIBINEM), University of Bologna, Bologna , Italy; 7) Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI; 8) Department of Neurology, University of Michigan Medical School, Ann Arbor, MI; 9) Policlinico S. Orsola-Malpighi Department of Biomedical and NeuroMotor Sciences (DIBINEM), University of Bologna, Bologna, Italy; 10) Department of Otorhinolaryngology Head & Neck Surgery, School of Medicine, University of Maryland, Baltimore, MD.

Alexander J. Abrams2,3,4, Robert B. Hufnagel1, Adriana Rebelo2,3, Claudia Zanna5,6, Neville Patel4, Michael A. Gonzalez2,3, Ion J. Campeanu4, Laurie B. Griffin7,8, Saskia Groenewald4, Alleene V. Strickland2,3, Feifei Tao2,3, Fiorella Speziani2,3, Lisa Abreu2,3, Rebecca Schüle2,3, Leonardo Caporali5, Chiara La Morgia5,6, Alessandra Maresca5,6, Rocco Liguori5,6, Raffaele Lodi9, Zubair M. Ahmed10, Kristen L. Sund1, Xinjian Wang1, Laura A. Krueger1, Yanyan Peng1, Carlos E. Prada1, Cynthia A. Prows1, Kevin Bove11, Elizabeth K. Schorry1, Anthony Antonellis7,8, Holly H. Zimmerman12, Omar A. Abdulrahman12, Yaping Yang14, Susan M. Downes17, Jeffery Prince4, Flavia Fontanesi15, Antonio Barrientos15,16, Andrea H. Nemeth17,18, Valerio Carelli5,6#, Stephan Zuchner#$2,3, Julia E. Dallman#4, Taosheng Huang1#$# These authors contributed equally$ These are the corresponding authorsOptic nerve atrophy[1] and axonal peripheral neuropathy (CMT2A)[2] are hereditary neurodegenerative disorders primarily caused by mutations in the canonical mitochondrial fusion genes *OPA1* and *MFN2*, respectively[3]. Interestingly, some patients present symptoms of both diseases[4,5], indicating mechanistic overlap. In yeast homologs of OPA1(Mgm1p) and MFN2 (Fzo1p) work in concert with a third protein, Ugo1p, whose ortholog remains to be identified in mammals. We found recessive mutations in the putative mitochondrial carrier gene, *SLC25A46*, in three families with both optic atrophy and axonal CMT phenotypes. Furthermore, we present evidence that SLC25A46 is the closest equivalent to Ugo1p in vertebrates and demonstrate its role in mediating mitochondrial morphology *in vitro* and *in vivo*. In zebrafish we found that loss-of-function affects the development and maintenance of neuronal processes and causes abnormal mitochondrial fusion morphology. Our results attest that identifying and characterizing rare disease genes is a relevant approach to elucidate common pathways of neuronal degeneration.

58

Neuronal aneuploidy and associated apoptosis in familial and sporadic frontotemporal lobar degeneration indicate that FTLD, like Alzheimer's disease and Niemann-Pick C1, is a cell cycle disorder. *H. Potter[1,2], J. Caneus[1,2], A. Granic[3], D. Dickson[4].* 1) NeurDepartment of Neurology and Linda Crnic Institute for Down Syndrome, University of Colorado Anschutz Medical Campus Aurora CO, USA; 2) Neuroscience Program, University of Colorado Anschutz Medical Campus Aurora CO, USA; 3) Institute of Health and Society and Newcastle Institute for Ageing, Campus for Ageing and Vitality, Newcastle University, Newcastle upon Tyne, NE4 5PL United Kingdom; 4) Neuropathology Laboratory, Mayo Clinic, 4500 San Pablo Rd., Jacksonville FL 32224.

The mechanism(s) responsible for neuronal cell death and cognitive decline in neurodegenerative diseases remain unclear. Chromosome-specific FISH and other analyses by our laboratory and others have shown the presence of high levels of mosaic aneuploidy, including up to 10% trisomy 21, in brains and peripheral tissues from sporadic and familial Alzheimer's disease (AD) and Niemann-Pick C1 patients. Mitotic spindle abnormalities and aneuploidy also arose in mouse and cell culture models of these disorders. In AD, we determined the mechanism of chromosome mis-segregation: the Aβ peptide, either endogenous or exogenous, competitively inhibits specific microtubule motors, particularly Eg5/kinesin 5, that are essential for mitotic spindle structure and function. Here we report that mosaic aneuploidy is also evident among neurons and glia in cortical samples from frontotemporal lobar degeneration (FTLD) patients carrying mutations in either the MAPT, progranulin, or C9ORF72 genes or who are apparently non-genetic/sporadic. Introduced mutant MAPT genes induce chromosome mis-segregation and aneuploidy in cell cultures and mouse models of FTLD. In both the human brain samples and the transfected cells, apoptosis is strongly associated with the aneuploidy, with 80% of aneuploid cells being TUNEL+. Based on these findings, it appears that defects in mitosis leading to aneuploidy may constitute a pathological mechanism contributing to neuronal loss and cognitive impairment in individuals with AD and FTLD and potentially other neurodegenerative diseases. To strengthen this hypothesis, we compared the level of mosaic aneuploidy in brain tissues from three groups of individuals: those characterized as having (**1**) AD, (**2**) AD pathology and normal cognition (ADPNC) and (**3**) normal cognition with no AD pathology (control). The data revealed a significant increase in numbers of aneuploid cells (trisomy and monosomy for chromosome 12 and 21) in the individuals with AD compared to the non-demented (ADPNC) and control patients, whereas there was no significant difference between ADPNC and controls, indicating that aneuploidy is better correlated to cognitive decline than is amyloid or tau pathology. Together, these data indicate that mosaic aneuploidy is involved in the progression of multiple neurodegenerative diseases and that understanding the mechanism by which it arises will set the foundation for the development of novel preventative therapeutics.

59

Results from the largest GWAS of Autism Spectrum Disorder to date. *J. Grove[1,2], The iPSYCH-SSI-Broad/MGH collaboration and Psychiatric Genomics Consortium Autism Working Group.* 1) Department of Biomedicine, Aarhus University, Aarhus , Denmark; 2) Bioinformatics Research Centre, Aarhus University, Aarhus , Denmark.

Autism Spectrum Disorder (ASD) is a childhood onset psychiatric disorder characterized by qualitative impairments in social interaction and communication, repetitive stereotypic behavior, and in some cognitive deficits. Worldwide prevalence hover around 1 percent. Its etiology is largely unknown, but ASD is highly heritable and it has been estimated that common variation explains about half of the genetic risk. Initial genome-wide association studies (GWAS) of ASD have reported a few significant associations but these have not led to robust replication in subsequent GWAS. In collaboration between iPSYCH, Statens Serum Institut (SSI) and Broad/MGH we have conducted the largest GWAS of ASD up to now, and followed up and meta-analysed with the large GWAS of ASD from the Psychiatric Genomics Consortium (PGC) in a combined analysis of 32132 subjects.The Danish sample is a population sample where cases were identified in the Danish Psychiatric Central Research Register and the controls constitute a random population sample. All subjects were then identified in the Danish Neonatal Screening Biobank, their DNA extracted, whole-genome amplified and genotyped on the PsychChip, a customized HumanCoreExome chip. The data is processed using the Ricopili pipeline of PGC. Heritability and genomic correlations are estimated by LD score regression and GCTA.We report here on the first data freeze consisting roughly of 60% of the total Danish sample, including 7783 cases and 11359 controls, showing 2 genome-wide significant loci. Meta-analysis with the PGC, comprising a total of 32132 subjects, reveals 4 genome-wide significant loci. Estimates of SNP heritability on the liability scale are 12% for the Danish sample and 13% for the combined sample. The genetic correlation between the Danish and the PGC samples is 82% for those of European ancestry and 75% for the full sample. Additional analyses of among other things chromosome X, sub-phenotypes and stratified heritability are ongoing and the results will be presented at the meeting.With a combined sample of just over 32k we report on the largest ASD GWAS to date. While still underpowered for a substantial dissection of the genetic architecture underlying ASD, the study is beginning to show robust signals. The present analyses identify 4 genome-wide significant loci, of which 3 appear to be novel. This stresses the importance of common variation in ASD etiology, and reveals new leads to understand the underlying biology.

60

De novo likely gene disrupting mutations and genic copy number variants increase the risk for Tourette's Disorder. *T.V. Fernandez[1], R.A. King[1], J. Xing[2], A.J. Willsey[3], A. Dietrich[4], J.A. Tischfield[2], G.A. Heiman[2], M.W. State[3], The TIC Genetics Collaborative Group.* 1) Yale Child Study Center and Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA; 2) Rutgers, the State University of New Jersey, Department of Genetics and the Human Genetics Institute of New Jersey, Piscataway, NJ, USA; 3) Department of Psychiatry, University of California, San Francisco, USA; 4) University of Groningen, University Medical Center Groningen, Department of Child and Adolescent Psychiatry, Groningen, The Netherlands.

Tourette's Disorder (TD) is an often-disabling developmental neuropsychiatric syndrome, characterized by persistent motor and vocal tics, with strong evidence for a genetic etiology. While current treatments have limited efficacy and may carry significant long-term adverse effects, the fundamental challenge in identifying novel therapeutic targets is our limited understanding of underlying biological mechanisms. Studying *de novo* (DN) mutations in cases versus controls has rapidly proven to be a powerful approach for gene discovery in complex neuropsychiatric disorders. We evaluated 325 TD parent-child trios using whole-exome sequencing to detect rare coding and splice site single nucleotide (SNVs) and insertion-deletion (indel) variants. BWA, SAMtools, and GATK were used to detect DN variants which were confirmed by Sanger sequencing. Additionally, we evaluated 412 trios for DN copy number variants (CNVs) using genome-wide SNP microarrays, CNVision, and qPCR confirmation. For the first time, we identified a significant excess of DN sequence and structural variation in TD versus controls. Likely gene disrupting (LGD) sequence mutations (premature stop codons, canonical splice site mutations, frameshift indels) confer an approximately 2-fold increase in risk for TD (OR 1.98 [1.14-3.44], p=0.01). Missense mutations predicted to be damaging to gene function also carry increased risk, though of lesser magnitude (OR 1.8 [1.22-2.66], p=0.002). Similarly, we find that DN genic CNVs are over-represented in affected individuals (OR 3.9 [1.2-12.7], p=0.01). These variants cluster within genes that are more intolerant to variation and those that are more highly expressed in the thalamus and striatum between birth and adolescence, brain regions and time periods that have long been implicated in TD etiology. Furthermore, we replicate our earlier finding of significant overlap between genes harboring DN LGD variants in TD and autism, reinforcing the hypothesis that there may be shared biological risk mechanisms between these two neurodevelopmental disorders. Finally, we find that the DN variants in TD converge on biological pathways that play key roles in dopaminergic signaling in the CNS. These findings indicate that continued genomic investigations of DN sequence and structural variation in TD are likely to identify specific risk genes through the identification of recurrent mutations.

61

Rare copy number variants implicate neuronal cell adhesion molecules in Tourette Syndrome. *A. Huang[1], D. Yu[2], L. Davis[3], C. Mathews[4], P. Paschou[5], N. Freimer[1], J. Scharf[2], G. Coppola[1], TSA International Consortium for Genomics, and the GTS GWAS Replication Initiative.* 1) University of California, Los Angeles, Los Angeles, CA; 2) Massachusetts General Hospital, Boston, MA; 3) University of Chicago, Chicago, IL; 4) University of California San Francisco, San Francisco, CA; 5) Democritus University of Thrace, Alexandroupoli, Greece.

Tourette syndrome (TS) is a complex neuropsychiatric disorder characterized by repetitive, involuntary motor and phonic tics of childhood onset. Although it has been well-established that genetics play a considerable role in TS, the identification of strong candidate genes has escaped research efforts for the past three decades. Published genome-wide association studies based on common genetic variation have thus far failed to establish any firm susceptibility loci. Therefore, in the current study, we focused on the analysis of rare (frequency < 1%) copy number variations (CNVs) in TS in a large cohort of European descent. Following extensive quality control, the dataset used in this analysis contained 6,042 unrelated samples consisting of clinically diagnosed TS cases (n=2,585) and ethnically matched controls (n=3,457), all assayed on dense, genome-wide Illumina OmniExpress single nucleotide polymorphism genotyping arrays, making ours the largest and most comprehensive CNV study in TS to date. We created a set of high-confidence CNV calls based on the consensus of two separate calling algorithms, and with these data, conducted several different types of analysis. We compared the CNV calls in our sample to those with convincing evidence for involvement in a variety of neuropsychiatric disorders, as well as those previously implicated in TS. To search for novel TS loci, we performed both segmental and gene-based tests of CNV association. Finally, we conducted a burden analysis of CNVs between cases and controls, stratified by both CNV type and size. We demonstrate evidence supporting the pathogenicity of rare exonic deletions in Neurexin 1 (*NRXN1*), which have previously been implicated in other neuropsychiatric conditions including autism spectrum disorder, epilepsy, and schizophrenia. Additionally, we find that rare duplications spanning the contactin 6 gene (*CNTN6*) are significantly enriched in TS cases (P=5.4 x 10-3, gene-wise test, after genome-wide empirical correction for multiple testing). Finally, burden analysis demonstrates a significant increase (p=0.0011) in the rate of rare, large, and likely pathogenic CNVs (>500kb) in TS patients compared to controls. Taken together, our study strongly supports a role for rare CNVs in the genetic etiology of TS and moreover, suggests additional CNVs that contribute to disease susceptibility to remain to be found.

62

Whole exome sequencing with simultaneous analysis of both parents has a high diagnostic yield for patients with epilepsy and neurodevelopmental disorders. *M. Stosser, T. Brandt, K. Retterer, J. Juusola, G. Richard, S. Suchy, D. McKnight.* GeneDx, Gaithersburg, MD.

Choosing an optimal testing strategy for patients with epilepsy and neurodevelopmental disorders (NDD) can be challenging and depends on the positive diagnostic rate (PDR) of the different genetic tests. The objective of this retrospective study was to evaluate the PDR of whole exome sequencing (WES) and multi-gene panels (NextGen sequencing and exon-level array CGH) for patients with epilepsy and NDD who were referred for testing to our molecular diagnostic laboratory between December 2011 and December 2014. A positive result was defined as one or two pathogenic or likely pathogenic variants in a single gene, depending on the mode of inheritance of the disorder. The highest diagnostic yield (171/450=38%) was achieved when WES analysis was performed on both the proband and the parents (WES-Trio). The diagnostic yield (18/79=23%) observed for proband-only WES analysis (WES-Proband) was strikingly lower and comparable to that of targeted multi-gene panels for epilepsy-related disorders. The PDR of all targeted epilepsy panels performed at our laboratory during this time period was 16% (899/5776). The infantile onset epilepsy panel (53 epilepsy-related genes) had a PDR of 21% (340/1620), and the PDR of the comprehensive panel (70 epilepsy-related genes) was 15% (437/2919). Using WES-Trio testing, the majority of autosomal dominant (83%) and X-linked disorders (69%) were determined to be caused by *de novo* mutations, which demonstrates the value of concurrent parental testing and explains the higher PDR of WES-Trio. The number of genes associated with epilepsy-related disorders is rapidly expanding and often the mutation spectrum of these newly-emerging genes is unknown. Defining the clinical significance of novel missense variants in one of these genes can be challenging but *de novo* occurrence can aid in proper variant classification. We made this observation in five new epilepsy-related genes, ALG13, GABRB2, KCNB1, NR2F1, and WDR45, where novel *de novo* changes accounted for 9% (15/171) of the positive WES-Trio results in our cohort. These data demonstrate that the highest PDR for patients with epilepsy and NDD was achieved by WES-Trio testing, which allows for the identification of *de novo* mutations. When both parents are not available for genetic testing, the PDR of WES only marginally exceeds that of a targeted epilepsy panel. These data may assist clinicians in determining the most effective testing strategy for their patients with epilepsy and NDD.

63

Gene discovery and high-throughput resequencing of candidate genes in epileptic encephalopathies. *C.T. Myers[1], J.M. McMahon[2], A. Schneider[2], R.K. Møller[3,4], G.L. Carvill[1], I.E. Scheffer[2], H.C. Mefford[1], Epi4K Consortium.* 1) University of Washington, Department of Pediatrics, Seattle, WA, USA; 2) University of Melbourne, Department of Medicine, Florey Institute of Neurosciences and Mental Health, Austin Health, Melbourne, Australia; 3) University of Southern Denmark, Institute for Regional Health Services, Odense, Denmark; 4) Danish Epilepsy Centre.

Objective: Whole exome studies in patients with epileptic encephalopathies (EE) have demonstrated the breadth of genetic heterogeneity in these severe childhood epilepsy syndromes. Our previous study identified 329 *de novo* mutations in 305 genes when 264 trios (affected child and unaffected parents) were sequenced. We aimed to identify additional patients with *de novo* mutations in 27 of these candidate genes to confirm the role of each gene in EE and to further define the phenotypic spectrum. **Methods**: We performed targeted capture and high-throughput resequencing of 27 genes in which a *de novo* mutation was identified in one or more proband with Infantile Spasms (IS) or Lennox-Gastaut syndrome (LGS) in our prior study. 537 patients with diverse EE phenotypes were screened. **Results**: We identified 16 patients with *de novo* mutations in 7 genes, thus establishing a genetic diagnosis for ~3% of our cohort. Among these are recurrent and novel mutations in *ALG13*, *CACNA1A*, *DNM1*, *GABRB3*, *GNAO1*, *IQSEC2*, and *SLC1A2* highlighting the importance of these genes in EE. Notably, recurrent mutations accounted for 44% of the pathogenic variants identified in this study. For two independent families with multiple affected individuals, we identified a parent with a mosaic germline mutation. *GABRB3* accounted for the majority of pathogenic variants (n=6/537), explaining ~1% of our cohort. We will report the frequency of *de novo* mutations for each gene screened in our cohort as well as investigate genotype-phenotype correlations for genes in which multiple patients harbor mutations. **Conclusion**: We have confirmed the role of at least 7 additional genes in the genetic etiology of EE and expanded the phenotypic spectrum associated with these genes beyond IS and LGS in which they were first discovered.

64

Loss-of-function mutations in *SLC12A5* encoding the potassium-chloride co-transporter KCC2 in epilepsy of infancy with migrating focal seizures. *T. Stödberg[1,2], A. McTague[3,4], A. Ruiz[5], H. Hirata[6,7,8], J. Zhen[9], P. Long[5], I. Farabella[10], E. Meyer[3], A. Kawahara[11], G. Vassallo[12], S. Stivaros[13,14], M. Bjursell[15], H. Stranneheim[15,16], S. Tigerschiöld[15,16], B. Persson[17,18], I. Bangash[19], K. Das[4,20], D. Hughes[21], N. Lesko[16,22], J. Lundeberg[23], R. Scott[4,24,25,26], A. Poduri[27,28], I. Scheffer[29,30], H. Smith[31], P. Gissen[31,32,33], S. Schorge[34], M. Reith[9,35], M. Topf[10], D. Kullmann[34], R. Harvey[5], A. Wedell[15,16], M. A. Kurian[3,4].* 1) Department of Women's and Children`s Health, Karolinska Institutet, SE-171 76, Stockholm, Sweden; 2) Neuropediatric Unit, Karolinska University Hospital, SE-171 76 Stockholm, Sweden; 3) Molecular Neurosciences, Developmental Neurosciences Programme, UCL Institute of Child Health, London, WC1N 1EH, U.K; 4) Department of Neurology, Great Ormond Street Hospital, London, WC1N 3JH, U.K; 5) Department of Pharmacology, UCL School of Pharmacy, London, WC1N 1AX, U.K; 6) Department of Chemistry and Biological Science, Graduate School of Science and Engineering, Aoyama Gakuin University, Sagamihara, Kanagawa 252-5258, Japan; 7) Center for Frontier Research, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan; 8) PREST, Japan Science and Technology Agency, Tokyo, 102-0076, Japan; 9) Department of Psychiatry, New York University School of Medicine, New York, NY 10016, U.S.A; 10) Institute of Structural and Molecular Biology, Crystallography/Department of Biological Sciences, Birkbeck College, University of London, WC1E 7HX, U.K; 11) Laboratory for Developmental Biology, Graduate School of Medical Science, University of Yamanashi, Chuo, 409-3898, Japan; 12) Department of Neurology, Royal Manchester Children's Hospital, Manchester, M13 9WL, U.K; 13) Academic Department of Radiology, Royal Manchester Children's Hospital, Manchester, M13 9WL, U.K; 14) Imaging Science, School of Population Health, University of Manchester, Manchester, M13 9PL, U.K; 15) Department of Molecular Medicine and Surgery, Science for Life Laboratory, Center for Molecular Medicine, Karolinska Institutet, SE-171 76 Stockholm, Sweden; 16) Centre for Inherited Metabolic Disorders, Karolinska University Hospital, SE-171 76 Stockholm, Sweden; 17) Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-751 85 Uppsala, Sweden;; 18) Department of Medical Biochemistry and Biophysics,Science for Life Laboratory, Karolinska Institutet, SE-171 76 Stockholm, Sweden; 19) EEG Department, Royal Oldham Hospital, OL1 2JH, Oldham, Lancashire, U.K; 20) Young Epilepsy, RH7 6PW, Lingfield, Surrey, U.K; 21) Department of Molecular Neuroscience, UCL-Institute of Neurology, WC1N 3BG, London, U.K; 22) Department of Laboratory Medicine, Karolinska Institutet, SE-171 76 Stockholm, Sweden; 23) Science for Life Laboratory, School of Biotechnology, Royal Institute of Technology, SE-100 44 Stockholm, Sweden; 24) Department of Neurological Sciences, University of Vermont College of Medicine, Vermont, VT 05405, U.S.A; 25) Department of Paediatric Neurology, Fletcher Allen Health Care, Vermont, VT 05401, U.S.A; 26) Clinical Neurosciences, Developmental Neurosciences Programme, UCL Institute of Child Health, London, WC1N 1EH, London, U.K; 27) Department of Neurology, Boston Children's Hospital, Boston, Massachusetts, MA 02115, U.S.A; 28) Department of Neurology, Harvard Medical School, Boston, Massachusetts, MA 02115, U.S.A; 29) Departments of Medicine and Paediatrics, University of Melbourne, Austin Health and Royal Children's Hospital, Melbourne, Victoria, VIC 3052, Australia; 30) Florey Institute, Melbourne, Victoria, VIC 3010, Australia; 31) MRC Laboratory for Molecular Cell Biology, UCL, London, WC1E 6BT, U.K.; 32) Department of Metabolic Medicine, Great Ormond Street Hospital, London, WC1N 3JH, U.K.; 33) Genetics and Genomic Medicine, Institute of Child Health, UCL, London, WC1N 1EH, U.K.; 34) Department of Clinical and Experimental Epilepsy, UCL Institute of Neurology, London, WC1N 3BG, U.K.; 35) Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY 10016, U.S.A.

Epilepsy of infancy with migrating focal seizures (EIMFS) is an early infantile epileptic encephalopathy (EIEE). Previously known as migrating partial seizures in infancy (MPSI) it was first described by Coppola in 1995. EIMFS is a drug-refractory condition with frequent focal seizures, developmental regression and a characteristic ictal pattern on EEG with seizure activity migrating from one hemisphere to the other. Only recently several genes have been implicated as causes of EIMFS.The hyperpolarizing, inhibitory action of GABA depends on low intraneuronal chloride levels. The potassium-chloride co-transporter KCC2, encoded by *SLC12A5* [MIM:606726], is the main chloride extruder in neurons and maintains the hyperpolarizing gradient. KCC2 polymorphisms have been implicated as susceptibility factors for febrile seizures and generalized epilepsy.We identified two families with EIMFS. Whole exome sequencing in one family and autozygosity mapping followed by exome sequencing in another left *SLC12A5* as the only strong candidate gene. Two affected children in a Swedish family were compound heterozygotes for missense mutations c.1277T>C (L426P) and c.1652G>A (G551D) and two affected children in a family of Pakistani origin with consanguineous first cousin parents were homozygous for the missense variant c.932T>A (L311H). All three mutations showed appropriate segregation in the families and were not present in healthy control populations. Protein homology studies were performed and predicted damaging effects of all three mutations.We used an *in vitro* heterologous expression system for immunoblotting and surface protein biotinylation. When compared to WT, the mutants L311H, L426P, and G551D showed reduced expression at the cell surface as well as reduced glycosylation. In addition in HEK293 cells transfected with WT and mutant FLAG-tagged KCC2 all three mutants were much less detected at the cell surface. Thus all three mutations seem to impair cell surface localization and post-translational modification of KCC2.To further investigate disease mechanisms we recorded in voltage-clamp mode from HEK293 cells transfected with a plasmid either encoding the wild-type KCC2b isoform or one of the mutants L311H, L426P, G551D. The KCC2 mutants showed a depolarized chloride reversal potential and a slowed recovery after chloride load proving a defective KCC2 transporter activity.Our report is the first to describe *SLC12A5* mutations as a monogenic cause of human epilepsy..

65

Functional analysis of *GRIN2A* mutations in childhood epileptic encephalopathies. *L. Addis[1, 2], L.R. Vidler[2], D.A. Collier[2], D.K. Pal[1], D. Ursu[2].* 1) Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom; 2) Neuroscience Discovery, Eli Lilly Research Centre, Windlesham, Surrey, United Kingdom.

Epileptic encephalopathies are severe brain disorders characterized by seizures and abundant epileptiform activity that contribute to cognitive and behavioural impairments. Landau-Kleffner syndrome (LKS) and continuous spikes and waves during slow-wave sleep (CSWS) are closely related encephalopathies with regression in language and global cognitive skills respectively. They show electroclinical overlap with Rolandic epilepsy (RE), the most frequent childhood focal epilepsy, forming a clinical spectrum of epileptic, cognitive, language and behavioural disorders. Recently it was discovered that around 20% of cases in this spectrum are caused by mutations in the NMDA glutamate receptor *GRIN2A*. Here we set out to determine the disease mechanism of twelve missense *GRIN2A* mutations. Mutations were chosen based on predicted functional changes, segregation with disease, amino acid conservation and disease severity. Mutations were inserted into *GRIN2A* cDNA, and HEK cells transiently transfected with mutant *GRIN2A*, and standard *GRIN1* constructs allowing formation of heteromers. 1) Western blotting of total protein lysates revealed that eight mutations caused a decrease in GRIN2A protein levels. The loss of the disulphide-bond of the cysteine residue in C436R and C231Y drastically destabilises the protein and causes its degradation, resulting in 70% loss. P79R, G483R, E714K and D731N destabilising amino acid changes also resulted in >50% loss. M705V and I814T caused a 30% decrease in expression. 2) Single-cell and high-throughput calcium imaging assayed glutamate binding, maximum response to glutamate, ratio of cell responses and NMDAR function. Mutations C436R and D731N totally abolished glutamate binding, whilst P79R, C231Y and G483R severely decreased the glutamate affinity, meaning the mutant receptors can only be activated by higher concentrations of the agonist. The maximum response to glutamate was also decreased for C231Y, M705V and A716T. No alterations in glutamate binding were detected for E714K, D933N and N976S, whereas M705V and I814T appeared to increase glutamate affinity 3) Fluorescent imaging of cell surface and intracellular expression of the mutant proteins also revealed corresponding alterations. Taken together, these data suggest that mutations across *GRIN2A* affect the expression and function of the receptor in different ways, with the end result of altered NMDA receptor currents and neuronal excitability.

66

Hyperexcitability of neurons and cardiomyocytes in a mouse model of *SCN8A* epileptic encephalopathy. *J.L. Wagnon[1], C.R. Frasier[2], L.F. Lopez-Santiago[2], Y. Yuan[2], J. Hull[2], Y. Bao[2], M.H. Meisler[1].* 1) Department of Human Genetics, University of Michigan, Ann Arbor, MI; 2) Department of Pharmacology, University of Michigan, Ann Arbor, MI.

Early-infantile epileptic encephalopathy type 13 (EIEE13, OMIM # 614558) is caused by *de novo* missense mutations of the gene *SCN8A* encoding the voltage-gated sodium channel $Na_v1.6$. In addition to generalized tonic-clonic seizures, EIEE13 can include developmental delay, severely impaired motor control, and sudden unexpected death in epilepsy (SUDEP). The *de novo* mutation p.Asn1768Asp was identified in the first reported patient with EIEE13 (Veeramah et al, AJHG 2012). In a heterologous expression system, the mutation resulted in impaired inactivation of sodium current (I_{NA}) and elevated persistent current resulting in neuronal hyperexcitability. A knock-in mouse model carrying the patient mutation recapitulates seizures and SUDEP (Wagnon et al, HMG 2015). To understand the pathogenic mechanism of this mutation *in vivo*, we characterized the electrophysiological properties of neurons and cardiomyocytes from the knock-in mouse. Isolated sodium current (I_{NA}) was recorded from acutely dissociated CA3 hippocampal neurons using the standard whole-cell patch clamp techniques. Persistent I_{NA} density was elevated more than 2-fold in both excitatory pyramidal neurons and inhibitory bipolar neurons. Whole-cell patch clamp recordings from acute brain slices detected abnormal spontaneous firing of hippocampal neurons. In addition to its major site of expression in neurons, *SCN8A* is expressed at a much lower level in heart, where $Na_v1.6$ is localized to the t-tubules of ventricular myocytes. *In vivo* electrocardiogram (ECG) recordings detected a reduced heart rate and accelerated idioventricular rhythm in response to caffeine in the mutant mice. Acutely isolated ventricular myocytes exhibited an increased incidence of delayed afterdepolarizations. We observed prolongation of the early repolarization phase of the action potential and increased duration of calcium transients.Taken together our results indicate that the p.Asn1768Asp mutation causes hyperexcitability of both neurons and cardiomyocytes. The hyperactivity of hippocampal neurons is likely to contribute to seizures. The abnormalities in cardiomyocytes may increase susceptibility to arrhythmias and thus contribute to SUDEP. Increased persistent sodium current appears to underlie neuronal hyperexcitability and may play a role in the heart. Therapeutic agents that selectively inhibit persistent I_{NA} may become a valuable treatment option for prevention of seizures and SUDEP in patients with *SCN8A* mutations.

67

Genome-wide analysis of multi-ancestry cohorts identifies new loci influencing glaucoma-related endophenotypes. *A. Iglesias Gonzalez[1], H. Springelkamp[1,2], A. Mishra[3,4], T. Aung[5,6], C-Y. Cheng[5,6,7], J.E. Craig[8], C.J. Hammond[9,10], M. Hauser[11,12], A.W. Hewitt[13], R. Höhn[14,15], C.C.W. Klaver[1,2], A.J Lotery[16], D.A. Mackey[17], L.R. Pasquale[18], N. Pfeiffer[14], A.C. Viswanathan[19], J.L. Wiggs[18], T-Y. Wong[5,6,7], C.M. van Duijn[1], S. MacGregor[4], International Glaucoma Genetics Consortium.* 1) Department of Epidemiology, Erasmus MC, Rotterdam, Rotterdam, Netherlands; 2) Department of Ophthalmology, Erasmus MC, Rotterdam, Rotterdam, Netherlands; 3) Department of Complex Trait Genetics, VU University, Centre for Neurogenomics and Cognitive Research, Amsterdam, The Netherlands; 4) Statistical Genetics Group, QIMR Berghofer Medical Research Institute, Royal Brisbane Hospital, Brisbane, Australia; 5) Ophthalmology and Visual Sciences Academic Clinical Program, Duke-NUS Graduate Medical School, National University of Singapore, Singapore, Singapore; 6) Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore; 7) Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore; 8) Department of Ophthalmology, Flinders University, Adelaide, SA, Australia; 9) Department of Ophthalmology, King's College London, St. Thomas' Hospital, London, United Kingdom; 10) Department of Twin Research and Genetic Epidemiology, King's College London, London, UK; 11) Departments of Medicine, Duke University Medical Center, Durham, NC, USA; 12) Department of Ophthalmology, Duke University Medical Center, Durham, NC, USA; 13) Centre for Eye Research Australia (CERA), University of Melbourne, Royal Victorian Eye and Ear Hospital, Melbourne, Victoria, Australia; 14) Department of Ophthalmology, University Medical Center Mainz, Mainz, Germany; 15) Department of Ophthalmology, Inselspital, Bern, Switzerland; 16) Clinical Neurosciences Research Group, Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, University Hospital Southampton, Southampton, UK; 17) Centre for Ophthalmology and Visual Science, Lions Eye Institute, University of Western Australia, Perth, Australia; 18) Department of Ophthalmology, Harvard Medical School, Boston, MA and Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA; 19) NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK.

Purpose: Glaucoma is a progressive optic neuropathy and a leading cause of irreversible blindness worldwide. Genome-wide association studies (GWAS) imputed to Hapmap reference panels have identified over 70 loci associated with glaucoma-related traits.The aims of this study are to use the 1000 genomes (1000G) imputation to (1) identify low frequent variants associated with glaucoma endophenotypes and (2) investigate the pathways overlapping between the different endophenotypes.Methods: We conducted a meta-analysis of 19 GWAS that included 29,578 Europeans and 8,373 Asians. The outcomes included intraocular pressure (IOP), vertical cup-disc ratio (VCDR), cup area (CA) and disc area (DA). Genetic data were imputed using the 1000G (Phase 1 v3). We subsequently tested the effect of all genome-wide significant SNPs on primary open-angle glaucoma (POAG) in three independent case-control studies. Gene set enrichment analysis using the DEPICT framework and expression analysis in zebrafish were performed to determine pathways implicated by the identified loci, and to evaluate the biological context of our findings.Results: This meta-analysis identified a novel locus associated with IOP, nine loci with VCDR, five with CA and six with DA. Two loci were found overlapping between IOP and optic disc parameters (*ABO* and *ADAMTS8*), supporting a genetic overlap between these quantitative traits. Gene-based analysis identified one novel gene associated with IOP and two novel genes with DA. Eleven loci were associated with POAG, including a suggestive new association at *CDKN1A*. Enrichment analysis highlighted pathways involved in development and revealed a new pathway implicated in regulation of adiponectin and leptin levels, which are related to metabolic syndrome. Significance of metabolic-related pathways was driven by *CDKN2B* and *SIX6* among others. We used zebrafish to explore downstream effects of *SIX6* downregulation and found alterations in the levels of *CDKN2B* and *CDKN1A*. *PAX6* was associated with disc area and regulates developmental genes like *SIX6* and *ATOH7*, and is involved in glucose metabolism. Conclusions: We have identified 21 loci associated with multiple glaucoma endophenotypes. Of clinical relevance is a new metabolic pathway identified which relates glaucoma to metabolic syndrome. Zebrafish analyses support the role of glaucoma genes in the cell cycle and demonstrate an *in vivo* interaction of genes potentially involved in metabolism.

68

The genomic region harboring the type 2 diabetes presumed causal variant within *TCF7L2* forms long-range functional connections with *ACSL5*. *M.E. Johnson[1], Q. Xia[1], A. Chesi[1], B.T. Johnston[1], S. Lu[1], E.F. Rappaport[2], P. Huang[3], A.D. Wells[4,5], G.A. Blobel[3,4], S.F.A. Grant[1,4].* 1) Divisions of Human Genetics and Endocrinology, Children's Hospital of Philadelphia, Philadelphia, PA; 2) NAPCore, Children's Hospital of Philadelphia, Philadelphia, PA; 3) Division of Hematology, Children's Hospital of Philadelphia, Philadelphia, PA; 4) Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 5) Division of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA.

Genome wide association studies (GWAS) only report genomic signals associated with a given trait and not necessarily the precise localization of culprit genes. Chromatin conformation capture techniques can aid in the identification of causal genes by characterizing genomic regions that make physical contact with a GWAS-implicated locus. The strongest associated type 2 diabetes (T2D) locus reported to date, residing within *TCF7L2*, presents a particular opportunity for such analyses as previous studies point to the T allele of rs7903146 in intron 3 as the causal variant at this location.We carried out 4C-seq and Capture C in parallel libraries using the immediate sequence harboring rs7903146 as the bait to elucidate the genomic regions it interacted with. Given that TCF7L2 mediates cell specific regulation of proglucagon in the intestinal tract, which in turn cleaves to the insulinotropic hormone GLP-1, we employed the human colon mucosal epithelial NCM460 cell line for this effort.When overlapping both sets of data, the bait region consistently interacted with 5 promoters, of which 4 resided in the same topologically associating domain (TAD) as *TCF7L2,* namely - and in order of peak score strength - *ACSL5, HABP2, LOC143188* and *TDRD1* (the latter being in a separate sub-TAD); furthermore, we observed interaction within *TCF7L2* itself. An additional promoter was observed on chromosome 6 corresponding to *MMS22L*. We went on to precisely gene edit this genomic element using CRISPR/Cas9. Leveraging sgRNAs targeting flanking sites located both upstream and downstream of rs7903146, we generated constructs that successfully removed either a 66bp or 1.4kb genomic segment. Following mRNA gene expression analysis, we observed a particularly dramatic impact on *ACSL5* levels (approx. 30x decrease) in both deleted homozygous settings, and subsequent Western blotting revealed that protein levels were almost entirely ablated. Furthermore, the gene expression changes for *HABP2* were also very notable.Interestingly, *ACSL5* encodes 'acyl-CoA synthetase long chain family, member 5', an enzyme known to play a role in mammalian fatty acid metabolism. *HABP2* encodes 'hyaluronan binding protein 2' and has not been strongly implicated in metabolic processes previously.As such, our data point to the immediate genomic location harboring rs7903146 as being a putative locus control region for a number of genes playing a role in the pathogenesis of T2D, in particular *ACSL5*.

69

Genome-wide Association Studies Identify *RAB38* and *HS6ST1* Associated with Albuminuria in Diabetes. *A. Teumer[1,2] on behalf of the CKDGen Consortium.* 1) Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany; 2) DZHK (German Center for Cardiovascular Research), partner site Greifswald, Greifswald, Germany.

Elevated concentrations of albumin in the urine (albuminuria) are associated with an increased risk of kidney disease progression, end-stage renal disease (ESRD) as well as cardiovascular events and mortality. Albuminuria is a hallmark of diabetic kidney disease (DKD), the leading cause of ESRD. No novel effective therapies for DKD have been approved in the past two decades.To gain insight into the pathophysiological mechanisms underlying albuminuria, we conducted a meta-analysis of 21 genome-wide association studies and an independent replication in 9 studies of urinary albumin-to-creatinine ratio (UACR) in individuals of European ancestry with (n=7,370) and without (n=46,061) diabetes. We characterized novel findings experimentally using *Rab38* knockout, transgenic and congenic Fawn Hooded Hypertensive (FHH) rats with induced diabetes.We identified and replicated associations between variants in *HS6ST1*, encoding heparan sulfate 6-O-sulfotransferase 1, and near *RAB38*/*CTSC*, encoding ras-related protein Rab-38 and dipeptidyl peptidase 1, respectively, with UACR among >7,000 individuals with diabetes. The change in average UACR per minor allele was 22% for *HS6ST1* and 14% for *RAB38*/*CTSC* ($p<1E-6$ for both). The genetic variants showed an effect on UACR in individuals with diabetes but not in those without diabetes (p-interaction <1E-5), representing examples of gene-by-diabetes interactions in a complex phenotype. The *Rab38* knockout rats had significantly higher albumin excretion than controls ($p<0.001$) despite similar blood glucose concentrations. RNA-seq data from micro-dissected human kidney samples showed higher gene expression in human tubuli compared to glomeruli for both *RAB38* and *HS6ST1*. The difference between tubular and glomerular expression was more pronounced for *RAB38* ($p=1.1E-8$) than for *HS6ST1* ($p=0.015$).The genes identified highlight novel pathways influencing albuminuria in diabetes and may represent novel targets for its modulation.

70

Chromosome interaction analysis of risk loci in related autoimmune diseases reveals complex, long-range promoter interactions implicating novel candidate genes. *P. Martin[1], A. McGovern[1], G. Orozco[1], K. Duffus[1], A. Yarwood[1], S. Schoenfelder[2], N. Cooper[3], A. Barton[1,5], C. Wallace[3,4], P. Fraser[2], J. Worthington[1,5], S. Eyre[1].* 1) Arthritis Research UK Centre for Genetics and Genomics. Centre for Musculoskeletal Research. Institute of Inflammation and Repair. Faculty of Medical and Human Sciences. Manchester Academic Health Science Centre. The University of Manchester. Stopford Building, Manchester, UK; 2) Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, UK; 3) JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, NIHR Cambridge Biomedical Research Centre, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge Biomedical Campus, Cambridge, UK; 4) MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK; 5) NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester Foundation Trust, Manchester Academic Health Science Centre, Oxford Road, Manchester M13 9WL.

Genome wide association studies (GWAS) have been tremendously successful in identifying genetic variants associated with complex diseases; however, since the majority of these variants map to intergenic enhancer regions, their functional annotation has proved elusive. Linking these enhancers to their target genes is essential if we are to realise the full translational potential of GWAS. It is established that enhancer regions regulate expression of their target genes through physical interaction via chromatin looping. We have used Capture Hi-C to investigate the potential interactions between disease-associated variants of four autoimmune diseases (rheumatoid arthritis, type 1 diabetes, psoriatic arthritis and juvenile idiopathic arthritis) and their functional targets in two relevant cell lines (B and T cells).

Uniquely, we designed two complementary experiments, the first using enrichment baits for disease associated regions and the second, baits for all promoters within 1Mb of associated variants, providing independent validation of captured interactions. We identified 8,594 interactions (764 fragments) in the first capture and 18,285 (1,938 fragments) in the second at 5% FDR which included well-established interactions in control regions (e.g. *HBA* locus). Unexpectedly, around 80% of significant interactions occurred at distances exceeding 500kb, whilst these could not be validated in this experiment it does reinforce the idea of long-range gene regulation. Among the 146 validated interactions we found evidence that firstly, SNPs associated with different autoimmune diseases, separated by distances of up to 1Mb, interact with each other and the same promoter suggesting common autoimmune gene targets (e.g. *PTPRC*, *DEXI, ZFP36L1)*. Secondly, that GWAS associated SNPs within enhancers interact with compelling candidate genes (e.g. *FOXO1*, *AZI2*), often situated several megabases away, whilst skipping closer genes more likely to be considered as functionally associated, strongly suggesting that SNPs uncovered by GWAS cannot simply be assigned to the nearest genes. Finally, long-range chromosomal interactions are often detected in a cell type specific manner. These results provide new insights into complex disease genetics and change the way we view the causal genes in disease.

71

A rare *A2ML1* variant confers susceptibility to otitis media and causes changes in the middle ear microbiome. *R.L.P. Santos-Cortez[1], C.M. Chiong[2,3], M.R.T. Reyes-Quintos[2,3], M.L.C. Tantoco[2], X. Wang[1], A. Acharya[1], I. Abbe[1], A.P. Giese[4], J.D. Smith[5], E.K. Allen[6,7], B. Li[1], E.M. Cutiongco-de la Paz[8,9], M.C. Garcia[3], E.G.D.V. Llanes[2,3], P.J. Labra[3], N.J. Ajami[10], J.F. Petrosino[10], G.T. Wang[1], K.A. Daly[11], J. Shendure[5], M.J. Bamshad[5], D.A. Nickerson[5], J.A. Patel[12], S. Riazuddin[4], M.M. Sale[6,7,13], T. Chonmaitree[12], Z.M. Ahmed[4], G.T. Abes[2,3], S.M. Leal[1], University of Washington Center for Mendelian Genomics.* 1) Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA; 2) Philippine National Ear Institute, University of the Philippines Manila - National Institutes of Health, Manila, Philippines; 3) Department of Otorhinolaryngology, University of the Philippines College of Medicine - Philippine General Hospital, Manila, Philippines; 4) Department of Otorhinolaryngology Head & Neck Surgery, School of Medicine, University of Maryland, Baltimore, Maryland, USA; 5) Department of Genome Sciences, University of Washington, Seattle, Washington, USA; 6) Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, Virginia, USA; 7) Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, USA; 8) Institute of Human Genetics, University of the Philippines Manila - National Institutes of Health, Manila, Philippines; 9) Department of Pediatrics, University of the Philippines College of Medicine - Philippine General Hospital, Manila, Philippines; 10) Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, USA; 11) Department of Otolaryngology, Head and Neck Surgery, University of Minnesota, Minneapolis, Minnesota, USA; 12) Division of Pediatric Infectious Disease and Immunology, Department of Pediatrics, University of Texas Medical Branch, Galveston, Texas, USA; 13) Department of Public Health Sciences, University of Virginia, Charlottesville, Virginia, USA.

Otitis media causes significant morbidity in a large proportion of the global population and is a major cause of hearing loss across all age groups. Strong evidence exists for genetic susceptibility to otitis media, but no rare variants have been previously associated with otitis media. In a large indigenous Filipino pedigree with ~50% prevalence of otitis media, exome sequencing and linkage analysis led to the identification of a duplication variant *A2ML1* c.2478_2485dupGGCTAAAT (p.(Ser829Trpfs*9)) segregating with nonsyndromic otitis media, with a significant maximum LOD score of 7.5 at reduced penetrance. The same duplication was identified in 3 out of 79 European- or Hispanic-American (EA/HA) otitis-prone children. All three children required tympanostomy tube insertion by age 6 months. On the other hand, the duplication was absent in 84 EA/HA non-otitis-prone children, in 1,378 exomes of EA/HA descent, and in 67,630 European non-Finnish and 11,606 Latino alleles from the Exome Aggregation Consortium database (p=3.34x10^{-14}). Both indigenous Filipino and EA/HA individuals with the duplication share a short founder haplotype that is estimated to be 1,800 years old. Seven additional *A2ML1* variants were identified in six otitis-prone children, and none of these variants were identified in non-otitis samples. A2ML1 is localized to middle ear mucosal epithelium and is 59% similar to alpha-2-macroglobulin, a marker of vascular permeability of middle ear mucosa during infection. Six of the identified A2ML1 variants are predicted to remove or affect function of the thiol-ester and receptor-binding domains, which are important for protease trapping and lysosomal clearance. To further determine if *A2ML1* variants cause shifts in microbial profiles, middle ear samples were collected from 16 indigenous Filipino individuals with otitis media, of whom 11 carry the *A2ML1* duplication. 16S rRNA gene profiling revealed bacteria that are rarely reported from otitis media samples. Differences in microbial profiles were detected based on carriage of the *A2ML1* duplication. *Haemophilus* is more abundant in otitis media individuals who are wild type, while *Fusobacterium* has greater abundance in carriers of the *A2ML1* duplication. In conclusion, we report identification of rare *A2ML1* variants that confer susceptibility to otitis media based on both family and association studies, including a duplication variant that induces changes in the middle ear microbiome.

72

Systems Genetics Approach Unravels the Molecular Mechanisms Underlying Lung Function Variation and Uncovers Novel Therapeutic Targets for COPD. *M. Obeidat[1], Y. Nie[1], K. Hao[2], Y. Bossé[3,4], M. Laviolette[4], D. Nickle[5], D. Postma[6], W. Timens[7], S. Gharib[8,9], L. Wain[10], M. Artigas[10], M. Tobin[10], I. Hall[11], S. London[12], D. Sin[1], P. Paré[1], SpiroMeta, CHARGE and Lung eQTL Consortia.* 1) Centre for Heart Lung Innovation, University of British Columbia, Vancouver, BC, Canada; 2) Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, USA; 3) Department of Molecular Medicine, Laval University, Québec, Canada; 4) Institut universitaire de cardiologie et de pneumologie de Québec, Laval University, Québec, Canada; 5) Merck Research Laboratories, Boston, MA, USA; 6) University of Groningen, University Medical Center Groningen, Department of Pulmonology, GRIAC research institute, Groningen, The Netherlands; 7) University of Groningen, Groningen, University Medical Center Groningen, Department of Pathology and Medical Biology, GRIAC research institute, Groningen, The Netherlands; 8) Center for Lung Biology, University of Washington, Seattle, Washington, USA; 9) Department of Medicine, University of Washington, Seattle, Washington, USA; 10) Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, UK; 11) Division of Therapeutics and Molecular Medicine, The University of Nottingham, Nottingham, UK; 12) Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina, USA.

Background: The SpiroMeta-CHARGE consortium undertook the largest GWAS for forced expiratory volume in one second (FEV_1) and its ratio to forced vital capacity (FEV_1/FVC) in ~48,000 individuals. We hypothesized that a subset of SNPs associated with FEV_1 and FEV_1/FVC act as expression quantitative trait loci (eQTL) to change the level of mRNA in lung tissue. Mapping lung function eQTLs will identify genes and pathways regulating lung function and uncover therapeutics targets for chronic obstructive pulmonary disease (COPD).**Methods**: Lung eQTLs were derived from genome-wide genotyping and gene expression analysis of 1,111 non-tumor lung tissue samples from UBC, Laval and Groningen universities. The eQTL study identified ~470,000 SNPs related to the level of gene expression in *cis* (within 1 Mb from gene) and ~17,000 SNPs in *trans*. We integrated SNPs associated with either FEV_1 or FEV_1/FVC at $P < 0.001$ from the SpiroMeta-CHARGE GWAS with eQTLs (at 10% FDR) to identify lung function eSNPs. Lung tissue mRNA levels of lung function eSNP- regulated genes (LFERG) were tested for association with COPD in the eQTL study (n=330 cases, n=428 controls) using logistic regression adjusted for age, gender, and smoking status, and results were used as input for Connectivity Map (https://www.broadinstitute.org/cmap/) to identify drugs/compounds that could reverse or augment the COPD gene signature.**Results**: For FEV_1, 3419 *cis* -eSNPs mapping to 271 genes, and 1568 *trans*-eSNPs mapping to 29 genes were identified. For FEV_1/FVC, 2214 *cis*-SNPs mapping to 275 genes and 442 *trans*-eSNPs mapping to 21 genes were identified. These figures represent a significant over-representation for eQTLs; enrichment ranged from 2.5 fold for *cis* to 32 fold for *trans* acting eSNPs. The identified genes were enriched in biological processes related to lung development and inflammatory processes. The lung mRNA levels of 51 of LFERG were associated with COPD at P<0.05. Connectivity Map (Cmap) analysis using these 53 COPD signature genes identified a number of compounds that reversed the COPD gene signature, including adiphenine; a local anesthetic which also acts as a nicotinic receptor antagonist.**Conclusion**: Systems genetics translated GWAS into genes and pathways underlying impaired lung function. *In silico* drug repurposing analysis identified adiphenine as a candidate drug that reverses the COPD associated gene signature. Future *in vitro* and *in vivo* studies are warranted to validate the finding.

73

Fine-mapping and molecular assays identify multiple functional variants at the *ANGPTL8* HDL-C GWAS locus. *M.E. Cannon[1], C.K. Raulerson[1], Q. Duan[1], Y. Wu[1], A. Ko[2], P. Pajukanta[2], M. Laakso[3], Y. Li[1,4], K.L. Mohlke[1].* 1) Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; 2) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; 3) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 4) Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA.

Recent European, African American, and Mexican genome-wide association studies (GWAS) have identified different lead variants (P<5E-8) associated with high-density lipoprotein cholesterol (HDL), located in or near the *ANGPTL8* gene (rs737337, rs4804154, rs2278426). *ANGPTL8* expression is limited to liver and adipose, despite being contained within one intron of ubiquitously expressed *DOCK6.* A rare coding variant in *ANGPTL8* is associated with HDL, and *ANGPTL8* knockout studies suggest a role in metabolism. In a preliminary analysis of RNA-seq data from a combined 22 adipose samples heterozygous for rs737337 and coding variant rs2278426 (*ANGPTL8* R59W), we observed an allelic expression imbalance (binomial P<0.05), consistent with a regulatory effect; alleles associated with decreased HDL appear to be associated with decreased *ANGPTL8* expression. To fine map the *ANGPTL8* locus, we examined the linkage disequilibrium (LD) structure of the HDL-associated variants in Europeans and African Americans. Given the extensive sharing of GWAS loci across populations, we assume the presence of at least one shared functional variant. Interestingly, the African American signal is wider than in Europeans, due to a ~20% frequency haplotype absent in Europeans. Using a typical LD threshold of r^2>0.8 (1KG) to define candidate variants, 4 exist in Europeans (EUR r^2 with rs737337), whereas 29 exist in Africans (AFR r^2 with rs4804154), and only two are shared. Using a threshold of r^2>0.5, 14 are shared, including missense variant rs2278426, which may contribute to the signal. To determine drivers of *ANGPTL8* tissue specificity, we tested the *ANGPTL8* promoter and 8 HDL-associated variants that overlap epigenetic marks indicative of regulatory regions in transcriptional reporter assays. The 400-bp promoter demonstrated transcriptional activity in liver cells (5-fold increase in HepG2) but not other cell types, suggesting the promoter is critical to expression in liver. Of the 8 candidate variants, 6 showed increased transcriptional activity (3-20 fold) in adipose and/or liver cells and modest allelic differences (1.2-1.5 fold), suggesting that additional drivers exist, especially in adipocytes. Electrophoretic mobility shift assays showed that 4 variants alter transcription factor binding. Despite different lead variants, the HDL association at *ANGPTL8* can be explained by a single shared regulatory signal with a complex mechanism including multiple potentially functional variants.

74

Functional analyses of type 2 diabetes-associated loci provides mechanistic insight into genetic susceptibility. *E.A. O'Hare, L.M. Yerges-Armstrong, J.A. Perry, A.R. Shuldiner, N.A. Zaghloul.* Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

Genetic susceptibility plays an important role in type 2 diabetes (T2D), but a clear understanding of how genes mediate disease onset has been elusive. To address this problem, we examined the role of individual genes in a single aspect of T2D etiology: production and maintenance of pancreatic β-cells. In light of the depletion of β-cell mass that often accompanies T2D, we hypothesized that genetic susceptibility to T2D may be mediated by the loss of β-cell mass. Using a transgenic zebrafish line allowing for *in vivo* visualization of β-cells, we systematically suppressed 67 identified zebrafish orthologs for genes at T2D-associated human loci and assessed defects in the production of β-cells. We found that 25 were necessary for proper production of β-cell mass at embryonic stages ($p \leq 1x10^{-20}$). Of these, 16 were novel for roles in β-cell production (e.g., *ankrd55*, *fitm2*, *klf14*, *klhdc5*, *pepd*, *slc30a8*, *thada*, *zfand3*). We used this systematic approach to functionally interrogate multi-gene associated genomic loci and identified single genes at five of six loci for which multiple candidates were tested. Deficits in embryonic production of β-cells point to the possibility that β-cell mass adaptive capacity over time may be impaired. We tested this possibility by examining the response of β-cells to either glucose-induced expansion or regeneration following β-cell ablation in animals depleted of T2D gene expression. These assays revealed significant roles in regeneration of β-cells for 17 of the 25 genes, suggesting that genetically programmed deficiencies in β-cell mass may be directly related to impaired maintenance. Finally, we investigated the relevance of our findings to human T2D onset in diabetic individuals from the Old Order Amish (n=45). Risk alleles in the 67 genes were classified as either necessary for β-cell mass (n=23) or not (n=37). We found that risk alleles in β-cell mass genes were significantly associated with younger age of onset (p=0.03). In addition, these risk alleles were enriched (1.4-fold; $p \leq 0.01$) in young, lean diabetics, suggesting that individuals with deficiencies in the functions of these genes may be predisposed to developing the disorder in the presence of fewer additional risk factors. Taken together, our study offers proof-of-principle for large-scale investigation of T2D gene function *in vivo.* These findings suggest that mediation of β-cell production may be a mechanism of genetic susceptibility to T2D for certain genotypes.

75

Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve replication rates in genome wide association studies. *Q. Lu[1], R. Powles[2], Q. Wang[2], J. He[3], H. Zhao[1,2,4].* 1) Department of Biostatistics, Yale University, New Haven, CT; 2) Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT; 3) Division of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT; 4) VA Cooperative Studies Program Coordinating Center, West Haven, CT.

Functionally annotating the human genome is a major goal in human genetics. Extensive efforts have been made to understand genomic function through both experimental and computational approaches, yet proper annotation still largely remains a challenge, especially in non-coding regions. Cell-type-specific high-throughput epigenetic data provided by large consortia, e.g. Roadmap Epigenomics Project, make it possible to systematically predict functional modules in the human genome in a tissue-specific manner. We have developed an unsupervised learning framework to predict tissue-specific functional regions through integrating diverse types of epigenetic annotations, and provide extensive case studies to demonstrate the biological insights it could bring to the broad scientific community. We first validate our method's ability to predict tissue-specific functionality using experimentally validated studies of non-coding elements. Our method is able to successfully identify a variety of non-coding regulatory machinery including enhancers, regulatory miRNA, and hypomethylated transposable elements. Next, we used tissue-specific functional annotations to prioritize SNPs in genome-wide association studies (GWAS). Brain-specific functional predictions most effectively increase the signal replication rate of schizophrenia GWAS data, while several other tissue types also show superior performance compared to using p-values only. Finally, we calculate the GWAS signal enrichment in various tissue types for 14 different human traits and diseases, and identify the highly relevant tissues for each genome-wide significant locus. Brain and blood were found to be the most relevant tissue types for schizophrenia in both global and local analyses. Examining coronary artery disease, heart-specific functional regions showed the strongest overall enrichment, but most genome-wide significant loci were found to be most functional in the gastrointestinal system, suggesting not only the large effect sizes of variants functional in the gastrointestinal system, but also a substantial proportion of undetected heart-related loci. In summary, our tissue-specific functional annotations can guide genetic studies at multiple resolutions and provide valuable insights in post-GWAS prioritization and disease etiology studies.

76

A General Analytical Framework to Identify Pathogenic Genes Underlying Complex Diseases. *P. Shooshtari[1,2], C. Cotsapas[1,2].* 1) Broad Institute of MIT and Harvard, Cambridge, MA; 2) Department of Neurology, Yale University, New Haven, Connecticut.

Genome-wide association studies have identified thousands of loci mediating risk to common, complex diseases, and we and others have shown that the majority of these effects map to regulatory DNA. These bulk analyses support a model where disease risk is driven by alterations to gene regulation. We have developed a framework to identify (a) specific regulatory elements driving disease risk; (b) the genes these elements are regulating; and (c) the cell types in which this regulation occurs. Unlike previous bulk analyses, we are thus able to identify pathogenic genes and thus uncover mechanisms of pathogenesis.Our approach is to identify regulatory features harboring causal disease variants, and through these the disease-causing gene(s) in each risk locus. We first construct a map of regulator:gene relationships across the genome using matched expression and regulatory region assays (initially, paired expression and DNase I Hypersensitivity site data from 22 cell types in the Roadmap Epigenomic Project). Uniquely, we identify robustly detected regulatory regions across samples. Only ~54% of peaks pass (covering 8% of the genome compared to 14% by all peaks), but explain the disease heritability attributable to all peaks, suggesting the non-replicating peaks are spurious. We then associate them to transcript levels to find genes under their control, and overlay likely causal variants from GWAS loci onto this genome-wide landscape of gene regulation. We combine the disease and regulatory association statistics into a joint likelihood that any gene is causal for disease, which we assess by permutation.We applied this approach to 97 loci from a large fine-mapping study of multiple sclerosis. 50/97 loci had at least one peak harboring a candidate causal variant. 133/997 genes in these loci associated to regulatory regions harboring risk variants ($p < 0.05$). For 121/133 genes we are able to localize disease risk to a single regulatory region, with the remainder showing evidence of complex regulatory alterations in pathogenesis. The latter include known pathogenic genes with altered regulation: STAT4, CD40 and CD58. We found that risk-mediating DHS are significantly enriched in activity in $CD3^+$, $CD8^+$, $CD4^+$, $CD19^+$, $CD14^+$ and fetal thymus cell populations ($p < 0.05$ after Bonferroni correction). We are currently extending this approach to a further five autoimmune and inflammatory diseases and will present the overall results of regulatory perturbation in disease.

77

metaCCA: Summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *A. Cichonska[1,2], J. Rousu[2], P. Marttinen[2], A.J. Kangas[3,4], P. Soininen[3,4], T. Lehtimäki[5], O.T. Raitakari[6,7], M.R. Järvelin[8,9,10], V. Salomaa[11], M. Ala-Korpela[3,4,12], S. Ripatti[1,13,14], M. Pirinen[1].* 1) Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland; 2) Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland; 3) Computational Medicine, Institute of Health Sciences, University of Oulu and Oulu University Hospital, Oulu, Finland; 4) NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland; 5) Department of Clinical Chemistry, Fimlab Laboratories, University of Tampere School of Medicine, Tampere, Finland; 6) Department of Clinical Physiology and Nuclear Medicine, University of Turku and Turku University Hospital, Turku, Finland; 7) Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku and Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland; 8) Department of Epidemiology and Biostatistics, MRC Health Protection Agency (HPA) Centre for Environment and Health, School of Public Health, Imperial College London, United Kingdom; 9) Institute of Health Sciences, University of Oulu, Finland; 10) Biocenter Oulu, University of Oulu, Finland; 11) National Institute for Health and Welfare, Finland; 12) Computational Medicine, School of Social and Community Medicine and the Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; 13) Public Health, University of Helsinki, Finland; 14) Department of Human Genetics, Wellcome Trust Sanger Institute, United Kingdom.

A dominant approach to genome-wide association studies (GWASs) is to perform univariate tests between genotype-phenotype pairs. However, analysing related traits together results in increased statistical power and certain complex associations become detectable only when several variants are tested jointly. Currently, modest sample sizes of individual cohorts and restricted availability of individual-level genotype-phenotype data across the cohorts limit conducting multivariate tests. Here, we introduce *metaCCA*, a computational framework for multivariate analysis of a single or multiple GWASs based on univariate regression coefficients. To our knowledge, it is the first summary statistics-based approach that allows multivariate representation of both phenotype and genotype.*metaCCA* extends the statistical technique of canonical correlation analysis to the setting where the original individual-level data are not available. Instead, *metaCCA* operates on three pieces of the full data covariance matrix: *Cxy* of univariate genotype-phenotype association results, *Cxx* of genotype-genotype correlations, and *Cyy* of phenotype-phenotype correlations. *Cxx* is estimated from a reference database matching the study population, e.g. the *1000Genomes*, and *Cyy* is estimated from *Cxy*. We employ a covariance shrinkage algorithm to add robustness to the method.A multivariate meta-analysis of two Finnish studies of nuclear magnetic resonance metabolomics (total *n*=7092) by *metaCCA,* applied to standard univariate output from the program SNPTEST, shows an *excellent agreement with the pooled individual-level multivariate analysis* of original data sets. Root mean squared error (RMSE) between *metaCCA*'s -log10 p-values and original ones is 0.02 when 455,521 SNPs from chromosome 1 are tested for an association with a cluster of 9 correlated lipid measures, and 0.45 when also genotypes are treated multivariately.We demonstrate the potential of multivariate methods with several known lipid genes. For example, moving from the univariate analyses of 9 lipid measures to the multivariate test changes the top p-value from 10^{-10} to 10^{-24} for *CETP,* and from 10^{-9} to 10^{-12} for *APOE*. Motivated by these results, we envisage that the multivariate association testing using *metaCCA* has a great potential to provide novel insights from already published summary statistics of large GWAS meta-analyses.

78

A systematic analysis of differential pathway architectures in diverse functional genomics networks for large-scale prediction of pathways from GWAS and exome-sequencing projects. *J.D. Mercer[1], S. Rosenbluh[1,4], A. Liberzon[1], J. Grabarek[1], D. Thompson[1], T. Eisenhaure[1,5], S. Carr[1], J. Jaffe[1], J. Boehm[1], A. Tsherniak[1], A. Subramanian[1], R. Narayan[1], T. Natoli[1], T. Liefeld[1], B. Wong[1], J. Bistline[1], T. Li[1,3,12], S. Calvo[1,7], Y. Li[1,6,7], J. Mesirov[1,8], N. Hacohen[1,9], A. Regev[1,10], K. Lage[1,2,3,11].* 1) Broad Institute, Cambridge, MA; 2) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, USA; 3) Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts, USA; 4) Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA; 5) Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Boston, Massachusetts, USA; 6) Department of Statistics, Harvard University, Cambridge, Massachusetts , USA; 7) Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA; 8) University of California, San Diego, California, USA; 9) Center for Immunology and Inflammatory Diseases and Division of Rheumatology, Allergy, and Immunology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA; 10) Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; 11) Harvard Medical School, Boston, Massachusetts, USA; 12) Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

High-throughput technologies in genomics, genetics, epigenetics, transcriptomics, and proteomics have led to the generation of heterogeneous biological networks that connect genes if they are functionally correlated in any of the aforementioned data types. These networks share global design features by being scale-free, small world, and modular and have the potential to catalyze genomic interpretation, systems biology, and therapeutic intervention. However, it remains challenging to systematically map and understand how biological signal is organized within and between networks. We design a statistical method based on machine learning to test 18 architectural metrics across 1,592 pathways in gene networks of correlated mRNA expression, phylogenetic patterns, cancer synthetic lethality relationships, cell perturbation profiles, and protein-protein interactions. We show that pathway architectures diverge significantly between networks, and between classes of pathways (e.g., cell signaling, metabolism, and cell regulation) within each network illustrating that despite similar global designs, pathway relationships are differentially organized in heterogeneous networks. We provide a web platform (GeNets) for exploiting network-specific pathway architectures to optimize biological discovery and for the scientific community to compare, visualize, and share genome-scale networks through a standardized framework. Finally, we use GeNets to systematically test and visualize non-obvious pathway relationships based on genetic variants discovered in hundreds of GWAS and exome sequencing projects and make the resulting interactive networks available to the genetics community through our web platform.

79

Identifying genetically-driven clinical phenotypes using linear mixed models. *J.D. Mosley[1], J.S. Witte[2], S.J. Hebbring[3], E.K. Larkin[1], L. Bastarache[4], C.M. Shaffer[1], J.H. Karnes[1], C.M. Stein[1], J.C. Denny[1,4], D.M. Roden[1].* 1) Department of Medicine, Vanderbilt University, Nashville, TN; 2) Department of Epidemiology and Biostatistics, University of California, San Francisco, CA; 3) Biomedical Informatics Research Center, Marshfield Clinic Research Foundation Marshfield, WI; 4) Biomedical Informatics, Vanderbilt University, Nashville, TN.

Electronic medical records (EMR) linked to genetic data are increasingly used to discover and validate genotype-phenotype associations. With the large numbers of phenotypes available for analysis, selecting those with important genetic components is a key first step in EMR-based discovery. We hypothesized that genetic linear mixed models (GLMMs), which quantify the additive phenotypic genetic variation attributable to a collection of common single nucleotide polymorphisms (SNPs), could be used to prioritize a large set of EMR-derived phenotypes for genetic studies. We used GLMMs to estimate the genetic liability for 1,309 binary clinical phenotypes extracted from a set of 29,349 EMRs of unrelated European ancestry subjects with SNP genotyping on the Illumina Exome Beadchip. Phenotype factors predictive of a highly significant genetic liability estimate were (1) an autoimmune etiology or (2) replicating a SNP association reported in the NHGRI GWAS Catalog, consistent with the large representation of SNPs in the immune-related human leukocyte antigen (HLA) region and of SNPs reported in the NHGRI GWAS Catalog on this platform. Phenotypes representing sequelae or subtypes of GWAS Catalog phenotypes had lower liability estimates. SNP variation around the HLA region was associated with 44 phenotypes with FDR $q<0.05$, including five not previously reported in the GWAS catalog: sicca syndrome, cholangitis, dermato/polymyositis, polymyalgia rheumatica and dermatophytosis. Among the HLA-associated phenotypes, there were significant (FDR $q<0.1$) genetic correlations between Type I diabetes and both celiac disease ($p=7\times10^{-5}$) and hypothyroidism ($p=3\times10^{-3}$) and between juvenile rheumatoid arthritis and primary biliary cirrhosis ($p=3\times10^{-3}$). We investigated the hypothyroidism and polymyalgia rheumatica phenotypes by a SNP association study to further delineate their HLA signals. We identified novel, replicating SNP associations with hypothyroidism near *HLA-DQA1/HLA-DQB1* at rs6906021 (OR=1.2 [95% CI:1.1-1.2], $p=9.8\times10^{-11}$) and with polymyalgia rheumatica near *C6orf10* at rs6910071 (OR=1.5 [95% CI 1.3-1.6], $p=1.3\times10^{-10}$). Association testing for hypothyroidism using imputed HLA alleles demonstrated a pattern of association consistent with an HLA-DR3 haplotype. In summary, phenome-wide application of GLMMs can identify phenotypes with important genetic drivers, and focusing on these phenotypes can lead to new discovery of variants associated with clinical disease.

80

Genetic validation and application of pathway-based annotation for unknown signals in untargeted metabolomics. *Y.H. Hsu[1,2,3], T. Esko[2,3,4], T.H. Pers[2,3], A. Metspalu[4], J.N. Hirschhorn[1,2,3].* 1) Department of Genetics, Harvard Medical School, Boston, MA; 2) Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA; 3) Broad Institute of the MIT and Harvard, Cambridge, MA; 4) Estonian Genome Center, University of Tartu, Tartu, Estonia.

Metabolomics, or the global profiling of small molecules in biological samples, has become a powerful approach for discovering biomarkers for traits and diseases. Genome-wide association studies have shown that many metabolites have strong associations with SNPs (metabolic quantitative trait loci, mQTL). Current untargeted profiling methods can measure thousands of signals in a single experiment. While computational tools have been developed to match these signals to known metabolites, many signals remain as unknowns after this identification step and are usually excluded from further analysis.We have developed a method to annotate unknown signals from untargeted metabolite profiling data and demonstrated that this method can be used to extract biologically meaningful information from the unknown signals. We previously performed untargeted LC-MS (liquid chromatography followed by mass spectrometry) profiling of plasma samples from 100 normal weight, 100 lean, and 100 obese individuals drawn from the extremes of a population of 50,000 individuals. We calculated pairwise correlations between the profiled signals (335 known metabolites, 16,607 unknowns) across the 300 samples. Starting with 1,649 predefined metabolic pathways from the ConsensusPathDB database, we reconstituted these pathways by using correlations between signals to compute a membership score for each signal in each pathway. To validate and test our annotation approach, we performed pathway enrichment analyses of SNP- or trait-associated signals, using the computed pathway membership scores. As a proof of principle, we identified 71 unknown signals that are associated (p < 1e-5) with a known mQTL in the *SLCO1B1* gene. We observed that these signals are enriched (p < 1e-10) in several pathways related to transport of organic anions and bile acid metabolism, consistent with previous literature evidence for this gene. Next, we identified signals whose plasma levels differed between lean and obese individuals, and showed that they are enriched (empirical p < 1e-3) in many metabolic pathways, including those related to amino acid metabolism, neurotransmission, lipoprotein-mediated lipid transport, vitamin digestion and absorption, and energy metabolism. Overall, these results indicate that our method for annotating unknown signals can be a useful tool for gaining biological insights from the large amount of data generated in untargeted metabolomics studies.

81

Quantifying Uncertainty in Heritability Estimation using Linear Mixed Models. *R. Schweiger[1], E. Halperin[1,2,3].* 1) The Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel; 2) Molecular Microbiology and Biotechnology Department, Tel-Aviv University, Tel-Aviv, Israel; 3) International Computer Science Institute, 1947 Center St., Berkeley, CA 94704, USA.

Linear mixed models (LMMs) are commonly used in several key areas of genetics. They have been instrumental in genome-wide association studies and in heritability estimation. A key component of LMMs is the estimation of variance components - the proportion of phenotypic variance explained by individ- uals' genetic similarity vs. non-genetic environmental factors. State-of-the-art methods use the Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) approaches. Currently, estimates for variance components as well as their confidence intervals rely on asymptotical properties. However, these assumptions are often violated, due to the bounded parameter space, dependencies, and limited sample size, leading to biased estimates, and inflated or deflated confidence intervals and p-values.

Here, we show that ML and REML estimates may be biased, and that often the probability that the genetic component is estimated as 0 is high even when the true heritability is bounded away from 0. We demonstrate how these biases depend on the kinship matrix. We also show that the estimation of confidence intervals by state-of-the-art methods is incorrect, especially when the true heritability is either relatively low or relatively high. We show that such biases are present in estimates of heritability of gene expression in GTEX and in other genotype data. We discuss how common practices on genotype data, such as the inclusion of principal components as fixed effects, may affect these biases. We propose a parametric bootstrap method to construct rigorous confidence intervals for variance components, p-value calculations, and for heritability estimates. Our method can be used as an add-on to existing methods for heritability and variance components estimation, such as GCTA, Fast-LMM, GEMMA, or EMMA. .

82

Genotype imputation with millions of reference samples. *B.L. Browning[1,2,3], S.R. Browning[2].* 1) Dept of Medicine, Div of Medical Genetics, University of Washington, Seattle, WA; 2) Dept of Biostatistics, University of Washington, Seattle, WA; 3) Dept of Genome Sciences, University of Washington, Seattle, WA.

The pace of whole-genome sequencing is accelerating, and tens of thousands of whole genome sequences are now being produced each year. This makes it possible to assemble reference panels of unprecedented size that can be used to accurately impute very low frequency variants. We present a new genotype imputation method that is highly accurate and that scales to reference panels with millions of individuals. The new imputation method, implemented in Beagle v4.1, is highly parallelized and has extremely low memory requirements, making it ideally suited to analyze large data sets using multi-core computer processors.We compare Beagle with the published Minimac2 and Impute2 executables on the chromosome 20 data from the 1000 Genomes and UK10K projects, and on 10Mb of simulated European data. All three methods impute into pre-phased samples and have very similar accuracy on these data, but there are significant differences in memory use and computation time that affect scalability. With 50,000 reference samples and short 2 Mb analysis windows, Minimac2 required 43 GB or memory and Impute2 required 110 GB of memory for single-threaded execution. These memory requirements prevent Minimac2 from using more than one core on computers with 64 GB memory and Impute2 from using more than one core on computers with 128 GB memory. In contrast Beagle required only 12 GB of memory to analyze the entire 10 Mb region with 12 parallel threads. Compute time summed over all threads for Beagle was 45x less than Minimac2 and 15x less than Impute2. Wall-clock computation time for Beagle was 175x less than Minimac2 and 165x less than Impute2 due to Beagle's ability to use all computer processing cores. We show that this new imputation method scales to much larger reference panels by performing imputation using a simulated reference panel having 2 million samples and a mean variant density of one variant per 8 bp.We estimate that with sequence data for 1 million reference samples and 1M SNP array data for 1000 target samples, Beagle v4.1 can accurately impute minor allele dose ($r^2 \geq 0.8$) in variants down to minor allele frequency 0.00001 at a computational cost of approximately $1 per imputed sample.

83

Germline Mutations in Cancer-predisposition Genes in 1,120 Children with Cancer: a Report from the Pediatric Cancer Genome Project. *J. Zhang[1,2], M. Walsh[2,3], G. Wu[1,2], M. Edmonson[1,2], T. Gruber[2,3,4], J. Easton[1,2], D. Hedges[1], X. Ma[1,2], X. Zhou[1], D. Yergeau[1,2], M. Wilkinson[1], B. Vandor[1,2], X. Chen[1,2], R. McGee[2,3], S. Hines-Dowell[2,3], R. Nuccio[2,3], E. Quinn[2,3], S. Shurtleff[2,3], M. Rusch[1,2], J. Becksfort[1,2], M. Weaver[1,2], L. Ding[5,6], E. Mardis[5,6], R. Wilson[5,6], C.-H. Pui[2,3], A. Gajjar[2,3], D. Ellison[2,4], A. Pappo[2,3], K. Nichols[2,3], J. Downing[2,4], St. Jude/Washington University Pediatric Cancer Genome Project.* 1) Computational Biology, St Jude Children's Research Hospital, Memphis, TN; 2) Pediatric Cancer Genome Project, St Jude Children's Research Hospital, Memphis, TN; 3) Department of Oncology, St Jude Children's Research Hospital, Memphis, TN; 4) Department of Pathology, St Jude Children's Research Hospital, Memphis, TN; 5) Department of Genetics, Washington University School of Medicine; 6) McDonnell Genome Institute, Washington University School of Medicine.

The incidence of germline mutations in cancer predisposition genes in pediatric cancer patients is largely unknown and these findings are of critical importance for therapeutic management and genetic counseling of patients and family members. To characterize the incidence and spectrum of predisposing mutations in children with cancer, we analyzed the germline genomes of 1,120 pediatric cancer patients using whole-genome sequencing (n=595), whole-exome sequencing (n=456), or both (n=69) as part of the St. Jude/Washington University Pediatric Cancer Genome Project. A total of 5,463 non-silent, non-polymorphic variations (single-nucleotide variants, small insertions/deletions, and copy-number variations) were detected across 565 cancer genes, including 60 clinically relevant autosomal dominant (AD) and 29 autosomal recessive (AR) cancer predisposition genes. Preliminary assessment of variant pathogenicity within these 89 AD/AR genes was carried out by an automated classifier based on interpretation in the context of 15 cancer or locus specific databases, *in silico* prediction, the presence of second-hits in the tumor genome, and the medical literature. Final classification was determined by panel review. Ninety-eight pathogenic (P) or likely pathogenic (LP) mutations were identified in 96 patients (8.6%) including 4 germline mosaic patients and 2 patients with bi-allelic mutations in *ATM* or *PMS2*. The most frequently mutated genes were *TP53* (n=50), *APC* (n=7), *BRCA2* (n=6), *NF1* (n=4), *PMS2* (n=4), *RB1* (n=3) and *RUNX1* (n=3). The incidence of P/LP germline mutations was highest in solid tumors (48 of 287, 16.7%) followed by brain tumors (22 of 245, 9.0%) and leukemia (26 of 588, 4.4%). Among the 58 patients with a P/LP mutation and family data, only 23 (40%) had a positive cancer history, suggesting that family history alone is insufficient to predict the presence of an underlying predisposition syndrome. An additional 35 patients carried protein truncation mutations in other AR or tumor suppressor genes (TSG), with *CHEK2* (n=4), *PML* (n=4) and *BUB1B* (n=3) being the most frequent recurrent mutated TSG. Our novel analytical approach enabled characterization of predisposing gene mutations in 8.6% of pediatric cancer patients. The mutation data identified in this study inform clinical management and can be explored via our newly developed pediatric cancer data portal (http://pecan.stjude.org), which serves as a resource for the research community.

84

Germline variants among unselected patients enrolled in a tumor/normal cancer genomic sequencing project identifies a high percentage with inherited risk. *V.M. Raymond[1], J.N. Everett[1], E.M. Stoffel[1], J.W. Innis[3], Y.M. Wu[4], D.R. Robinson[2, 4], P. Vats[4], R.J. Lonigro[4], R. Mody[3], A.M. Chinnaiyan[2, 4].* 1) Department of Internal Medicine, University of Michigan, Ann Arbor, MI; 2) Department of Internal Pathology, University of Michigan, Ann Arbor, MI; 3) Department of Pediatrics, University of Michigan, Ann Arbor, MI; 4) Michigan Center for Translational Pathology, University of Michigan, Ann Arbor.

Introduction: The MI-Oncoseq project is a tumor-normal whole genome sequencing project with the goal of identifying therapeutic targets. We aimed to understand the germline variant spectrum in study participants.Methods: Adults (age 18+) with refractory tumors and children (≤ age 25) with all tumor types are study eligible. Patients meet with a genetic counselor (GC) at the time of consent to the IRB-approved protocols for discussion of potential secondary germline findings and collection of a cancer focused pedigree. Germline results are reviewed for 161 genes involved in cancer pathways, which are categorized and reviewed by the study team, including a GC and Medical Geneticist. Category 1 includes genes associated with autosomal dominant, high risk cancer syndromes. Category 2 includes genes associated with autosomal dominant, moderate risk cancer syndromes. Autosomal recessive genes are grouped separately. Pathogenic/likely pathogenic (P/LP) variants in the 80 Category 1/2 genes are considered for disclosure. Results: Between August 2011 - October 2014, 463/500 enrolled patients (400 adult, 100 pediatric) completed sequencing (92.6%). 1594 total variants (insertions-deletions, single nucleotide variants, copy number variants), 631 unique, were identified in Category 1/2 genes in 431 patients (mean 3.7 variants/patient). 32 patients had no Category 1/2 gene variants. P/LP variants were identified in 20 Category 1 genes in 20 patients (APC, BRCA1/2, BRIP1, DICER1, FH, MLH1, MSH2, PDGFRB, SBDS, SMARCA4, TP53) and 23 Category 2 genes in 23 patients (APC p.I1307K, BAP1, BARD1, CHEK2, HOXB13, MITF, MYH, PALB2). One patient had P/LP variants in both categories. 255/395 Category 1/2 variants of uncertain significance (VUS) were not previously reported in public databases. 1159 variants were benign. Conclusion: Of the 1594 total Category 1/2 variants reviewed, 24.8% were VUS, of which 64.6% were not previously reported. This underscores the importance of collaborative data sharing and interpretation. 9.1% of patients with cancer referred to this tumor-normal sequencing project, unselected for family history, had a P/LP germline variant associated with a high/moderate cancer risk highlighting he importance of prior genetic counseling on the potential germline findings in tumor-normal sequencing. NIH Clinical Sequencing Exploratory Research (CSER) Award NIH 1UM1HG006508.

85

Somatic *TP53* mutations detected in germline testing: The importance of phenotypic correlation in cancer predisposition testing. *J.N. Weitzel[1], K.R. Blazer[1], H. LaDuca[2], B. Nehoray[1], T. Slavin[2], T. Pesaran[1], C. Rybak[1], I. Solomon[1], M. Neil-Swiller[1], E. Chao[2].* 1) Clinical Cancer Genetics/Population Sciences, City of Hope, Duarte, CA; 2) Amby Genetics, Aliso Viejo, CA.

Purpose: Analysis of DNA isolated from peripheral blood/saliva is typically used for diagnosis of hereditary cancer predisposition. Results from testing are accepted as representing a patient's germline, as acquired somatic mutations are rare. While somatic *TP53* mutations are detected in multiple cancer types, germline mutations are exceedingly rare and result in Li-Fraumeni Syndrome (LFS). *TP53* mutations are increasingly detected on multi-gene panels, across diverse patient scenarios, suggesting either a broader phenotype or possible detection of clonal populations with an acquired *TP53* mutation. This study evaluated whether somatic interference may be more common in genetic testing than previously anticipated.**Methods:** Among patients with pathogenic *TP53* results from multi-gene panel testing, cases were selected with potentially abnormal next-generation sequencing (NGS) metrics, including decreased ratio of mutant to wild-type allele, >2 detected alleles or haplotypes, or large copy-number alterations. Clinical data was obtained from test requisition forms and compared to LFS testing criteria (classic, Chompret, or BC<age 36 years).**Results:** Among 166 *TP53* positive cases, 25 were identified as higher risk for somatic interference based on abnormal NGS metrics. None of these families met Chompret or classic diagnostic LFS criteria. Four probands were diagnosed with breast cancer <36 y.o.; this is not significantly different from the testing cohort (n=21,306) as a whole (Fisher's exact test; p=0.16). To date testing additional tissues confirmed somatic origin for 4/25 cases; two were subsequently diagnosed with a hematologic disorder. Although this cohort was defined primarily based on abnormal NGS metrics, we also identified a 63 yo woman with lobular breast cancer who did not meet any LFS criteria, wherein the NGS metrics were unremarkable but subsequent testing, prompted by the clinician because of the discordant phenotype, identified a low-grade lymphoma and absence of the *TP53* mutation in DNA isolated from breast tissue. Investigation of additional cases is underway. **Conclusions:** We suggest that somatic *TP53* mutations in blood/saliva may be more common than previously thought. Beyond using NGS quality control measures, clinician recognition of test results inconsistent with a LFS phenotype should create an index of suspicion, and caution is urged in the medical management of patients in whom the only criterion for LFS is a *TP53* mutation.

86

ChIP-Seq analysis of lymphocytes from Li-Fraumeni patients reveals the drastic impact on p53 DNA binding of heterozygous *TP53* mutations associated with early-onset cancers. *T. Frebourg[1], Y. Zerdoumi[1], R. Lanos[1], A. Bouzelfen[2], F. Charbonnier[1], G. Bougeard[1], J-M. Flaman[1].* 1) Dept Genetics, Inserm U1079, Rouen Univ Hosp, Rouen, France; 2) Inserm U918, Rouen University, France.

Li-Fraumeni Syndrome (LFS), one of the most severe inherited forms of cancer, results from heterozygous germline mutations of the *TP53* gene, encoding a key transcriptional factor activated in response to DNA damage. We have recently shown, from a large LFS series including 415 *TP53* mutations carriers, that the most severe form of the disease, characterized by childhood tumors, is associated to missense mutations with negative dominant activity (Bougeard *et al*., Journal of Clinical Oncology 2015 in press). Thanks to a new p53 functional assay that we developed in lymphocytes, we showed that dominant-negative missense mutations drastically alter the transcriptome in response to DNA damage. To study, at the genome scale, the functional impact of heterozygous *TP53* mutations on p53 DNA binding, we set up a chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) analysis. After exposure of lymphocytes to doxorubicin, a powerful DNA damaging agent, chromatin was shared by sonication, immunoprecipitated with the anti-p53 DO-1 antibody, DNA fragments were sequenced on an Illumina HiSeq 2000 platform and the peaks, corresponding to the ChIP-enriched regions, were identified and quantified using the HOMER algorithm. ChIP-seq analyses of exposed control lymphocytes, with wild-type *TP53* genotype, accurately mapped 706 p53 binding sites. New p53 binding sites were validated using a gap repair *ADE2* reporter vector and a functional assay in yeast. ChIP-seq analysis of LFS lymphocytes with heterozygous *TP53* dominant-negative missense mutation reveals only 28 binding sites and, for these sites the depth of the corresponding peaks was reduced. Altogether, our results show for the first time that the clinical severity of *TP53* dominant-negative missense mutations is explained by a global alteration of p53 binding at the genome scale and support that LFS results from a defect of the transcriptional response to DNA damage.

87

Tumour risks and genotype-phenotype-proteotype analysis of 800 patients with germline mutations in the succinate dehydrogenase subunit genes *SDHB*, *SDHC* and *SDHD*. *E.R. Maher[1,9], K.A. Andrews[1,9], L. Vialard[2], D.B. Ascher[3], D.E.V. Pires[3,4], N. Bradshaw[5], T. Cole[2], F. Lalloo[6], M. McConachie[7], P.J. Morrison[8], V. Murday[5], S.M. Park[9], Y. Wallis[2], D. Goudie[7], R.S. Lindsay[5], C.G. Perry[5], L. Izatt[10], E.R. Woodward[2], SDH-UK Consortium.* 1) Department of Medical Genetics, University of Cambridge, UK; 2) West Midlands Regional Genetics service, Birmingham Women's Hospital, Birmingham, UK; 3) Department of Biochemistry, University of Cambridge, Cambridge, UK; 4) Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Brazil; 5) University of Glasgow, Glasgow, UK; 6) Department of Medical Genetics, Central Manchester Hospitals, Manchester, UK; 7) East of Scotland Regional Genetics Service, Ninewells Hospital and Medical School, Dundee, UK; 8) Department of Genetics, Northern Ireland Regional Genetic Service, Belfast City Hospital, UK; 9) Department of Clinical Genetics, Cambridge University Hospital, Cambridge, UK; 10) Clinical Genetics Department, Guy's Hospital, London, UK.

Germline mutations in the succinate dehydrogenase subunit genes *SDHB*, *SDHC* and *SDHD* are the most frequent causes of inherited phaeochromocytomas and paragangliomas. Since these genes were identified over a decade ago, genetic testing for mutations has become a standard investigation for many of these patients. However, insufficient information regarding incomplete penetrance and phenotypic variability hinders optimum management of mutation carriers. We undertook a retrospective survey of 800 individuals (401 previously reported) with germline mutations in *SDHB/C/D* (620 *SDHB,* 31 *SDHC* and 149 *SDHD*). In our UK-based cohort *SDHC* exon 6 deletions and *SDHD* exon 4 deletions were significantly more common than in other populations. Analysis of age-related tumour risks provided novel estimates of penetrance and genotype-phenotype correlations. *In silico* structural prediction analyses were performed to evaluate the functional effects of *SDHB* and *SDHD* mutations. In addition to tumour specific differences in risk for individual genes, we confirmed that the *SDHD* p.Pro81Leu mutation has a distinct phenotype and found evidence suggesting higher penetrance with *SDHB* p.Ile127Ser mutations. These genotype-phenotype associations were correlated with structural prediction studies. The penetrance in *SDHB* and *SDHD* mutation-positive non-probands by age 60 years was 21.9% [95% CI 15.3-28.0%] and 50.4% [95% CI 31.0-64.3%] respectively and the risk of malignant disease at age 60 years in non-proband *SDHB* mutation carriers was 3.8% [95% CI 0.8-6.7%]. Increased knowledge of the molecular basis of the phenotypic variability commonly observed in individuals with germline *SDHB/C/D* mutations will facilitate the development of personalised management based on gene and mutation-specific tumour risks.

88

A parent-of-origin effect impacts the phenotype in low penetrance retinoblastoma families segregating the p.Arg661Trp mutation of *RB1*. *C. Houdayer[1, 6, 7], P. Eloy[1], C. Dehainault[1], M. Sefta[2], I. Aerts[3], L. Lumbroso le Rouic[4], D. Stoppa Lyonnet[1, 6, 7], F. Radvanyi[2], G. Millot[5], M. Gauthier Villars[1].* 1) Dept Genetics, Institut Curie, Paris, France; 2) CNRS UMR144, centre de recherche de l'Institut Curie, Paris; 3) Departement d'oncologie pédiatrique, adolescents jeunes adultes, Institut Curie, Paris; 4) Departement d'oncologie chirurgicale, service d'Ophtalmologie, Institut Curie, Paris; 5) UMR 3244, centre de recherche de l'Institut Curie, Paris; 6) INSERM U830, centre de recherche de l'Institut Curie, Paris; 7) Université Paris Descartes, Sorbonne Paris Cité, Paris.

Retinoblastoma (Rb), the most common pediatric intraocular neoplasm, results from inactivation of both alleles of the *RB1* tumor suppressor gene. The second allele is most commonly lost, as demonstrated by loss of heterozygosity studies. *RB1* germline carriers usually develop bilateral tumors, but some Rb families display low penetrance and variable expressivity. In order to decipher the underlying mechanisms, 23 unrelated low penetrance pedigrees segregating the common p.Arg661Trp mutation and other low penetrance mutations were studied. In families segregating the p.Arg661Trp mutation, we demonstrated, for the first time, a strong correlation between the gender of the transmitting carrier and penetrance, as evidenced by Fisher's exact test: the probability of being unaffected is 90.3% and 32.5% when the mutation is inherited from the mother and the father, respectively (p-value = 7.10^{-7}). Interestingly, a similar correlation was observed in families segregating other low penetrance alleles.Consequently, we investigated the putative involvement of an imprinted, modifier gene in low penetrance Rb. We first ruled out a *MED4*-driven mechanism by *MED4* methylation and expression analyses. We then focused on the differentially methylated CpG85 island located in intron 2 of *RB1* and showing parent-of-origin-specific DNA methylation. This differential methylation promotes expression of the maternal p.Arg661Trp allele. We propose that the maternally inherited p.Arg661Trp allele retains sufficient tumor suppressor activity to prevent retinoblastoma development. In contrast, when the mutation is paternally transmitted, the low residual activity would mimic a null mutation and subsequently lead to retinoblastoma. This implies that the p.Arg661Trp mutation is not deleterious *per se* but needs to be destabilized in order to reach pRb haploinsufficiency and initiate tumorigenesis. We suggest that this phenomenon could be a general mechanism to explain phenotypic differences in low penetrance Rb families.

89

Germline and somatic inactivating *SMARCA4* mutations in small cell carcinoma of the ovary, hypercalcemic type (SCCOHT): diagnostic and therapeutic implications. *W.P.D. Hendricks[1], P. Ramos[1], H. Yin[1], A.N. Karnezis[2,3], Y. Wang[2,3], M.L. Russell[1], D.W. Craig[1], V.L. Zismann[1], A. Sekulic[4], B.E. Weissman[5], D.G. Huntsman[2,3], J.M. Trent[1].* 1) Division of Integrated Cancer Genomics and Division of Neurogenomics, Translational Genomics Research Institute (TGen), Phoenix, AZ; 2) Department of Pathology and Laboratory Medicine, The University of British Columbia, Vancouver, BC, Canada; 3) Centre for Translational and Applied Genomics, British Columbia Cancer Agency, Vancouver, BC, Canada; 4) Department of Dermatology, Mayo Clinic, Scottsdale, AZ, USA; 5) Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, NC, USA.

Small cell carcinoma of the ovary of hypercalcemic type (SCCOHT) is a rare, aggressive, and poorly-differentiated tumor, whose average age of diagnosis is 24yo versus 63yo for most ovarian cancers. Recent studies in our laboratory and others have now identified inactivating germline and somatic *SMARCA4* mutations and concomitant protein loss in ~90% of SCCOHT's (*Nature Genetics* 46:427,2014). We have also identified one SCCOHT tumor with intact SMARCA4 that instead harbored a homozygous inactivating mutation in *SMARCB1*, a SWI/SNF subunit commonly mutated in malignant rhabdoid tumors. The SMARCA4 alterations observed in SCCOHT tumors are characteristic of a tumor suppressor (e.g. typically bi-allelic, inactivating and LOH). Further, sequencing reports of 5 SCCOHT-affected families have shown segregation of *SMARCA4* mutations, and consistently younger age of disease onset in successive generations. Our recent results indicate that SCCOHT patients may share the same cancer predisposition syndrome as malignant rhabdoid tumors (MRT) and atypical teratoid/rhabdoid tumors (AT/RT), that is, germline mutation of a core SWI/SNF subunit, *SMARCB1* in MRTs and AT/RTs and *SMARCA4* in AT/RTs. Based on the early age at presentation, diploid cytogenetics, mutational spectra, and the presence of rhabdoid-like cells by histology, it has been proposed that SCCOHT represents another type of extra-renal rhabdoid tumor. Unifying the classification of these three tumors may have therapeutic implications. As in MRT, targeting epigenetic modifications may have therapeutic potential in SCCOHT. We have examined the SCCOHT cell lines BIN67 and SCCOHT1 against a panel of 14 epigenetic drugs that target all classes of regulators of histone modifications or DNA methylation. Unlike MRT, we found that SCCOHT cells are insensitive to EZH2 inhibition despite observing EZH2 overexpression in both cell lines. Also, as we expected, the absence of both SMARCA4 and SMARCA2 SWI/SNF ATPases in SCCOHTs renders the use of bromodomain inhibitors as a synthetic-lethality strategy for SMARCA4-deficient tumors ineffective. Of interest, BIN67 and SCCOHT1 were shown to be hypersensitive to the HDAC inhibitors romidepsin, quisinostat and panobinostat. We will present our progress to elucidate the cell of origin, identify therapeutic vulnerabilities (by siRNA and drug screening), and design initial clinical trials to further advance treatment options for patients with SCCOHT.

90

Common variants in *MMP20* at 11q22.2 predispose to 11q deletion and impact neuroblastoma risk. *X. Chang[1], L. McDaniel[2], C. Hou[1], M. Diamond[2], K. Thomas[1], J. Li[1], Y. Guo[1], F. Mentch[1], H. Qiu[1], C. Kim[1], S. Diskin[2,3], P. Sleiman[1,3,4], E. Attiyeh[2], J. Maris[2,3], H. Hakonarson[1,3,4].* 1) Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA; 2) Division of Oncology and Center for Childhood Cancer Research, The Children's Hospital of Philadelphia; Philadelphia, PA; 3) Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 4) Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA.

Neuroblastoma is a malignant tumor of the sympathetic nervous system presenting in early childhood. Somatically acquired chromosomal aberrations such as *MYCN* amplification (MNA), 17q gain, 1p deletion and 11q deletion are key prognostic factors of poor neuroblastoma outcome. Among them, MNA and 11q deletion have been shown to be almost mutually exclusive suggesting sub-groups of neuroblastoma classified by cancer genomics. Besides MNA, genes such as *ALK* (3.3%), *CCND1* (2.5%) and *LIN28B* (1.1%) have been shown to be targeted by somatic amplification in neuroblastoma primary tumors. To identify additional chromosomal structural changes in neuroblastoma, we genotyped 641 neuroblastoma tumor DNA samples, using the HH550/610Q SNP arrays from Illumina, including 442 samples with matched blood DNA samples for germline genome analysis. Tumor-based copy number segments were first calculated by ASCAT and OncoSNP genome wide, and then curated manually. Following thorough QC, 437 tumor samples were kept for analysis, including 300 with matched blood DNA genotyping data. Here, we report on novel focal amplifications of genes, *MYC*, *ZFHX3*, *KRAS*, *RRAS2* and *CYTH1*. Interestingly, we observed a considerable number of cases with low-level amplification of *CCND1* (45/434, 10.4%). In most of those cases, the amplification of *CCND1* co-occurred with 11q deletion (43/45, 95.6%), suggesting these two may be interacting. We further confirmed that the amplified genes were significantly overexpressed in neuroblastoma tumors and cell lines by examining gene expression data from 100 primary tumors and 29 cell lines. Given the intriguing phenomena that MNA and 11q deletion are almost mutually exclusive in neuroblastoma, we additionally conducted a genome wide association study (GWAS) in 120 European-American individuals with 11q deletion and 4,857 ancestry-matched controls. Our results demonstrate that common variants at 11q22.2 within *MMP20* are associated with neuroblastoma in patients with 11q deletion (rs3781788, $P=4.67 \times 10^{-9}$, OR=2.482), suggesting these variants impact neuroblastoma risk in 11q deleted patients.

91

Epigenetic and transcriptional dysregulation of Oxytocin receptor (Oxtr) in Tet1 methylcytosine dioxygenase deficient mouse brain. *A.J. Towers[1], XL. Li[2], A.L. Bey[3], P. Wang[2], S.K. Siecinski[1], S. Xu[2], X. Cao[2], L.J. Duffney[2], YH. Jiang[1,2,3].* 1) Program in Genetics and Genomics, Duke University, Durham, NC; 2) Pediatrics Dept, Duke University, Durham, NC; 3) Neurobiology Dept, Duke University, Durham, NC.

Oxytocin (OXT), the brain's most abundant neuropeptide, acts as a neuromodulator and a hormone with effects throughout the body. One major action is its activation of the brain's reward circuitry by increasing dopamine release from the ventral tegmental area to the ventral striatum, amygdala, and hippocampus. Oxytocin signaling is mediated by ligand binding to its receptor (OXTR), which is widely distributed throughout the limbic region and prefrontal cortex. OXT and OXTR modulate a variety of behaviors, including stress and anxiety responses, social memory and recognition, sexual and aggressive behaviors, and bonding and maternal behavior. Genetic and epigenetic studies have implicated OXT and OXTR in neuropsychiatric disorders, such as autism and anxiety. The mechanism underpinning the epigenetic regulation of *Oxtr*, however, is poorly understood. Recently, the methylcytosine dioxygenases of the Ten-eleven translocation (TET) family, including TET1, were suggested to play a role in DNA demethylation. We have generated and characterized *Tet1* knockout (-/-) mice to test the role of TET1-mediated DNA demethylation in gene expression and cognitive and social behaviors. RNA-seq expression profiling of *Tet1*-/- mouse hippocampus revealed downregulation of *Oxtr*, while qRT-PCR confirmed consistent downregulation in multiple brain regions of the adult *Tet1*-/- mouse. Conversely, *Oxtr* expression in *Tet1*-/- embryonic stem cells (ESCs) did not significantly differ from *Tet1*+/+ ESCs, suggesting a developmental window for the emergence of *Oxtr* downregulation. Interestingly, we discovered hypermethylation of the CpG island (CGI) located within *Oxtr* exon 3 in *Tet1*-/- mice. While CGI hypermethylation was not observed in ESCs, consistent with normal gene expression, hypermethylation was detected as early as E14.5. This suggests TET1 is necessary for preventing hypermethylation of *Oxtr* within the first few days post conception in mice. Consistent with the *Oxtr* results, we observed impaired maternal care in virgin *Tet1*-/- female mice, as evidenced by a longer latency to pup retrieval and less time spent huddling with the pups. This data supports a role for TET1 in regulating *Oxtr* expression through maintenance of appropriate levels of DNA methylation with implications on maternal behavior. This work offers novel insight into understanding the epigenetic regulation of *Oxtr* expression and supports further studies of epigenetic dysregulation in neuropsychiatric disorders.

92

GWAS meta-analysis identifies susceptibility loci for epigenetic age acceleration in human cerebellum. *A. Lu[1], E. Hannon[2], M. Levine[1], K. Hao[3], E. Crimmins[4], A. Kozlenkov[5], K. Lunnon[2], J. Miller[2,6], S. Dracheva[5], S. Horvath[1,7].* 1) Human Genetics, UCLA, Los Angeles, CA; 2) University of Exeter Medical School, Exeter, UK; 3) Department of Genetics and Genomic Sciences, The Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY; 4) Davis School of Gerontology, University of Southern California,Los Angeles, CA; 5) Department of Psychiatry, The Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY; 6) Institute of Psychiatry, King's College London, UK; 7) Biostatistics, School of Public Health, UCLA, Los Angeles, CA.

DNA methylation (DNAm) levels are particularly promising biomarkers of aging since chronological age has a profound effect on DNAm levels in most human tissues and cell types. Several recent studies support measuring accelerated aging effects using DNAm levels. Earlier, we developed epigenetic clock" (based on 353 CpG markers) to estimate DNAm levels with a noteworthy aspect of its broad applicability to most human cell types, tissues, and organs. Evidence has been accumulating that the epigenetic clock is related to chronological age more strongly than previously proposed biomarkers, and it appears to be capturing aspects of the biological aging process by its ability to predict aging-related outcomes such as all-cause mortality. However, the biological and genetic determinants of epigenetic clock remain elusive to explain its differentiation from chronological age, suggesting studies to identify associated genes and biological pathways.Here we aim to identify the genetic variants associated with epigenetic age acceleration in human cerebellum (CRBLM). To define age acceleration, we used the residuals resulting from a linear regression model of DNAm age on chronological age. Thus this measure has no correlation with chronological age. In total, we gathered 555 European-ancestry individuals across 5 studies available for SNP data and DNA methylation profiled in CRBLM tissues. We performed GWAS for each study on ~5.7 genotyped or imputed SNPs based on the 1000 Genome haplotypes then combined the results with fixed-effect models weighted by inverse variance. A total of 5 SNPs were identified at genome-wide significance ($P < 5.0 \times 10^{-8}$) and 234 at suggestive levels ($P < 1.0 \times 10^{-5}$). Strikingly, the whole GWAS signals highly significantly overlapped with age-related diseases including Alzheimer's disease ($P = 4.4 \times 10^{-15}$) and age-related macular degeneration ($P = 6.4 \times 10^{-6}$). In parallel with the GWAS-meta analysis, we conducted transcriptomic meta-analysis to profile ~ 2% genes mostly correlated with age acceleration. Functional annotation analysis suggests those genes are enriched with Polycomb-Group Protein targets in stem cells ($P = 1.42 \times 10^{-15}$), the genes bound by RNA polymerase II in stem cells ($P = 2.10 \times 10^{-25}$), and the genes annotated by human brain transcriptome ($P < 5 \times 10^{-7}$). In conclusion, our study provides a new biological insight to study aging and suggests that epigenetic age acceleration is under genetic control.

93

Methylomic aging as a window on lifestyle impact: tobacco and alcohol alter the rate of biological aging. *M.V Dogan[1,2], S.R.H Beach[3,4], M.K. Lei[4], C.E. Cutrona[5], M. Gerrard[6], F.X. Gibbons[6], R.L. Simons[4,7], G.H. Brody[4], R.A. Philibert[1,2].* 1) Biomedical Engineering, University of Iowa, Iowa City, IA; 2) Psychiatry, University of Iowa, Iowa City, IA; 3) Psychology, University of Georgia, Athens, GA; 4) Center for Family Research, University of Georgia, Athens, GA; 5) Psychology, Iowa State University, Ames, IA; 6) Psychology, University of Connecticut, Storrs, CT; 7) Sociology, University of Georgia, Athens, GA.

The rate of biological aging varies from person to person and is influenced by lifestyle choices. Since cigarette smoking and excessive alcohol consumption are leading causes of morbidity and mortality worldwide, our study was designed to understand the association between substance use and biological aging using DNA methylation based indices in two independent cohorts. The first cohort consisted of 656 European Americans whose average age was 63.4 years and the second cohort consisted of 180 African American individuals whose average age was 48.9 years. Genome-wide DNA methylation was assessed using the Illumina HumanMethylation 450k BeadChip and these data were obtained from the public database, Gene Expression Omnibus. In order to circumvent using unreliable self-reported substance use data, we instead utilized DNA methylation biomarkers to quantify cigarette and alcohol consumption. Specifically, cg05575921 from the *AHRR* gene and cg23193759 found in chromosome 10 open reading frame 35, C10orf35, were used as biomarkers for cigarette and alcohol consumption, respectively. Demethylation at these loci is observed with increasing substance use. We used the epigenetic "clock" consisting of 71 DNA methylation loci to examine the association between these DNA methylation biomarkers to deviation of biological age from chronological age. All levels including low levels of smoke exposure were associated with accelerated biological aging. In contrast, a mixed effect was observed for alcohol consumption. Moderate use of alcohol was associated with healthy aging. This study demonstrates the potential utility of methylomic indices of biological aging as a tool in elucidating the impact of lifestyle on health.

94

A survey of DNA methylation polymorphism in the human genome identifies environmentally responsive co-regulated networks of epigenetic variation. *R. Joshi, P. Garg, C. Watson, A. Sharp.* Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY.

While studies such as the 1000 Genomes Projects have resulted in detailed maps of genetic variation in humans, to date there are no robust maps of epigenetic variation. We have defined sites of common epigenetic variation, that we term Variable Methylation Regions (VMRs). To avoid the confounder of cellular heterogeneity, we utilized Illumina 450k DNA methylation data from populations of five purified cell types: T-cells, B-cells, fibroblasts, neurons and glia (n=58 to 111 individuals). Using a robust approach we identify hundreds of VMRs in each cell type that show common variability in DNA methylation levels in the normal population.We find that VMRs occur preferentially at enhancers and in 3' UTR regions, consistent with a role in regulating expression. We observed that at the majority of VMRs methylation is highly heritable. However, we also observed a subset of VMRs distributed across the genome that show highly correlated variation in *trans*, and form co-regulated networks. These VMRs tend to have low heritability, differ between cell types and are enriched for specific biological pathways of direct functional relevance to each tissue. For example, in T-cells we defined a network of 61 co-regulated VMRs enriched for genes that play roles in T-cell activation; in fibroblasts a network of 21 co-regulated VMRs comprising all four *HOX* gene clusters that is highly enriched for control of tissue growth; and in glia a network of 66 VMRs enriched for roles in postsynaptic membrane organization. These VMR networks are significantly enriched for TF binding sites, indicating that the network epigenetic state is responsive to molecular signaling cascades. By culturing genetically-identical fibroblasts under varying conditions of nutrient deprivation and cell density, we experimentally demonstrate that some VMR networks are responsive to environmental conditions, with methylation levels at these loci changing in a coordinated fashion in *trans* dependent on cellular growth. Intriguingly these environmentally-responsive VMRs showed a 47-fold enrichment for imprinted loci ($p<10^{-94}$), including the differentially methylated regions associated with 9 imprinted genes, suggesting that imprinted loci are particularly sensitive to environmental conditions. Our study provides the first detailed map of common epigenetic variation in the human genome, showing that both genetic and environmental causes underlie this variation.

95

A novel predictive model of sexual orientation using epigenetic markers. *T.C. Ngun[1], W. Guo[2], N.M. Ghahramani[3], K. Purkayastha[1], D. Conn[4], F.J. Sanchez[5], S. Bocklandt[1], M. Zhang[2,6], C.M. Ramirez[4], M. Pellegrini[7], E. Vilain[1].* 1) Department of Human Genetics, David Geffen School of Medicine at University of California Los Angeles (UCLA), Los Angeles, CA, USA; 2) Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST, Tsinghua University, Beijing 100084, China; 3) Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; 4) Fielding School of Public Health, UCLA, Los Angeles, CA, USA; 5) Department of Counseling Psychology, The University of Wisconsin-Madison, WI, USA; 6) Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080, USA; 7) Department of Molecular, Cellular, and Developmental Biology, UCLA, Los Angeles, CA, USA.

Sexual orientation is one of the most pronounced sex differences in the animal kingdom. Although upwards of 95% of the general population is heterosexual, a small but significant proportion of individuals (3-5%) is homosexual. Male sexual orientation has been linked to several genomic loci, with Xq28 and 8p12 being the most replicated. As with other complex traits, environmental factors may also play an important role. Firstly, monozygotic twins show substantial levels of discordance for this trait. Secondly, each male pregnancy a woman has increases the chance that her next son will be homosexual by 33% (the fraternal birth order effect). Thirdly, early life androgen exposure in women is associated with increased rates of non-heterosexual identity. Taken together, the evidence suggests a role for non-genetic and, possibly, epigenetic influences on sexual orientation. Our aim in this study was to create a predictive model for sexual orientation using epigenetic markers. We created our model based on genome-wide DNA methylation patterns in 37 monozygotic male twin pairs that were discordant for sexual orientation. 10 monozygotic twin pairs concordant for homosexuality were included as a control population. Genomic sites where methylation occurred were consolidated into short regions based on proximity and correlation of their methylation patterns to increase the signal to noise ratio. We then applied the FuzzyForest algorithm to our dataset. Briefly, regions were clustered into modules using Weighted Gene Coexpression Network Analysis and recursive feature elimination was performed with the random forest algorithm (RF) to identify regions most relevant to sexual orientation. The highest prediction accuracy was achieved using information from just 9 regions. Some of these regions were associated with the regulatory domains of two genes, CIITA and KIF1A. The former is a transcriptional regulator that is sometimes referred to as the master control factor of class II major histocompatibility complex genes. The latter is a neuron-specific transport protein that is important for movement of synaptic vesicle precursors along axons. Our results demonstrate that studies of the epigenome can yield new insights into the biological underpinnings of sexual orientation and provide strong support to the hypothesis that epigenetics is involved in sexual orientation. To our knowledge, this is the first example of a biomarker-based predictive model for sexual orientation.

96

RNF12 is essential for X-inactivation in female mouse embryonic stem cells, is required for female mouse development, and might be a target for future therapies to treat X-linked disorders in females: evidence from a mouse knockout model. *S. Barakat[1,2], J. Gribnau[2].* 1) MRC Centre for Regenerative Medicine, University of Edinburgh, Edinburgh, United Kingdom; 2) 1Department of Reproduction and Development, Erasmus MC, University Medical Center, Rotterdam, The Netherlands.

X-chromosome inactivation (XCI) in females is a crucial mechanism which equalizes X-linked gene-dosage between both sexes. In mice, a first wave of imprinted XCI occurs during cleavage stages of embryonic development, followed by X-chromosome reactivation (XCR) at the pre-implantation blastocyst, and subsequent random XCI (rXCI) in the post-implantation epiblast. rXCI can be simulated in differentiating mouse ES cells. We have previously shown that the X-encoded RNF12 protein act as a dosage-sensitive XCI-activator (PMID:19945382, PMID:21298085, PMID:24613346). When RNF12 becomes up-regulated during differentiation, it targets the pluripotency factor Rex1 for proteasomal degradation (PMID:22596162). As Rex1 is a repressor of the non-coding Xist RNA, which is crucial for initiating chromosome-wide gene-silencing, down-regulation of Rex1 by RNF12 allows female-specific Xist-expression and XCI-initiation in a stochastic manner (PMID:20083102). Here we present the generation and analysis of a novel *Rnf12* knockout mouse model, and provide evidence that RNF12 is also crucial for iXCI and rXCI *in vivo*. Whereas *Rnf12*-/Y males are viable, *Rnf12*-/- female mice fail to undergo XCI leading to lethality at post-implantation. *Rnf12*-/+ animals inheriting the maternal knockout allele are lethal due to silencing of the paternal *Rnf12* allele upon iXCI. In addition, RNF12 stored in the oocyte and produced *de novo* from both the maternal and paternal derived X chromosome influence XCI kinetics in pre-implantation embryos. *Rnf12*+/- females inheriting the paternal knockout allele are healthy but show an XCI-defect, with adult cells displaying XCR. Bi-allelic expression of X-linked transcripts is observed in various tissues, but is compensated by an XCI-independent form of dosage compensation. This peculiar finding, together with our recent results on XCR in human induced pluripotent stem cells (PMID: 25640760), opens a new area of research which might lead to novel approaches for treating X-linked diseases, such as Rett syndrome, in females.

97

Deep learning the relationship between chromatin architecture, chromatin state and transcription factor binding. *A. Kundaje[1,2], C.S. Foo[2], J. Israeli[3], A. Shrikumar[2], J. Buenrostro[1], A. Schep[1], W. Greenleaf[1].* 1) Dept. of Genetics, Stanford University, Stanford, CA; 2) Dept. of Computer Science, Stanford University, Stanford, CA; 3) Biophysics Program, Stanford University, Stanford, CA.

Assays such as DNase-seq and MNase-seq that profile genome-wide chromatin accessiblity and nucleosomal patterns have allowed comprehensive identification of regulatory elements (REs) and characterization of their local chromatin architecture. Different classes of active and repressed REs such as promoters, enhancers and insulators have been found to be associated with distinct combinations of histone modifications defining their chromatin state. Multivariate hidden Markov models are typically used to learn chromatin states from multiple histone modification ChIP-seq experiments for discovery and annotation of diverse REs. However, histone ChIP-seq experiments are time-consuming, costly and require large amounts of input material; limiting their applicability in rare cell types. In this study, we decipher a predictive relationship between chromatin architecture and chromatin state at REs and leverage it to simultaneously predict histone modifications, chromatin states and transcription factor (TF) binding from a single low-cost assay known as ATAC-seq. ATAC-seq simultaneously profiles accessibility, nucleosomes and TF footprints at REs from low input samples based on direct in vitro transposition of sequencing adaptors into native chromatin. We train novel deep learning methods based on convolutional neural networks on a novel two-dimensional representation of ATAC-seq data to learn a direct mapping from chromatin architecture to chromatin state by leveraging subtle patterns in insert-size distributions. Using a multi-task, multi-modal formulation we integrate ATAC-seq data, DNA-shape and DNA-sequence to simultaneously predict multiple histone modifications, chromatin state and TF binding with high accuracy (80-90%). Models trained on a combination of DNase-seq and MNase-seq data also achieve similar high accuracy supporting a fundamental predictive mapping between chromatin architecture and chromatin state. We develop novel feature visualization methods to peer into the deep neural networks and identify interesting architectural features such as TF footprints that are automatically learned from raw data. We explore the feasibility of cross-cell type prediction and determine the minimum sequencing depth requirements for predictive power. In conclusion, our method enables characterization of REs from low quantities of input material using a single assay, potentially enabling detailed regulatory maps in rare cell populations in primary tissue.

98

Inter-planetary Systems Biology Reveals Differences in Twin Astronauts at the Genetic, Epigenetic, Transcriptional, and Epitranscriptomic Levels. *C. Mason.* Physiology and Biophysics, Weill Cornell Medical College, New York, NY.

Next-generation sequencing (NGS) technologies have recently been used to detect and classify epigenetic changes at an unprecedented level. Epigenetic patterns help define cellular phenotypes and can even be used to more accurately sub-classify human tumors such as acute myeloid leukemias (AML). Indeed, since some phenotypes such as cancer and aging can be better differentiated with epigenetic signatures (vs. purely genetic), and epigenetic profiling is the focus of large-scale NGS efforts at the NIH like the Epigenomics Roadmap. There is also evidence that epigenetic changes accompany DNA damage and that sites of methylation correspond to DNA damage sites. Finally, there is an emerging field of "epitranscriptomics," where modifications of RNA, like the epigenetic modification of DNA, have been shown to impact brain development, gene regulation, and protein translation rates. Such ongoing research in DNA and RNA methylation has shown that these mediators of gene expression and modification represent critical regulatory layers in cell biology and genetics. Yet, these key epigenetic mechanisms of cellular regulation have barely, or never, been studied in the context of human space travel or during various gravity transitions ("G-transitions"). While previous work has addressed gene expression and proteomic dynamics during space travel, here we report data from the NASA Twins Study, regarding two identical twins, which provides an extraordinary opportunity to finally examine the impact of space travel on epigenetic, epitranscriptomic, and expression changes for identical twins. We have generated genome-wide maps of DNA and RNA methylation sites for the twins, as well a complete catalog of coding and noncoding RNA expression, which is mid-way through a 13 time-point, integrated portrait of gene regulation before, during, and after space travel. We used open-source algorithms to profile the methylation state of these samples, which revealed thousands of inter-twin differences and intra-person temporal changes in DNA and RNA. This thorough, genome-wide molecular portrait of the epigenetic, epitranscriptomic, and gene regulatory landscape of these twins will help address the potential epigenetic and long-term genome regulatory risk of space travel, and also discern if there is similarity of the twins' epigenetic changes to other phenotypes like aging and cancer.

99

Vitamin D and risk of Multiple Sclerosis: a Mendelian Randomization Study. *L.E. Mokry[1], S. Ross[1], O.S. Ahmad[1,2], V. Forgetta[1], G. Davey Smith[3], A. Leong[4], C.M.T. Greenwood[5], J.B. Richards[1,2,6,7].* 1) Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montreal, Quebec, Canada; 2) Department of Medicine, McGill University Montreal; 3) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, UK, School of Social and Community Medicine, University of Bristol,; 4) Division of General Internal Medicine, Massachusetts General Hospital and Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States; 5) Department of Oncology, Epidemiology, Biostatistics and Occupational Health, and Human Genetics, McGill University, Montreal, Quebec, Canada; 6) Department of Human Genetics, McGill University, Montréal, Québec, Canada; 7) Department of Twin Research and Genetic Epidemiology, King's College London, United Kingdom.

Background: Observational studies have demonstrated an association between decreased vitamin D levels and risk of multiple sclerosis (MS), however it remains unclear whether this relationship is causal. We undertook a Mendelian randomization (MR) study to evaluate whether genetically lowered vitamin D levels influence risk of MS. **Methods and Findings:** We identified single nucleotide polymorphisms (SNPs) associated with 25-hydroxyvitamin D (25OHD) levels from the SUNLIGHT consortium, the largest (n = 33,996) genome-wide association study (GWAS) to date for vitamin D. Four SNPs were genome-wide significant for 25OHD levels (P-values ranging from $6x10^{-10}$ to $2x10^{-109}$) and all SNPs lay in, or near, genes strongly implicated in separate mechanisms influencing 25OHD. We then ascertained their effect on 25OHD levels in the Canadian Multicentre Osteoporosis Study, a population-based cohort (n = 2,347), and tested the extent to which the 25OHD-decreasing alleles explained variation in 25OHD levels. We found that the count of 25OHD-decreasing alleles across these four SNPs was strongly associated with lower 25OHD levels (F-test statistic = 49.7, P = $2.4x10^{-12}$). Next, we conducted an MR study to describe the effect of genetically lowered 25OHD on the odds of MS in the International MS Genetics Consortium, the largest genetic association study to date for MS (n = 14,498 cases, n = 24,091 controls). Alleles were weighted by their relative effect on 25OHD levels and sensitivity analyses were performed testing MR assumptions. MR analyses found that each genetically determined standard deviation decrease in log transformed 25OHD levels conferred a 2.0-fold increase in odds of MS (95% CI: 1.7-2.5; P= $7.7x10^{-12}$; I^2 = 63%, 95% CI: 0%-88%) . These results persisted after sensitivity analyses excluding SNPs possibly influenced by population stratification or pleiotropy (OR = 1.7, 95% CI: 1.3-2.2; P = $2.3x10^{-5}$; I^2 = 47% 95% CI: 0%-85%) and including only SNPs involved in 25OHD synthesis or metabolism ($OR_{synthesis}$= 2.1, 95% CI: 1.6-2.6; P=$1x10^{-9}$ and $OR_{metabolism}$= 1.9, 95% CI: 1.3-2.7; P=0.002). While these sensitivity analyses decrease the possibility that pleiotropy may have biased the results, residual pleiotropy is difficult to exclude entirely. **Conclusions:** Genetically lowered 25OHD levels are strongly associated with MS susceptibility. Whether vitamin D sufficiency can delay, or prevent, MS onset merits further investigation in long-term randomized controlled trials.

100

Genome-wide interaction study of Parkinson disease and vitamin D deficiency implicates immune system pathways. *W.K. Scott[1,2], L. Maldonado[2], G.W. Beecham[1,2], E.R. Martin[1,2], M.L. Evatt[3], J.C. Ritchie[4], J.L. Haines[5], C.P. Zabetian[6,7], H. Payami[8,9], M.A. Pericak-Vance[1,2], J.M. Vance[1,2], L. Wang[1,2].* 1) Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami MIller School of Medicine, Miami, FL; 2) John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami FL; 3) Department of Neurology, Emory University, Atlanta, GA; 4) Department of Pathology, Emory University, Atlanta, GA; 5) Department of Epidemiology and Biostatistics, and Institute for Computational Biology, Case Western Reserve University, Cleveland, OH; 6) Veterans Affairs Puget Sound Health Care System, Seattle, WA; 7) Department of Neurology, University of Washington, Seattle, WA; 8) Departments of Neurology and Genetics, University of Alabama-Birmingham, Birmingham, AL; 9) HudsonAlpha Institute for Biotechnology, Huntsville, AL.

Parkinson disease (PD) is influenced by both genetic and environmental risk factors. Meta-analysis of 15 genome-wide association studies (GWAS) identified 28 common loci associated with PD. These loci explain only part of the overall risk of PD; the remainder is likely composed of rare variants, epistasis, and gene-environment (GxE) interaction. Vitamin D deficiency (VDD; plasma concentration <20 ng/mL) is reproducibly associated with PD. To identify novel loci, we conducted a genome-wide interaction study (GWIS) in samples from two GWAS: one with 477 PD cases and 430 controls and one with 482 PD cases and 412 controls, both imputed to 5.3 million single nucleotide polymorphisms (SNP) using the 1000Genomes reference panel. Vitamin D concentration was measured by tandem mass spectrometry. VDD was associated with PD in both datasets (OR=2.7, p<0.0001; OR=1.9, p=0.009). GWIS analysis compared two logistic regression models from each dataset: a full model with SNP allele dosage, VDD, SNP*VDD interaction, and covariates (sex, age, sampling season), and a restricted model of VDD and covariates. Then a 2-df joint meta-analysis (JMA) of SNP main effect and GxE interaction was conducted using JMA implemented in METAL. The only genome-wide significant result ($p<5x10^{-8}$) was due to SNP main effects in *SNCA*, a known PD gene. Three other loci had results with $p<10^{-6}$. Two were due to main effects (in *MAPT*, a known PD gene, and a novel intergenic locus on 20q13.33 between *MIR548AG2* and *LOC100506470*. The third resulted primarily from GxE and was located at 5q11.2 in *LOC401188*. To evaluate potential clustering of nominally significant results in pathways, we selected 465 genes with at least one joint test p<0.05 after adjusting for the number of SNPs in the gene. These genes were tested for enrichment of KEGG pathways using WebGestalt. In total, 11 pathways were over represented (p<0.05 after FDR correction for multiple comparisons); 8 were immune system-related, the most significant of which was antigen processing and presentation (p=0.0026). These results show that GWIS might reveal novel loci (such as *LOC401188*) associated with PD, and that consideration of GxE interactions across pathways might reveal biological mechanisms for disease. The clustering of nominally significant joint tests in immune system pathways is consistent with the role vitamin D plays in immune response and our prior GWAS results associating PD with SNPs in the HLA class II region.

101

***DNAJC6* mutations associated with early-onset Parkinson's disease.** *S. Olgiati[1], M. Quadri[1], M. Fang[2], J.P.M.A. Rood[3], J.A. Saute[4], H.F. Chien[5], C.G. Bouwkamp[1,6], J. Graafland[1], M. Minneboo[1], G.J. Breedveld[1], J. Zhang[2], F.W. Verheijen[1], W. Mandemakers[1], A.J.W. Boon[3], J.A. Kievit[1], L.B. Jardim[4,7], E.R. Barbosa[5], C.R.M. Rieder[8], K.L. Leenders[9], J. Wang[2,10,11,12], V. Bonifati[1].* 1) Department of Clinical Genetics, Erasmus MC, Rotterdam, the Netherlands; 2) BGI-Shenzhen, Shenzhen, China; 3) Department of Neurology, Erasmus MC, Rotterdam, the Netherlands; 4) Medical Genetics Service, Hospital de CIÌnicas de Porto Alegre, Porto Alegre, Brazil; 5) Department of Neurology, University of São Paulo, São Paulo, Brazil; 6) Department of Psychiatry, Erasmus MC, Rotterdam, the Netherlands; 7) Department of Internal Medicine, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil; 8) Neurology Service, Hospital de CIÌnicas de Porto Alegre, Porto Alegre, Brazil; 9) Department of Neurology, University Medical Center Groningen, the Netherlands; 10) Department of Biology, University of Copenhagen, Copenhagen, Denmark; 11) Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia; 12) Macau University of Science and Technology, Avenida Wai long, Taipa, Macau, China.

Background *DNAJC6* mutations were recently described in two families with autosomal recessive juvenile parkinsonism with onset before age 11, prominent atypical signs, poor or absent response to levodopa, and rapid progression to wheelchair-bound state within ~10 years from onset. However, *DNAJC6* mutations have not been associated with Parkinson's disease (PD) so far.**Methods** The entire *DNAJC6* open reading frame was analyzed in 274 patients with early-onset sporadic PD (n=182), or familial PD compatible with autosomal recessive inheritance (n=92). Selected variants were followed up by cosegregation, homozygosity mapping, linkage analysis, and whole-exome sequencing.**Findings** We identified two families and a sporadic patient with different biallelic *DNAJC6* mutations segregating with PD. In each of the two families, novel *DNAJC6* homozygous substitutions (c.2779A>G and c.2223A>T) were flanked by long runs of homozygosity within significant linkage peaks (LOD score 3.07 and 3.18). High-coverage whole-exome sequencing did not detect additional pathogenic variants within the linkage regions. The sporadic patient carried two rare variants (c.2038+3A>G, c.1468+83del) with a possible effect on RNA splicing. All these cases fulfilled the criteria for a clinical diagnosis of early-onset PD, had symptoms onset in the third-to-fifth decade, and slow disease progression. Response to dopaminergic therapies was prominent but, in some patients, limited by psychiatric side-effects. The clinical phenotype overlaps that of other monogenic forms of early-onset PD, caused by mutations in *parkin*, *PINK1*, or *DJ-1*. Furthermore, another 11 PD probands carried rare *DNAJC6* variants in single heterozygous state, including a frameshift mutation cosegregating with PD in two siblings.**Interpretation** Our findings delineate a novel form of hereditary early-onset PD. Screening of *DNAJC6* is warranted in all patients with early-onset PD compatible with autosomal recessive inheritance. Our data provide further evidence for the involvement of synaptic vesicles endocytosis and trafficking in PD pathogenesis.**Funding** Stichting ParkinsonFonds; Beijing Genomics Institute.

102

A novel protein aggregation mechanism triggered by translation of cryptic amyloidogenic elements in the 3' UTR of neurofilament genes. *A. Rebelo[1], A. Abrams[1], E. Cottienne[2], A. Horga[3], M. Gonzalez[1], A. Sanchez-Mejias1[1], L. Matilde[2], H. Houlden[2], J. Blake[2], C. Woodward[2], M. Sweeney[2], J. Holton[2], M. Hanna[2], J. Dallman[3], M. Auer-Grumbach[4], M. Reilly[2], S. Zuchner[1].* 1) Dr. John T Macdonald Department of Human Genetics and John P. Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, FL; 2) MRC Centre for Neuromuscular Diseases, UCL Institute of Neurology, Queen Square, London,UK; 3) Department of Biology, University of Miami, Miami, FL; 4) Department of Orthopaedics, Medical University Vienna, Vienna, Austria.

Abnormal protein aggregation is commonly observed in an expanding number of neurodegenerative diseases including amyotrophic lateral sclerosis (ALS), Alzheimer's and Parkinson's diseases. We describe a novel mechanism for intracellular toxic aggregates induced by an unusual mutational event in axonal neuropathy families. We identified families with Charcot-Marie-Tooth disease type 2 (CMT2) carrying frameshift mutations in the neurofilament heavy gene (NEFH) leading to stop-loss and extended translation of 40 amino acids in the 3'UTR. *In silico* aggregation prediction analysis combined with experimental results demonstrated that the terminal 20 residues in the alternative protein is amyloidogenic and critical for the formation of aggregates. Expression of the mutant NEFH in neuro-2a cells resulted in prominent abnormal protein aggregates, leading to disruption of the neurofilament network, altered cell morphology and abnormal mitochondrial distribution. Zebrafish embryos injected with mRNA encoding the mutant *NEFH* resulted in significantly decreased lengths of motor neuron axons. We also identified a similar aggregation induced mechanism triggered by translation of a cryptic amyloidogenic element present in the 3'UTR of the neurofilament light (NEFL), also known to cause aggregation in CMT and other diseases. Our studies expand on the hypothesis of neurofilament aggregation and dysfunction as a disease mechanism for axonopathies. In addition, we present a novel protein aggregation-triggering mechanism, which should be taken into consideration when evaluating stop-loss variations.

103

Identification of a founder mutation in *ABCA7* in a Belgian cohort of Alzheimer's disease patients. *K. Sleegers[1,2], E. Cuyvers[1,2], T. Van den Bossche[1,2,3,4], A. De Roeck[1,2], C. Van Cauwenberghe[1,2], S. Vermeulen[1,2], M. Mattheijssens[1,2], K. Peeters[1,2], S. Engelborghs[2,4], M. Vandenbulcke[5], R. Vandenberghe[6], P.P. De Deyn[2,4], C. Van Broeckhoven[1,2], BELNEU Consortium.* 1) VIB and University of Antwerp, Antwerp, Belgium; 2) Institute Born-Bunge, Antwerp, Belgium; 3) Department of Neurology, Antwerp University Hospital, Edegem, Belgium; 4) Department of Neurology and Memory Clinic, Hospital Network Antwerp, Middelheim and Hoge Beuken, Antwerp, Belgium; 5) Department of Old Age Psychiatry and Memory Clinic, University of Leuven and University Hospitals Leuven Gasthuisberg, Leuven, Belgium; 6) Laboratory for Cognitive Neurology, Department of Neurology, University of Leuven and University Hospitals Leuven Gasthuisberg, Leuven, Belgium.

ABCA7 was identified as a risk gene for Alzheimer's disease (AD) in genome-wide association studies (GWAS) and is one of the genes most strongly associated with AD risk in the Belgian population. Recently, significant association of *ABCA7* loss-of-function (LOF) variants with AD was reported in an Icelandic population. In the context of a targeted massive parallel re-sequencing of GWAS-identified AD risk genes in our Belgian AD cohort (772 unrelated Belgian patients; 757 Belgian healthy elderly), we could substantiate an increased frequency of predicted LOF mutations in AD patients compared to control individuals (RR 4.03, 95%CI 1.75-9.29), with 4 frameshift and 2 nonsense mutations present only in patients. One frameshift mutation (p.E709fs) was observed in 9/772 patients and not in control individuals. Moreover, this mutation was found to segregate with disease in a Belgian family with autosomal dominant inheritance of AD. This mutation causes a reading frameshift leading to a premature stop codon and predicting an N-truncated protein which is degraded. As expected a decrease of ABCA7 expression was observed in brain and/or lymphoblast cell lines of the mutation carriers.To explore a founder effect, additional unrelated AD patients (n=183, mean AAO 78·8±6·0 years) and control individuals (n=265) from the source population as well as 356 Belgian AD patients originally referred for molecular diagnostic screening of monogenic dementia (mean AAO 61.8±10.2 years), were screened for p.E709fs using Sanger sequencing. We identified 6 additional carriers among patients, of whom 4 from the diagnostic cohort, and no carriers among controls. The mutation is located on a common haplotype indicating that all carriers are descendants of a common genetic founder. The diagnostic cohort further included 1 carrier (0.3%) of a probable pathogenic *PSEN1* mutation, and 1 carrier (0.3%) of a pathogenic *GRN* mutation, which was lower than the frequency of the *ABCA7* founder mutation in this cohort (1.1%). All 15 patients carrying the *ABCA7* founder mutation showed a classical AD phenotype, with a relatively late onset age (73.7 ± 8.1 years) but a wide range of 36 (54-90) years, and an increased proportion of patients with positive family history.In conclusion, we describe a founder effect of a rare LOF mutation in the GWAS-identified AD risk gene *ABCA7* which may underlie unexplained familial clustering of AD in patients with an otherwise classical late-onset AD phenotype.

104

***SORL1* rare variants: a major risk factor for familial early onset Alzheimer disease.** *G. Nicolas[1, 2, 3], C. Charbonnier[2], D. Wallon[2, 3, 4], O. Quenez[2], C. Bellenguez[5], B. Grenier-Boley[5], S. Rousseau[2], A.-C. Richard[2], A. Rovelet-Lecrux[3], K. Le Guennec[3], D. Bacq[6], J.-G. Garnier[6], R. Olaso[6], A. Boland[6], V. Meyer[6], J.-F. Deleuze[6], P. Amouyel[5], H.-H. Munter[7], G. Bourque[7], M. Lathrop[7], T. Frebourg[1, 3], R. Redon[8], L. Letenneur[9], J.-F. Dartigues[9], E. Génin[10], J.-C. Lambert[5], D. Hannequin[1, 2, 3, 4], D. Campion[11].* 1) Department of Genetics, Rouen University Hospital, Rouen, Normandy, France; 2) National Reference Center for Young Onset Alzheimer Patients, Rouen, Normandy, France; 3) Inserm U1079, IRIB, Normandy Univ, Rouen, France; 4) Department of Neurology, Rouen University Hospital, Rouen, Normandy, France; 5) Pasteur Institute and Inserm U1167, Lille, France; 6) Centre National de Génotypage, Evry, France; 7) McGill University and Génome Québec Innovation Centre, Montréal, Canada; 8) Inserm, UMR 1087, l'institut du thorax, Nantes, France, CNRS, UMR 6291, Nantes, France, Université de Nantes, Nantes, France, CHU Nantes, l'institut du thorax, Service de Cardiologie, Nantes, France,; 9) Inserm, U897, Bordeaux France; Univ Bordeaux, Bordeaux, France; 10) Inserm UMR1078, CHU Brest, Univ Bretagne Occidentale, Brest, France; 11) Department of Research, Rouvray Psychiatric Hospital, Sotteville-lès-Rouen, France.

Causative variants within the *APP*, *PSEN1* or *PSEN2* genes are detected in 77-85% of families with Autosomal Dominant Early Onset Alzheimer Disease (ADEOAD). All these mutations result in increased aggregation of the Aβ peptide in the brain, leading to AD. In 5/14 ADEOAD unrelated patients with no pathogenic variant within these genes, we previously reported potentially pathogenic variants in the *SORL1* gene. *SORL1* encodes the Sortilin-related receptor with A-type repeats (also known as SORLA or LR11) which plays a key role in the trafficking of the Amyloid β (Aβ) Precursor Protein (APP), as an APP neuronal sorting receptor, and redirects nascent Aβ peptides to the lysosome. It therefore plays a protective role against Aβ neuronal secretion. Based on these results obtained in a small set of families, we aimed to determine if rare predicted damaging *SORL1* variants may contribute to EOAD genetic risk. For that purpose, we conducted a whole exome analysis in 484 patients with EOAD and 498 ethnically-matched controls from France. We used the following strategy: (1) selection of extreme forms of the disease (age of onset before 65 years, 52% had an onset before 55 years), (2) stringent diagnostic criteria, (3) enrichment in cases with positive family history (42%), (4) rigorous matching of controls on ethnicity, (5) high depth of coverage (mean ~120x), (6) filtering of rare (MAF<1%) variants to retain only disruptive (nonsense, frameshift, splice site) or missense variants predicted damaging by three algorithms, and (7) collapsing of variants at the gene level. Exome wide, *SORL1* was the top hit: we detected an enrichment of disruptive and predicted damaging missense rare *SORL1* variants in cases (Odds Radio (OR) = 5.03, 95% Confidence Interval [CI] =[2.02-14.99], p= $7.49.10^{-5}$). This enrichment was even stronger when restricting the analysis to the 205 cases with positive family history (OR=8.86, 95% CI=[3.35-27.31], p=$3.82.10^{-7}$). These results remained highly significant after adjustment on *APOE* status or restriction to the cases with the highest level of diagnostic evidence. We conclude that predicted damaging rare *SORL1* variants are a strong risk factor for EOAD and that the association signal is mainly driven by familial cases. This is the first study reporting a gene-based association of rare variants with an exome wide significance in AD, using case-control whole exome sequencing data.

105

Loss-of-function mutations in *TBK1* are a frequent cause of frontotemporal dementia and amyotrophic lateral sclerosis in Belgian and European cohorts. *C. Van Broeckhoven[1, 2], I. Gijselinck[1, 2], S. Van Mossevelde[1, 2], J. van der Zee[1, 2], A. Sieben[1, 2, 3], B. Heeman[1, 2], S. Engelborghs[2, 4], M. Vandenbulcke[5], R. Vandenberghe[6], P. De Jonghe[1, 2], P. Cras[2, 7], P. De Deyn[2, 4], J-J. Martin[2], M. Cruts[1,2], BELNEU Consortium, EU EOD Consortium.* 1) Department of Molecular Genetics, VIB, University of Antwerp, Antwerp, Belgium; 2) Institute Born-Bunge, University of Antwerp, Antwerp, Belgium; 3) Department of Neurology, University Hospital Ghent and University of Ghent, Ghent, Belgium; 4) Department of Neurology and Memory Clinic, Hospital Network Antwerp Middelheim and Hoge Beuken, Antwerp, Belgium; 5) Department of Psychiatry, University Hospitals Leuven and University of Leuven, Leuven, Belgium; 6) Department of Neurology, University Hospitals Leuven and University of Leuven, Leuven, Belgium; 7) Department of Neurology, Antwerp University Hospital, Edegem, Belgium.

In the Belgian frontotemporal dementia (FTD) patient cohort, mutations in known FTD genes accounted for 30% of the familial FTD patients and 75% of familial FTD-ALS patients, with several autosomal dominant families remaining genetically unresolved. In an informative, 4-generations FTD-ALS family, whole-genome- sequencing identified the tank-binding kinase 1 gene (*TBK1*) as the strongest candidate gene. Further, other groups recently published *TBK1* as a gene for amyotrophic lateral sclerosis (ALS), FTD-ALS and FTD. Hence, we aimed to assess the genetic contribution of *TBK1* in a Belgian cohort of 481 FTD and FTD-ALS patients, and 147 ALS patients. We used multiplex parallel sequencing of *TBK1* in the family and the patient cohorts, and in a control group.We identified a loss-of-function (LOF) mutation, p.Glu643del, in *TBK1* that segregated in the extended FTD-ALS family. In the cohorts, we observed 10 patients carrying a mutation resulting in an overall mutation frequency of 1.6% (10/628): 0.9% in FTD patients (4/459), 3.4% in ALS patients (5/147) and 4.5% in FTD-ALS patients (1/22).Mean onset age of *TBK1* carriers was 62.1 ± 8.9 years with ALS carriers being much younger. Disease penetrance is incomplete, with 70% of the carriers affected by age 70 years and two carriers remaining unaffected aged over 80 years. Five of the unrelated patients were carrying the p.Glu643del mutation with the mutation located on 3 different haplotypes sized between 3 and 17.5 Mb. The p.Glu643del mutation carriers had a signficant later onset age and longer disease duration suggesting a milder pathological effect. The LOF mutations (frameshift, nonsense, amno acid deletions), including the p.Glu643del mutation, led to loss of transcript and/or protein in blood and/or brain. Preliminary data of the *TBK1* mutation screening in a large European cohort of FTD and ALS patients (approx. n=2500), ascertained with the European Early-Onset dementia (EU EOD) consortium, confirmed our findings and will provide information regarding mutation spectrum and frequencies in different countries.These findings are in line with published studies and extend and reiterate that FTD and ALS belong to one disease continuum. Decreased expression of TBK1 in brain suggests (partial) haploinsufficiency as an underlying disease mechanism. Several FTLD and ALS proteins are in the same pathway as TBK1, including optineurin and p62, stressing a role for autophagy and inflammation in neurodegeneration.

106

The *C9orf72* repeat expansion modulates onset age of FTD-ALS through increased DNA methylation and transcriptional downregulation. *M. Cruts[1,2], I. Gijselinck[1,2], S. Van Mossevelde[1,2], J. van der Zee[1,2], A. Sieben[1,2,3], S. Engelborghs[2,4], J. De Bleecker[3], A. Ivanoiu[5], O. Deryck[6], D. Edbauer[7,8], M. Zhang[9], B. Heeman[1,2], E. Rogaeva[9,10], P. De Jonghe[1,2,11], P. Cras[2,11], J.-J. Martin[2], P.P. De Deyn[2,4], C. Van Broeckhoven[1,2], BELNEU Consortium.* 1) Department of Molecular Genetics, VIB, University of Antwerp , Antwerp, Antwerp, Belgium; 2) Institute Born-Bunge, University of Antwerp, Antwerp, Belgium; 3) Department of Neurology, University Hospital Ghent and University of Ghent, Ghent, Belgium; 4) Department of Neurology and Memory Clinic, Hospital Network Antwerp Middelheim and Hoge Beuken, Antwerp, Belgium; 5) Department of Neurology, Saint-Luc University Hospital and Institute of Neuroscience, Université Catholique de Louvain, Brussels, Belgium; 6) Department of Neurology, General Hospital Sint-Jan Brugge-Oostende, Bruges, Belgium; 7) German Center for Neurodegenerative Diseases (DZNE), Munich, Germany; 8) Adolf Butenandt Institute, Biochemistry and Munich Cluster of Systems Neurology (SyNergy), Ludwig-Maximilians University Munich, Munich, Germany; 9) Tanz Centre for Research in Neurodegenerative Diseases, University of Toronto, Toronto, Canada; 10) Department of Medicine, Division of Neurology, University of Toronto, Toronto, Canada; 11) Department of Neurology, Antwerp University Hospital, Edegem, Belgium.

Pathological expansion of a G_4C_2 repeat in the 5' regulatory region of *C9orf72* (MIM *614260) is the most common genetic cause of frontotemporal dementia and/or amyotrophic lateral sclerosis (FTD-ALS MIM #105550). *C9orf72* patients have highly variable onset ages suggesting the presence of modifying factors and/or anticipation. We assessed the effect of G_4C_2 repeat expansion size on onset age, the role of genetic anticipation and the effect of repeat size on DNA methylation and activity of the *C9orf72* promoter in 72 index patients and 61 affected or unaffected relatives with a *C9orf72* repeat expansion derived from a Belgian cohort of 593 patients with FTD including 40 FTD-ALS patients, and 227 ALS patients. Repeat expansion sizes measured in blood DNA varied between 45 and over 2100 G_4C_2 repeat units with short expansions of 45 to 78 units present in 5.6% of the 72 index patients carrying an expansion. Short expansions as little as 47 units co-segregated with disease in two families. A subject with a short expansion in blood but an indication of mosaicism in brain showed the same TDP-43 and dipeptide-repeat (DPR) pathology as those with a long expansion. Further, we provided evidence for an association of G_4C_2 expansion size with onset age ($p<0.05$) most likely explained by an association of methylation state of the 5' flanking CpG island and repeat expansion size in blood ($p<0.0001$) and brain ($p<0.05$). In several informative *C9orf72* parent-child transmissions, we identified earlier onset ages, increased expansion sizes and/or increased methylation state of the 5' CpG island across two generations reminiscent of disease anticipation. Also, intermediate repeats of 7 to 24 units showed a slightly higher degree of methylation ($p<0.0001$) and a decrease of *C9orf72* promoter activity ($p<0.0001$) compared to normal short repeats of 2 to 6 units, possibly explaining the associated risk with disease. Decrease of transcriptional activity was even more prominent in the presence of disease-related small deletions flanking G_4C_2 ($p<0.0001$). Together, we provided evidence that increased methylation of CpG sequences in the *C9orf72* promoter due to an increased G_4C_2 repeat size may lead to loss of function of *C9orf72* in FTD-ALS.

107

Missed connections: failure to recombine is a common feature of human oogenesis. *T. Hassold, H. Maylor-Hagen, J. Gruhn, P. Hunt.* Molecular Biosciences, Washington State University, Pullman, WA.

Failure to recombine is arguably the most important known cause of human nondisjunction, having been linked to maternally- or paternally-derived cases of sex chromosome trisomy and autosomal trisomies 13, 14, 15, 18, 21 and 22. However, almost all information on these "exchangeless" homologs has come from studies of trisomic conceptuses. Thus the incidence of this defect and its impact on gametogenesis is not clear; e.g., if oocytes or spermatocytes containing exchangeless homologs are selected against during meiosis, the incidence may be even higher in gametes than in zygotes. To address this, we initiated comparative studies of exchangeless chromosomes in fetal ovarian samples from elective terminations of pregnancy and already collected data from testicular biopsies involving males with obstructive azoospermia. To date we have examined over 8,000 oocytes from 191 females and over 4,000 spermatocytes from 56 males. We have identified striking male:female differences, with approximately 10% of occytes – but only 1-2% of spermatocytes – containing at least one exchangeless chromosome pair. Detailed analyses of oocytes indicates striking chromosome-specific differences, with almost all exchangeless homologs involving chromosomes 21 or 22. Further, the effect is linked to the overall level of recombination in the cell, with the presence of one or two exchangeless chromosomes per cell associated with 10% and 20% reductions, respectively, in the total number of crossovers in the cell; thus, extremely low levels of recombination increase the likelihood of one or more exchangeless chromosomes. Finally, our observations indicate significant inter-individual variation in the incidence of exchangeless chromosomes: although, on average, 10% of oocytes exhibited this recombination defect, in some samples over 25% of oocytes contained at least one exchangeless chromosome pair. Taken together, our observations indicate a remarkably high level of aberrant recombination in human oocytes and provide important initial evidence of a genetic predisposition to meiotic nondisjunction.

108

An isogenic trisomic-disomic model system using cells from people with mosaic Down syndrome unmasks trisomy 21 associated increases in age-related chromosomal instability and telomere shortening. *K. Rafferty[1], C. Charalsawadi[1,2], C. Jackson-Cook[1,3].* 1) Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA; 2) Department of Pathology, Faculty of Medicine, Prince of Songkla University, Songkhla, Thailand; 3) Department of Pathology, Virginia Commonwealth University, Richmond, VA.

It is known that age-related changes impacting multiple organ systems occur earlier in people with Down syndrome (Ds), but the biological basis underlying this trisomy 21-associated propensity for premature aging is poorly understood. Given that the trisomic/normal cells from people with mosaic Ds (mDs) are identical with regards to environmental exposures and genes (except for chromosome 21 copy number), comparisons of these isogenic trisomic/disomic cells allow one to "unmask" the cellular consequences of trisomy 21 by removing extraneous factors. The primary aim of this study was to determine if trisomy 21 results in an increase in the acquisition of age-related somatic chromosomal changes. To meet this aim chromosome-specific telomere lengths and/or instability frequencies were compared between the isogenic trisomic/normal cells of 24 people with mDs ranging from 1 to 44 years of age. Somatic chromosomal instability (CIN) frequencies were quantified by scoring 1000 cells using the cytokinesis-block micronucleus (MN) assay coupled with FISH (RUNX1 probe to distinguish trisomic/euploid nuclei). Chromosome-specific telomere lengths were quantified using a Q-FISH (pantelomeric probe) method. In the younger participants (ages 1-10; n=15), no significant difference was observed in CIN frequencies in the euploid (0.18 mean ± 0.02 SEM) compared to trisomic (0.19 ± 0.03) cells ($p>0.05$), but in the older participants (ages 16-44; n=9), the trisomic binucleates showed a significant increase in the relative proportion of cells with CIN (0.33 ± 0.06) when compared to their euploid (0.15 ± 0.03) cells ($p=0.01$). MN containing chromosome 21 occurred more often than expected by chance ($p<0.01$), but accounted for only 8.8% of the total MN detected, with the other MN containing chromatin from other chromosomes. Trisomic cells also had significantly shorter telomeres across all chromosomes compared to euploid cells (to date, n=13; $p<0.001$), with this difference being apparent as early as age 2. Overall, the shortest telomeres tended to be observed for 9q in both euploid and trisomic cells. Collectively, these results suggest that the cellular effects related to aging in Ds/mDs arise from a "network" involving multiple acquired chromosomal findings and are not limited to alterations involving chromosome 21. They also support the use of this isogenic mDs model system for providing new insight about cellular changes that arise from a trisomy 21 imbalance.

109

Runs of homozygosity (ROH) reveal correction of chromosomal imbalances during fetal development. *A.L. Penton[1], D. Segal[2], R. Burnside[1], P. Papenhausen[1], Runs of homozygosity (ROH) reveal correction of chromosomal imbalances during fetal development.* 1) Laboratory Corporation of America Holdings, Durham, NC; 2) Rutgers New Jersey Medical School.

Trisomy and monosomy rescue are well known processes that correct early embryonic aneuploidies resulting from non-disjunction. This occurs by loss of the extra chromosome in the case of trisomy and duplication of a single chromosome in the case of monosomy. Although the genomic imbalance is "corrected" this process often results in uniparental disomy (UPD) during trisomy rescue and always results in UPD during monosomy rescue. In contrast, segmental UPD (sUPD) is localized to a specific region of the chromosome and the etiology is not always clear. We present patients that display evidence of UPD by SNP array (visualized as ROH) due to correction of a genomic imbalance to disomy during fetal development. Fetal material from one patient showed potential UPD 12 as well as a low level of trisomy 12 (17%) consistent with incomplete trisomy rescue. A newborn with features of Down Syndrome displayed complete allele homozygosity of the most proximal portion of 21q contiguous with a 10.16 Mb duplication of the remaining distal portion of the long arm. This likely resulted from a (3:1) unbalanced meiotic segregation of a parental balanced translocation, and non-viable monosomy for the centromeric portion of chromosome 2. Subsequent duplication of the normal chromosome 21 resulted in viable trisomy for the Down Syndrome critical containing distal region of chromosome 21. A nine year old referred for developmental delay carried a 9.4 MB terminal ROH on chromosome 1p that was rescued from an unbalanced der(1)(t(1;17)(p36.3;q21) prenatally. This is consistent with a mitotic recombination based correction mechanism. A 42 year old woman referred for infertility carried a 58.54 Mb ROH on terminal 15q, yet did not have Prader-Willi or Angelman syndrome. This suggested segmental UPD15 which was confirmed by parental analysis. We suspect a deletion rescue mechanism similar to the derivative rescue of the previous patient. These cases show that UPD can result not only from correction of aneuploidy due to non-disjunction but also from correction of partial aneuploidy caused by the segregation of unbalanced translocation derivatives, or deletions. Clinical features in these individuals may be caused by residual effects from the early gestational presence of imbalance. If correction occurs early enough so that little or no clinical features are present, transmission of imbalance is possible from germ cells.

110

Genomic imbalances in fetuses and patients with congenital heart malformation. *I. Maya[1], S. Kahana[1], T. Tenne[1], S. Jakobson[1], J. Yeshaya[1], M. Shohat[1,2], L. Basel-Vanagaite[1,2,3,4].* 1) Recanaty Genetic Institute, Rabin Medical Center, Petah Tikva, Israel; 2) The Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel; 3) Pediatric Genetics, Schneider Children s Medical Center of Israel, Petah Tikva, Israel; 4) Felsenstein Medical Research Center, Rabin Medical Center, Petah Tikva, Israel.

Background:Congenital heart malformations(CHM) have been reported to be associated with pathogenic copy number variants (CNVs). The purpose of this study was to compare the frequency of clinically significant (likely pathogenic or pathogenic) submicroscopic CNVs in:1) fetuses with CHM 2)postnatal cohort of patients with CHM.We also report on prenatal and postnatal diagnostic yield of clinically significant CNVs in the specific types of CHM. Methods:During 2010-15 our clinical diagnostic hospital-based laboratory performed 10,000 chromosomal microarray tests on fetal and postnatal samples using CGH and SNP array platforms.In total,300 of cases were referred due to CMH including 200 prenatal and 100 postnatal cases.In the prenatal cohort 67% had isolated CHM.For postnatal cases, 10% of the individuals presented with an isolated CHM. Results:Overall,clinically significant CNVs were identified in 13% (39/300) of the specimens referred with normal or unknown karyotypes.Clinically significant CNVs were detected in 24% of the patients and 10% of the fetuses. Known clinically significant CNVs were detected in 62% and unique CNVs in 38% of the cases.Velo-Cardio-Facial syndrome comprised 25% of the known pathogenic CNVs.Out of the clinical significant CNVs,18% were also later detected by karyotyping. The most common CHM(n=121) was ventricular septal defect (VSD).80% of the prenatal cases and 10% of the postnatal cases presented as an isolated VSD.The detection rate of clinically significant CNVs was lowest in fetuses with isolated VSD(2%) and highest in postnatal cases with additional abnormalities(25%).In CHM group with aortic arch abnormalities (coarctation, dilatation,double aortic arch,overriding aorta,right aortic arch),the detection rate was 10% in both prenatal and postnatal group.In cases with Tetrallogy of Fallot,clinically significant CNVs were detected in 23%;in cases with transposition of great arteries(TGA),clinically significant CNVs were detected in 12%.Variants of unclear clinical significance(VUS) were identified in prenatal and postnatal CHM cases in 3 and 9%, respectively.In the control group of low risk pregnancies,-clinically significant CNVs were identified in 1.8%.Conclusion:The overall diagnostic yield of clinically significant CNVs in prenatal cases with CHM was significantly lower than in postnatal cases with CHM.This can be explained by to the presence of additional abnormalities,such as intellectual disability,in 90% of the postnatal cohort.

111

The role of copy number losses in non-syndromic cleft lip and palate. *L.A. Harney[1,2,3], B.W. Darbro[1,3], A. Long[2], J. Standley[1], R.A Cornell[3,4], J.C. Murray[1,3], J.R. Manak[1,2,3].* 1) Department of Pediatrics, University of Iowa, Iowa City, IA; 2) Department of Biology, University of Iowa, Iowa City, IA; 3) Interdisciplinary Program in Genetics, University of Iowa, Iowa City, IA; 4) Department of Anatomy and Cellular Biology, University of Iowa, Iowa City, IA.

Clefts of the lip and/or palate (CL/P) occur in about 1 in 700 live births and are categorized as non-syndromic (NSCL/P) or syndromic (SCL/P). Individuals with NSCL/P have isolated clefts and account for about 70% of clefting cases, whereas syndromic occurrences include cognitive or structural anomalies. Although genome-wide association, candidate gene, and animal model studies have been used to study CL/P, a large-scale analysis of copy number variation (CNV) in CL/P has yet to be performed. We performed a high resolution array-based comparative genomic hybridization study to identify copy number variants associated with NSCL/P in a cohort of 851 cases from the Philippines. Focusing on rare copy number losses, our preliminary analysis identified 74 genes that were deleted in greater than one individual while 213 genes were deleted in a single case; collectively, the majority of genes were not previously implicated in clefting. After comparing the list of deleted genes to OMIM, DECIPHER, NCBI, and MGI databases, four were selected for functional follow-up in zebrafish. These genes, *GALNT13* [MIM 608369], *PKP2* [MIM 602861], *MYO5C* [MIM 610022] and *ULK4* [MIM 615075], are all novel clefting candidates, are overlapped by a CNV loss in greater than one individual, and appear in less than 1% of the cohort. Six additional genes identified have been previously implicated in clefting through association studies (*NTN1* [MIM 601614], *PCYT1A* [MIM 123695]), variant analyses (*ZNF750* [MIM 610226], *CDH1* [MIM 192090], *OFD1* [MIM 300170]), or chromosomal microarrays (*IMMP2L* [MIM 605977]). Replication studies with a Caucasian cohort of over 300 individuals with NSCL/P (in addition to 302 individuals with SCL/P) are currently underway. Together, these studies will define the contribution of copy number variants to disease incidence of CL/P.

112

Interchromosomal core duplicons drive recurrent complex inversions within the chromosome 8p23.1 region. *K. Mohajeri[1], C.D. Campbell[1], J. Huddleston[1,2], B. Nelson[1], C.R. Catacchio[3], M. Ventura[3], E.E. Eichler[1,2].* 1) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; 2) Howard Hughes Medical Institute, Seattle, WA, USA; 3) Department of Biology, University of Bari, Bari, Italy.

The chromosome 8p23.1 region harbors one of the largest and most common inversion polymorphisms in the human population. The inversion haplotype is thought to predispose to recurrent rearrangements associated with congenital heart defects and the region is the source of extensive structural variation of beta-defensin gene families, which are risk factors for autoimmune disease. The complexity of the region and its enrichment in recent segmental duplications has complicated disease association, assembly and studies of natural genetic variation. We generated a high-quality sequence assembly of an inverted 6.2 Mbp haplotype of the 8p23.1 locus by single-molecule real-time sequencing of 70 large-insert clones from a CH17 human hydatidiform mole BAC resource. The alternate haplotype configuration shows nine genic differences and an increased propensity for non-allelic homologous recombination. We find two highly identical, directly orientated duplications ~385 kbp in length, which are largely missing from the human reference genome, mapping on either side of a congenital heart defect-associated critical region. Instead of a single inversion polymorphism, we identify three inversion events of 311 kbp, 442 kbp and 4.6 Mbp in length. Our phylogenetic analysis with outgroup non-human primate genomes suggests that all three events are specific to the human lineage arising between 0.6-0.9 million years ago. Remarkably, each inversion breakpoint within 8p23.1 is flanked by an inversion-associated repeat (IAR) that ranges in size from 54-62 kbp. We have identified a total of 16 genomic IARs and estimate the duplication expansions that distributed these interchromosomal cores throughout seven chromosomes to have occurred between 10-20 million years ago across primate evolution. In addition to flanking the 8p23.1 inversions, IARs are observed at the bounds of two additional inversion events that have led to the structural differences between primates on chromosomes 3 and 11. Through constructing an alternate reference assembly of the 8p23.1 locus, we were able to further understand structural variation within the locus, supply a framework for the role of this alternate haplotype in congenital heart defect-related deletion, and characterize interchromosomal cores we identify as sequence elements mediating inversions.

113

Characterization of Mosaic Chromothripsis in the Human Germline by Chromosomal Microarray and Whole Genome Sequencing. *F. Quintero-Rivera[1], C. Redin[2], N. Dorrani[3], J.A. Martinez-Agosto[4], M.E. Talkowski[2].* 1) Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, UCLA Clinical Genomic Center,Los Angeles,CA; 2) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 3) Dept. of Pediatrics and Human Genetics, UCLA David Geffen School of Medicine, Los Angeles, CA; 4) Dept. of Pediatrics, UCLA David Geffen School of Medicine, Los Angeles, CA.

Complex human germline chromosomal rearrangements, with more than three breakpoints, can originate from deconstruction of chromosomal segments via a single devastating event into multiple smaller fragments which are subsequently rejoined. This phenomenon, defined as ‘chromothripsis’, was first detected in cancer cells. We later defined examples of chromothripsis in the human germline of subjects with congenital anomalies, with the presence of oscillating reciprocal dosage changes that were characteristic in cancer cells. The reconstructed derivative chromosomes exhibited a complex genomic architecture comprised of translocations and/or inversions. However, to our knowledge, mosaic germline chromothripsis exhibiting complex balanced and unbalanced rearrangements has not been described.Here, we report a 29-year-old female with mosaic partial trisomy of eight regions of chromosome 19. The proband has facial dysmorphism, skeletal defects, hypotonia, delayed speech and motor skills, mild intellectual disability, progressive insidious decline in cognitive function, difficulty with immediate short-term memory, and mild loss of brain volume. G-banded karyotype and M-FISH analysis of peripheral blood revealed a *de novo* mosaic intrachromosomal duplication of the long arm of chromosome 19q13.11 due to insertion of material from the same chromosome 19 (q13.33q13.43) in 35% of cells examined [mos 46,XX,ins dup(19)(q13.11q13.33q13.43)[7]/46,XX[13]dn.ish ins dup(19)(wcp19+)]. Chromosomal microarray analysis (CMA) revealed eight *de novo* mosaic gains of part of the short arm of chromosome 19p13.11p12 and of part of the long arm of chromosome 19q12q13.42 with disomic regions in between duplicated intervals, in about 60% of cells. These regions encompass a total of 18.68 Mb; 6.7 Mb on p-arm and 11.98 Mb on q-arm. Whole-genome sequencing (WGS) of large-inserts, e.g. jumping libraries, was performed, generating 79.8-fold physical coverage of the genome. These analyses revealed still greater complexity, including multiple inserted inverted segments at cytobands 19p12 and 19p13.3. These data represent one of the most complex examples of germline chromothripsis described to date, and suggest that WGS may represent a powerful tool to accurately characterize mosaicism in routine cytogenetic practice.

114

Mechanistic Insights of Complex Insertional Translocations. *S. Gu[1], C.M.B. Carvalho[1], B. Yuan[1], W. Bi[1], A. Patel[1], S.W. Cheung[1], J.R. Lupski[1,2,3,4].* 1) Molecular & Human Genetics, Baylor College of Medicine, Houston, TX; 2) Texas Children's Hospital, Houston, TX; 3) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 4) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.

Chromothripsis-like chromoanasynthesis causes germline complex genomic rearrangements (CGRs) observed in patients with congenital disorders. In contrast to the oscillation between two copy number states (copy number neural and deletion) in chromothripsis, frequent copy number changes with interspersed regions containing deletions/duplications/triplications are found in chromoanasynthesis; such events are usually restricted to one chromosome or one chromosome arm. A very recent *in vitro* study of DNA damage in micronuclei in single human cells showed chromothripsis-like events involving translocations between two chromosomes (Zhang et al., *Nature* 2015, PMID: 26017310). Here, we observed chromoanasynthesis events involving two or three chromosomes in patients with complex insertional translocations (ITs). IT is defined as a segment of one chromosome being inserted into a new region of the same or a non-homologous chromosome, while complex IT may generate gain or loss of chromosome segments at the inserted loci. We identified 16 individuals with complex IT involving known disease genes or disease candidate genes among the 56,000 individuals tested from January 2007 to November 2014 at Baylor College of Medicine Medical Genetics Laboratories using a combination of clinical microarray and fluorescence *in situ* hybridization (FISH). Subsequently, customized high-density aCGH was performed on 10 individuals with available DNA, and breakpoint junctions were fine-mapped at nucleotide resolution by long-range PCR and DNA sequencing in 7 individuals. We observed that complex ITs involving two or three chromosomes could be part of a chromoanasynthesis event. In addition, microhomologies and small-scale complexities, in the form of insertion of fragments apparently templated from nearby sequence at the breakpoint junctions, were observed. These observations resemble the breakpoint junction signatures found in CGRs generated through chromoanasynthesis. We showed that chromoanasynthesis generated through replicative-based mechanism(s) could involve different chromosomes and may generate interchromosomal complex IT and thus not be restricted to one chromosome.

115

Pathogenesis, risk stratification and management of pregnancy-associated aortic dissection in Marfan syndrome and related connective tissue disorders. *M.L. Russo[1], G.L. MacCarrick[2], E. Sparks[2], H.C. Dietz[2], J.P. Habashi[3].* 1) Gynecology and Obstetrics & McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD; 2) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD; 3) Pediatric Cardiology, Johns Hopkins University, Baltimore MD.

Marfan syndrome (MFS) is an autosomal dominant connective tissue disorder caused by mutations in *FBN1* with a strong predisposition for aortic aneurysm and dissection. Studies in MFS mouse models and humans suggest that excessive TGFb signaling in the aortic wall contributes to disease pathogenesis through noncanonical activation of ERK. Pregnant MFS women have an elevated risk of aortic dissection in the peripartum period. We have strong evidence in MFS mice to suggest that production of oxytocin throughout pregnancy, particularly in the third trimester, and the sustained release with lactation, drives this increased risk through ERK signaling. Indeed, avoidance of lactation or use of an oxytocin antagonist abrogates risk of peripartum aortic dissection in MFS mouse models. Through an IRB approved patient survey and retrospective chart review, pregnancy, delivery and nursing data were collected in MFS (n=176) and Loeys-Dietz syndrome (LDS; n=53), a related connective tissue disorder. Pregnancies were included with survival past 20 weeks gestation. Risk of dissection in pregnancy was not significantly different between MFS (5.6%) and LDS (5.7%). Sub-analysis of patients in whom pre-pregnancy aortic root diameter was known revealed a significant increase in the incidence of aortic dissection if diameter was $\geq$4cm (4/16, 25%) versus <4cm (3/64, 4.7%; $p<0.05$). There were 10 dissections in MFS, 6 during pregnancy and 4 postpartum (PP, defined as birth-12 months post-delivery) and 3 PP dissections in LDS. 5/13 dissections were type B (MFS 3/10 and 2/3 LDS). Of the 7 PP dissections, there was no significant difference in mode of delivery (3/133 vaginal, 4/106 c-section) or induction with pitocin (2/53 induction vs. 5/186 no induction). PP aortic dissection (average 4.5 months PP) was uniquely seen among lactating women with MFS (4/145 versus 0/31), but greater numbers of study participants will be needed to reach firm conclusions. There was a significant linear trend towards increasing dissection rate with increasing pregnancy number ($p<0.05$). Aortic dissection in MFS and LDS remains a significant cause of morbidity and mortality and the risk persists through 12 months PP, particularly in lactating women and c-section does not appear to mitigate this risk. Type B dissections occur as often as type A in pregnant MFS women, which may limit the utility of prophylactic aortic root replacement in anticipation of pregnancy and modify imaging surveillance in pregnancy.

116

Williams-Beuren syndrome as an epigenetic disease: association of GTF2IRD1 with chromatin silencing complexes. *P. Carmona-Mora[1], F. Tomasetig[1], C.P. Canales[1], A. Alshawaf[2], M. Dottori[2], J.I. Young[3], R. Barres[4,5], E.C. Hardeman[1], S.J. Palmer[1].* 1) Cellular and Genetic Medicine Unit, School of Medical Sciences, UNSW Australia, Sydney, NSW, Australia; 2) Centre for Neural Engineering, The University of Melbourne, Melbourne, VIC, Australia; 3) John P. Hussman Institute for Human Genomics, University of Miami Leonard Miller School of Medicine, Miami, FL, USA; 4) The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark; 5) School of Medical Sciences, UNSW Australia, Sydney, NSW, Australia.

GTF2IRD1 is a member of the *GTF2I* gene family, located on 7q11.23 a region prone to genomic rearrangements. Hemizygous deletions in 7q11.23 cause Williams-Beuren syndrome (WBS) and duplications cause WBS duplication syndrome. Human mapping data and analyses of mouse knockouts implicate *GTF2IRD1* and *GTF2I* as the prime candidates for the craniofacial abnormalities, mental retardation, visuospatial construction deficits and hypersociability of WBS. Exploring the molecular role of GTF2IRD1 and GTF2I (TFII-I) therefore provides a means to understand the cellular cause of WBS, but also provides a unique opportunity to study the genetic and epigenetic mechanisms that contribute to many aspects of human behaviour. To understand the basis of the phenotypes related to GTF2IRD1, we have studied its molecular function *in vivo* and *in vitro*. In previous transgenic mouse studies, we showed that GTF2IRD1 repressed a specific functional group of genes in muscle tissue, but recent follow-up results from a DNA methylation analysis by MBD-Seq showed that DNA methylation is not the mechanism by which this repression occurs. Yeast two-hybrid screening has allowed us to identify a range of novel protein interaction partners for GTF2IRD1, which mainly include proteins involved in chromatin modification and transcriptional regulation. We have demonstrated the presence of GTF2IRD1 in chromatin modifier complexes and identified direct associations with histone deacetylases (HDACs) in human embryonic stem cell-derived neurons at the endogenous level; thus developing the GTF2IRD1 interactional network and creating testable hypotheses for GTF2IRD1 molecular function. These hypotheses predict a functional association with HDAC activity. Alteration of GTF2IRD1 levels in cell lines by forced expression or siRNA knockdown was found to affect the assembly of core HDAC complex proteins and the level of global HDAC enzymatic activity, suggesting that a major element of GTF2IRD1-mediated gene regulation operates via histone deacetylation. These findings are also consistent with much of the work emerging for TFII-I. On this basis, one may hypothesise that WBS and WBS duplication syndrome are largely epigenetic diseases, in which cellular abnormalities are caused by disturbances in the ratio of elements of the chromatin modifying machinery, leading to abnormalities of gene regulation that are cell context dependent.

117

Pseudoxanthoma elasticum: Expanding *ABCC6* mutation database and pathogenicity test of missense mutations by zebrafish mRNA rescue. *S. Raftari[1], H. Vahidnezhad[1,2,3], L. Youssefian[1,2], A. Mirzaei[1], M. Daneshpazhouh[2], M. Salehi[2], H.R. Mahmoudi[2], Q. Li[1], J. Uitto[1].* 1) Thomas Jefferson University, Philadelphia, PA; 2) Tehran University of Medical Sciences, Tehran, Iran; 3) Pasteur Institute of Iran, Tehran, Iran.

Pseudoxanthoma elasticum (PXE), the prototype of heritable ectopic mineralization disorders, is characterized by calcium hydroxyapatite deposition in the skin, eyes and the cardiovascular system with considerable morbidity and mortality. The disease is inherited in an autosomal recessive pattern, and the majority of cases is caused by mutations in the *ABCC6* gene which encodes a putative transmembrane efflux transporter expressed primarily in the liver and the kidneys. Attesting to the genetic heterogeneity are recent demonstrations of mutations also in the *ENPP1* and *GGCX* genes in patients with PXE-like phenotypes. In this study, we examined a cohort of 7 families of Iranian ancestry diagnosed by characteristic clinical features and by skin histopathology as PXE. Mutation detection strategy consisted of PCR amplification of all 31 exons of *ABCC6*, together with 50-70 bp flanking intronic sequences. Sequencing of the PCR products revealed sequence variants in all patients, two of them being homozygous for nonsense mutations (p.R1141X and p.Y1069X). The remaining sequence variants resulted in amino acid substitutions, two of them (p.G1405V and p.R760W) being pathogenic based on bioinformatics predictions by PolyPhen-2 (probably damaging, score 1.0) and SIFT programs (damaging, score 0). Furthermore, the wild-type amino acids, G1405 and R760, were conserved through evolution and they were not present in SNP databases. The pathogenicity of *ABCC6* missense mutations has been tested by an mRNA rescue assay in zebrafish. Specifically, morpholino-mediated knock-down of *Abcc6*, preventing splicing of pre-mRNA with subsequent formation of premature termination codon of translation, resulted in profound phenotype consisting of pericardiac edema, curved tail, and stunted growth, and demise of the embryos by day 7 of post-fertilization. Wild-type human or mouse *ABCC6* mRNA, when injected concomitantly with the morpholino, resulted in essentially complete rescue of the phenotype, while mutant mRNAs harboring the pathogenesis missense mutation did not elicit such rescue. This mRNA rescue assay can be utilized for further analysis of the putative missense mutations for cataloging additional *ABCC6* mutations in PXE towards expanding database, with implications for genetic counseling, prenatal testing and preimplantation genetic diagnosis, as well as for development of mutation-based personalized treatment.

118

Comparison of phenotypic features associated with *TGFBR1* and *TGFBR2* mutations: results of the Montalcino Aortic Consortium. *G. Jondeau[1,2], J. Ropers[3], E. Regalado[4], A. Braverman[5], A. Evangelista[6], G. Teixedo[6], J. De Backer[7], L. Muino Mosquera[7], S. Naudion[8], C. Zordan[8], T. Morisaki[9], H. Morisaki[9], Y. Von Kodolitsch[10], S. Dupuis-Girod[11], S.A. Morris[12], R. Jeremy[13], S. Odent[14], M. Langeois[1], M. Spentchian[1], M. Aubart[1,2], C. Boileau[1,15], R. Pyeritz[16], D. Milewicz[4], Montalcino Aortic Consortium.* 1) CNMR Syndrome de Marfan et apparentés, APHP Hopital Bichat, paris, paris, France; 2) INSERM U1148, LVTS, Hopital Bichat, 75017 France; 3)) Unité de Recherche Clinique HU Paris Œle-de-France Ouest, Hôpital Ambroise Paré, 92100 Boulogne, France; 4) Division of Medical Genetics, University of Texas at Houston Health Science Centre, Houston, Texas; 5) Cardiovascular Division, Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110; 6) Servei de Cardiologia, Hospital Universitari Vall d'Hebron, Barcelona, Spain; 7) Department of Cardiology and Center for Medical Genetics, University Hospital Ghent, Belgium; 8) Service de Génétique Médicale, Hôpital Pellegrin, 33076 Bordeaux Cedex, France; 9) Department of Bioscience and Genetics, National Cerebral and Cardiovascular Center Research Institute, Suita, Osaka, Japan; 10) German Aorta Centre Hamburg at the Centre of Cardiology and Cardiovascular Surgery, University Medical Centre Hamburg-Eppendorf, Hamburg; Germany; 11) Service de Génétique, Hôpital Femme-Mère-Enfant - Groupe Hospitalier Est, 69677 BRON, France; 12) Department of Pediatrics-Cardiology, Texas Children's Hospital / Baylor College of Medicine, Houston, TX 77030; 13) Marfan and Aortic Disease Clinic, Royal Prince Alfred Hospital, University of Sydney, Sydney, NSW, 2006, Australia; 14) Service de Génétique Clinique, CHU de Rennes; Institut de Génétique et Développement, CNRS UMR6290, Université de Rennes 1, Rennes, France; 15) Service de Genetique, AP-HP, Hopital Bichat, 75018 Paris, France; 16) Perelman School of Medicine at the University of Pennsylvania. Smilow Center for Translational Research 11-133. Philadelphia, PA 19104.

Background: Mutations in genes encoding the TGF-beta Type I and II receptors **(***TGFBR1* and *TGFBR2*) have been reported in patients with either syndromic or non-syndromic Heritable Thoracic Aortic Disease. ***Purpose***: We sought to determine the range of phenotypic expression and phenotypic differences between patients with *TGFBR1* vs. *TGFBR2* mutations.***Methods***: Data were collected using a standardized questionnaire and compared between the two populations.***Summary of results***: Clinical data from 397 patients from 188 families, assessed in 13 institutions were collected.

	TGFBR1 n=157 (40%)	TGFBR2 n=240 (60%)	p
Probands	61 (39)	96 (40)	0.9
Females	90 (57)	124 (52)	0.3
Presenting feature			0.7
Aortic root aneurysm	25 (16)	51 (21)	
Ao dissection	24 (15)	37 (15)	
Dysmorphy	13 (8)	16 (7)	
Familial screening	94 (60)	133 (55)	
Age at last FU (median [IQR])	33 [19, 51]	31 [19, 44]	0.3
Head and neck arterial tortuosity	47/93 (50)	67/126 (53)	0.8
Hypertelorism	31/123 (25)	53/179 (30)	0.4
Translucent skin	56/129 (43)	72/207 (35)	0.1
Wide scars	33 (26)	60/200 (30)	0.4
Broad or bifid uvula	28/126 (22)	61/199 (31)	0.1
Arched palate	81 (65.9)	97/206 (47)	0.02
Craniosynostosis	9/105 (9)	17/168 (10)	0.8
Marfan systemic score : mean (sd)	3.95 (3.44)	4.13 (2.99)	0.6
% systemic score 7 or greater	26/120 (21.7)	33/176 (18.8)	0.6
Cardiac defect (BAV, VSD, PDA…)	12/147 (8.2)	39/224 (17.4)	0.01
First aortic event			0.03
Thoracic Aneurysm repair	32 (52)	54 (53)	
Type A Ao dissec	28 (45)	33 (32)	
Type B Ao dissec	2 (3)	15 (15)	
Extra aortic arterial event	16 (10)	20 (9)	0.7

An aortic event (surgery and/or aortic dissection) occurred more frequently in males than females with *TGFBR1* mutations (54% of males (median age 27 years [21-39]) and 30% of females (34 years [28-43]), p=0.003), but gender differences for aortic events were not significant in patients with TGFBR2 mutations (55% of males (median age 31 [20-41]) and 40% of females (31 years [20-41], NS}.The last maximal aortic root diameter recorded before or at the time of type A aortic dissection was greater in *TGFBR1* mutation carriers than in *TGFBR2* (65mm (±22) vs. 50 (±12); p= 0.04). The prevalence of aortic dissection was similar in the 2 but type B aortic dissections were more frequent in patients with *TGFBR2* mutations.***Conclusions:*** Clinical features of patients carrying a *TGFBR1* and *TGFBR2* are similar but aortic disease presentation differs. Aortic dissection appears to be more frequent in men, is observed at a larger aortic diameter, and less frequently affects the descending aorta in patients with *TGFBR1* mutation.

119

Variations in non-coding regions of *TGFβ3*, *CDH2* and *IRF6*, affecting their expression, show association with cleft lip and palate (CL±P). *P. Kumari[1], S.K. Singh[2], R. Raman[1].* 1) Cytogenetics Laboratory, Department of Zoology, Banaras Hindu University, Varanasi, Uttar Pradesh-221005, India; 2) G.S. Memorial Plastic Surgery Hospital and trauma Center, Mehmoorganj, Varanasi, Uttar Pradesh-221010, India.

A number of genes and variations within their coding regions have shown association with CL±P. We report variations in non-coding regions of 3 candidate genes (*TGFβ3*-MIM 190230*, CDH2*-MIM 114020*, IRF6*-MIM 607199) and their association with non syndromic cleft lip with or without cleft palate (NSCL±P) or Van der Woude syndrome (VWS; MIM 119300), a syndromic condition of CL±P. Sequencing of *TGFβ3* showed 3 variants in promoter region and one in 3`UTR. In-silico analysis revealed each promoter variant creating a different transcription factor binding site. Possible effect of these variants on the promoter activity of *TGFβ3* was studied by luciferase assay. One variant exhibited 3.7 times greater activity than the normal one while in the construct having the other two variants expression was marginally higher than the control. In a family having the monogenic, autosomal disorder, VWS, whole genome sequencing detected a variation within the intronic region of the adhesion gene, *CDH2*, cosegregating with the disease. A case-control association study of the intronic variant with NSCL±P, using ARMS-PCR showed its association in the recessive model. RT-PCR, performed in the lip/palate tissue samples of NSCL±P cases to analyse the expression potential of this region, showed its expression which was confirmed by SLOT-BLOT RNA hybridization. In addition, 5` and 3` RACE along with Southern hybridization confirmed transcription within the intronic region. In-silico modelling of RNA folding, through RNA Fold Web Server showed a more stable stem and loop structure in the mutant rather than the normal intronic sequence. The mutant sequence also indicated possibility of the formation of miRNA from this region. In another VWS family, screening of *IRF6* gene exhibited a haplotype involving 3 non-coding region variations (2 intronic and one 3`UTR) that co-segregated with the disease. Association of *IRF6* with VWS has been established in Caucasans but not in Asian, especially Indian, cases. The qRT-PCR analysis revealed that the expression of the mutant was 2.27 times lower than the normal haplotype which was also confirmed by the luciferase assay. This report highlights the role of mutations in non-coding region in gene regulation and manifestation in disease conditions such as cleft lip±palate in this case.

120

Natural history of dermatan 4-*O*-sulfotransferase 1 (D4ST1)-deficient Ehlers-Danlos Syndrome (DDEDS): from an international collaborative clinical study by the International Consortium for EDS. *T. Kosho[1], D. Syx[2], T. Van Damme[2], H. Morisaki[3], H. Kawame[4], T. Sonoda[5], Y. Hilhorst-Hofstee[6], A. Maugeri[7], N. Voermans[8], R. Mendoza-Londono[9], K. Wierenga[10], P. Jayakar[11], K. Ishikawa[12], T. Kobayashi[13], Y. Aoki[14], S. Watanabe[15], T. Ohura[13], M. Kono[16], K. Mochida[17], T. Morisaki[3], N. Miyake[18], M. Malfait[2].* 1) Department of Medical Genetics, Shinshu University School of Medicine, Matsumoto, Japan; 2) Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium; 3) Department of Bioscience and Genetics, National Cerebral and Cardiovascular Center Research Institute; 4) Division of Genomic Medicine Support and Genetic Counseling, Tohoku Medical Megabank Organization; 5) Department of Pediatrics, University of Miyazaki, Miyazaki, Japan; 6) Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands; 7) Department of Clinical Genetics, Center for Connective Tissue Disorders, VU University Medical Center, Amsterdam, The Netherlands; 8) Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; 9) Clinical and Metabolic Genetics, the Hospital for Sick Children, Toronto, Canada; 10) Section of Genetics, Department of Pediatrics, the University of Oklahoma College of Medicine, Oklahoma City, OK; 11) Division of Genetics and Metabolism, Miami Children's Hospital, Miami, FL; 12) Department of Pediatrics, Iwate medical University School of Medicine, Morioka, Japan; 13) Department of Pediatrics, Tohoku University School of Medicine, Sendai, Japan; 14) Department of Medical Genetics, Tohoku University School of Medicine, Sendai, Japan; 15) Department of Orthopedics, Nakajima Hospital, Sendai, Japan; 16) Department of Dermatology, Nagoya University Graduate School of Medicine, Nagoya, Japan; 17) Department of Dermatology, University of Miyazaki, Miyazaki, Japan; 18) Department of Human Genetics, Yokohama City University Graduate School of Medicine.

Dermatan 4-*O*-sulfotransferase 1 (D4ST1)-deficient Ehlers-Danlos syndrome (DDEDS), caused by recessive loss-of-function mutations in *CHST14* encoding D4ST1, is a recently delineated form of EDS, characterized by various malformations and progressive multisystem fragility-related manifestations. At present, 31 patients (21 families) have been published. To establish the natural history, an international collaborative study was planned. We have collected detailed and comprehensive clinical information from 20 published patients (17 families) and 15 additional patients (14 families). Craniofacial features in infancy included large fontanelle, hypertelorism, short and downslanting palpebral fissures, blue sclera, low-set and rotated ears, high arched palate, long philtrum, small mouth, and micro-retrognathia. Facial shapes became slender and asymmetrical with protruding jaw from adolescence. All had congenital multiple contractures, with adducted thumbs and clubfoot in most. Finger shapes were characteristic (tapering, slender, cylindrical). Feet were progressively deformed to pes planus with valgus. Joint hypermobility was significant especially at small joints. (Kypho)scoliosis was frequently seen and surgically corrected in six. Skin hyperextensibility was observed from early infancy, progressing to redundancy into adulthood. Wrinkling palmar creases were more evident according to aging. Large subcutaneous hematomas, observed in most, were frequently the most serious problem, occurring after minor traumas, spreading in several hours with severe pain, and sometimes accompanying hemorrhagic shock. Two fatal complications were described: a massive skin necrosis resulting from a large subcutaneous hematoma all around the leg, caused by manual reposition of traumatic hip dislocation; perforation of descending colon diverticula followed by skin rupture, with deterioration of general conditions to death. Other important complications included congenital heart defects (atrial septal defects) and retinal detachment in some and cryptorchidism (male) and constipation in most. The current study illustrates the natural history that at birth patients manifest an arthrogryposis-like appearance and later fit the hallmark of EDS (skin hyperextensibility, joint hypermobility, tissue fragility) with a decrease of ADL/QOL and sometimes with potential lethality due to progressive skeletal deformities, large subcutaneous hematomas, and visceral/ophthalmological complications.

121

***Cbx3* and its role in craniofacial development: zebrafish as a model system for testing dysmorphology candidate genes.** *H.E. Edelman[1], C. Woods[1], J.E. Hoover-Fong[2], A.S. McCallion[1].* 1) McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD; 2) Greenberg Center for Skeletal Dysplasias, McKusick-Nathans Institute of Genetic Medicine, Department of Pediatrics, Johns Hopkins University, Baltimore, MD.

With the wide range of craniofacial abnormalities seen in children, it is important to have an efficient pipeline to identify and test the candidate genes involved in these developmental anomalies. Because we can easily manipulate the zebrafish genome using CRISPR/Cas9 technology and the cartilage structures of its head are well characterized, zebrafish make an excellent model system for studying craniofacial abnormalities. In 2003, Hoover-Fong et al. published a report of a patient with severe facial dysgenesis and a deletion of 7p15.1-p21.1. We compared a SNP array from this patient to that from a patient with a partially overlapping deletion on chromosome 7p, but no craniofacial abnormalities. This process narrowed the genomic region that was uniquely deleted in our patient with the extreme facial malformation. The result was a region with 5 relatively uncharacterized candidate genes. We have begun to systematically determine the biological requirement for the genes in the interval beginning with the gene *cbx3* (chromobox homolog 3), whose protein product is known as heterochromatin protein 1 gamma (HP1-gamma) and is thought to be involved in the switch from euchromatin to heterochromatin. In zebrafish, *cbx3* has been duplicated during evolution and exists as *cbx3a* and *cbx3b*. To check for biologically relevant expression of this candidate gene, we did whole-mount in situ hybridization at various time points in development. We saw that both *cbx3a* and *cbx3b* begin to express throughout the whole zebrafish embryo at 24 hours post fertilization (hpf) but expression becomes restricted to the head at 48 and 72 hpf. Since this confirmed a probable role for *cbx3* in the development of craniofacial structures, we used CRISPR/Cas9 technology to create a line of fish with a 2 base pair deletion in the first exon of *cbx3a*. This frameshift mutation results in a premature termination codon and predicted nonsense-mediated decay resulting in a loss of function of the allele. We are currently phenotyping the offspring of this line using acid-free double staining with alcian blue and alizarin red to examine the developing cartilage and bone of the head. Although our data are preliminary, they are suggestive of a requirement for *cbx3a* in normal facial development. We are using this work as a platform from which to build a systematic pipeline for the analysis of the genetic architecture of craniofacial anomalies.

122

Cerebro-costo-mandibular syndrome revisited: phenotype and outcome of twenty molecularly confirmed individuals. *D.C. Lynch[1], M. Tooley[2], E.J. Bhoj[3], E.G. Lemire[4], B.N. Chodirker[5], J.P. Taylor[6], D.K. Grange[7], E.H. Zackai[3], E.P. Kirk[8], J. Hoover-Fong[9], L. Fleming[10], R. Savarirayan[11], S.F. Smithson[2], A.M. Innes[1,12], J.S. Parboosingh[1,12], F.P. Bernier[1,12].* 1) Medical Genetics, University of Calgary, Calgary, Alberta, Canada; 2) Department of Clinical Genetics, St. Michael's Hospital, Bristol, UK; 3) Division of Genetics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA; 4) Division of Medical Genetics, Department of Pediatrics, University of Saskatchewan, Saskatoon, Saskatchewan, Canada; 5) Department of Pediatrics and Child Health, University of Manitoba, Winnipeg, Manitoba, Canada; 6) Genetic Health Service, Auckland, New Zealand; 7) Division of Genetics and Genomic Medicine, Washington University School of Medicine, St Louis, Missouri, USA; 8) School of Women's and Children's Health, University of New South Wales, Randwick, New South Wales, Australia; 9) Greenberg Center for Skeletal Dysplasias, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, USA; 10) National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA; 11) Victorian Clinical Genetics Services, Murdoch Childrens Research Institute, and University of Melbourne, Parkville, Victoria, Australia; 12) Alberta Children's Hospital Research Institute for Child and Maternal Health, Calgary, Alberta, Canada.

Cerebro-costo-mandibular syndrome (CCMS) is an autosomal dominant disorder with reduced penetrance that we recently identified to be caused by hypomorphic mutations in *SNRPB.* Smith *et al*. first described CCMS in 1966 as a triad of Pierre Robin sequence, posterior rib gaps, and intellectual disability. The prognosis was initially reported to be poor, with approximately 50% of patients dying in the first year of life from respiratory insufficiency. It has been suggested that the initial reports of poor CNS outcome were a result of neonatal hypoxic brain injury. However, the long term clinical and in particular neurological prognosis remains poorly documented. Here we discuss the clinical spectrum of CCMS in a cohort of 16 families (20 individuals) with mutations in *SNRPB*. All but two patients were born after 1990. All penetrant patients had the classic posterior rib gap anomalies. Of the 20 patients, four (20%) died before the age of one year. Improved survival is likely due to aggressive use of early tracheostomy, which was required in seven of our patients. Respiratory distress was noted in 13 patients (65%), and five required a period of mechanical ventilation. All patients who underwent mandibular distraction were subsequently able to breathe without ventilation. Scoliosis is present in 60% of patients in this cohort, with four patients requiring surgery. While three patients in the cohort have learning difficulties, all are cognitively normal, supporting that intellectual disability is in fact not a common or genetically determined feature of CCMS. We also note that severity and prognosis do not correlate with *SNRPB* genotype and observe variable disease expression across family members with the same mutation. While there is a broad range of severity associated with CCMS, in our cohort the prognosis for most affected children has improved in recent years, with most patients now surviving with little or no risk of neurodevelopmental or intellectual disabilities. Although the name of disorder will likely remain cerebro-costo-mandibular syndrome, deemphasizing the "cerebro" component initially reported to be common in this condition has important counselling implications for affected individuals and their families.

123

Development of the GLASS: Genetics Literacy Assessment for Secondary Schools. *R.J. Okamura[1], S.S. Lee[2], B. Bowling[3], K.E. Ormond[1,2].* 1) Genetics, Stanford University, Stanford, CA; 2) Stanford Center for Biomedical Ethics, Stanford, CA; 3) Biological Sciences, Northern Kentucky University, Highland Heights, KY.

With the advent of Next Generation Science Standards, the way genetics is taught in secondary schools is changing. There is currently no validated psychometric measure of secondary school genetics knowledge. Without a validated measure, there is no way to measure learning gains in genetics as a result of changes in education. To address this deficit, we developed the GLASS (Genetics Literary Assessment for Secondary Schools), an instrument to measure genetics knowledge at the secondary school level. We began with a list of core concepts created by the American Society for Human Genetics' (ASHG) Information and Education Committee. We used these concepts to create a psychometric instrument consisting of nine demographic and 25 knowledge-based questions assessing genetics literacy with an emphasis towards concepts important in healthcare. We assessed content validity via an expert review of members of ASHG's Information and Education Committee and the National Society of Genetic Counselors' education special interest group (n=17). Finally, to test the instrument, we conducted a pilot study at three high schools in the San Francisco Bay Area. We surveyed 105 subjects and conducted 15 think-aloud interviews. The average GLASS score was 12.70 points (out of 25 possible) with a standard deviation of 3.91. Data from the pilot study show evidence of construct validity. Subjects with higher grade point averages scored higher than subjects with lower grade point averages ($p<0.001$) and participants who completed honors biology scored higher than those who only completed biology ($p<0.001$). In addition, participants indicating an interest in health, medicine, and science performed better than those that did not ($p=0.023$). GLASS scores did not differ among genders, ethnic backgrounds, or racial identities. Based on our findings, we plan to conduct a large sample of participants with varying levels of genetics knowledge and geographic distribution in 2016 to complete validation of the GLASS.

124

New tool for measuring genetic variation knowledge among health professionals. *K. Abdallah[1], M. Moss[1], J. Jenkins[2], K. Calzone[3], S. Sellers[4], V. Bonham[1].* 1) Social and Behavioral Research Branch, National Human Genome Research Institute, NIH, Bethesda, MD; 2) Division of Policy, Communications, and Education, National Human Genome Research Institute, NIH, Bethesda, MD; 3) Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD; 4) Department of Family Studies and Social Work Miami University, Oxford, OH.

Background: As genomic medicine advances, it has become increasingly important for healthcare providers to integrate an understanding of genetic variation into clinical care. Yet, there is a dearth of survey measures that assesses healthcare providers' knowledge of genetic variation. The Bonham and Sellers Genetic Variation Knowledge Index [GKAI] was developed to address this gap. **Methods**: The GKAI was developed through focus groups, cognitive interviews, expert advisory panels (geneticists and survey methodologists), and exploratory factor analysis of pilot data. Items included questions such as "the DNA sequences of two randomly selected healthy individuals of the same sex are 90-95% identical." The items were scored true/false, with a "don't know" category. "Don't know" was considered an incorrect response for analytic purposes. Scores were obtained by summing the correct responses. Higher scores indicate greater scientific knowledge of human genetic variation.The index was validated with a national random sample of general internists in the USA (N=787) and modified for use with registered nurses (RNs) and nurse practitioners (NPs). The GKAI survey items were modified to conform with scope of practice standards for RNs and NPs. The final measure for physicians is a 6-item index with a 0-6 score range, while for RNs (N=694) and NPs (N=63), the index is 8-items with a 0-8 score range.**Results**: Physicians scored significantly higher on genetic knowledge than RNs (*p*-value<.0001) and NPs (*p*-value=0.01). The mean GKAI for physicians is 3.26/6.00 (54.3% correct, SD= 1.17), for RNs 3.63/8.00 (45.4% correct, SD= 1.99), and for NPs 3.76/8.00 (47% correct, SD=1.8). Previous genetic training was a significant predictor of GKA I scores for RNs only (*p*-value=0.02 RNs, 0.72 NPs, 0.78 physicians); while self-rating of genetic knowledge/perception of genetic knowledge positively associated GKAI scores for all three groups (*p*-value=0.06 RNs, 0.04 NPs, 0.04 physicians).**Conclusion**: The GKAI provides a quick and accurate tool for assessing human genetic variation knowledge across health professions. It can be administered as a pre-post education assessment. Additionally, analysis suggests that exploring individual-level provider characteristics associated with differences in knowledge may provide a more comprehensive understanding of specific healthcare provider genetic education needs.

125

Knowledge and attitudes of medical residents and fellows working in various hospitals of United States of America, on genetic testing for disease specific biomarkers and knowledge of Precision Medicine. *S. Ghavimi[1], H. Azimi[2,3,4], P. Sealy[1].* 1) Department of Medicine, Howard University Hospital , Washington, District of Columbia - DC; 2) Carleton University, Ottawa, Ontario, Canada; 3) Psychogenome, Ottawa, Ontario, Canada; 4) All Saints University School Of Medicine, Dominica.

Objectives: The aim of the study was to assess knowledge and attitudes of medical residents and fellows working in various hospitals of United States of America, on genetic testsing for disease specific biomarkers and knowledge of Precision Medicine. Methods: We distributed self-administered questionnaire to the residents and fellows either through email or by being in contact with them with phone. Logistic regression models were used to evaluate the determinants of knowledge and attitudes towards their knowledge of Genetics and Precision Medicine. Results: 3835 residents and fellows answered the survey questions. Among the fields which answered the questions Internal Medicine (31%) and Pediatrics residents (28%) wore among the highest number of residents which answered the questions. Around 10% answered correctly the questions on regarding the genes involved in developing genetic disorders. Knowledge of Precision Medicine was highest among residents and fellows which had prior genetic testing during graduate training (11%) and inversely associated with male gender. As for cancer screening and specific biomarkers, residents and fellows were more knowledgeable if they attended courses on cancer genetic testing for biomarkers (12%) or postgraduate training courses in epidemiology and evidence-based medicine (9%). More than 90% asked for the additional training on the genetic testing and genetic testing for cancer during the specialization training. Conclusion: The knowledge of Genetic testing for disease specific biomarkers and Precision Medicine among residents and fellows who answered the questionnaire appears to be insufficient. There is a need for additional training in this field. We suggest ACGME to incorporate at least one month of Genetic testing training for disease specific biomartkers and/or further Continuing medical education on Genetic testing and Precision Medicine.

126

Evolution from Expository to Open Inquiry Learning to Improve Genetic Literacy. *T.C. Tatum Parker, D. Karge-Galik.* Biological Sciences, Saint Xavier University, Chicago, IL.

Genetic literacy fosters dialogue about the scientific, social and ethical implications of genetic technologies. This literacy should respect that there are uncertainties inherent in science and that genomic research is an engine of innovation and job creation. Genetic science, and the technologies that are rapidly arising from it, are becoming increasingly more powerful in the world outside of labs and hospitals. Genomics has resulted in increased crop yields and also numerous GMO debates, DNA technology is edging open the door to personalized medicines while fueling debates over who 'owns' an individual's genetic information. Genetic literacy among the public and media is low; with the benefits often being over shadowed in the spotlight by the more sensational detrimental side effects. A study by Bates (2005) showed that the public processed a greater variety of messages than assumed by previous researchers, including documentaries, non-science fiction films, and popular television in addition to previous researchers' focuses on science fiction and news media. There were two overarching goals of this project 1) to improve the genetic literacy of students enrolled in the course; and 2) to familiarize students with genetics research and the impacts, both explicit and implicit, it can have on society. To do this we examined the efficacy of using an immersion exercise (I.E.) where students took on the role of the investigator. Students read primary literature, discussed amongst themselves, developed their own methods for recreating the experiment they read. Students then collected their data sets, 5 – 15 samples per student, following the protocol approved by the Institutional Review Board of Saint Xavier University. Then students compared their results with those of their peers and those of the published study. This study examined 3 years without I.E. and two years with the I.E. On the post assignment reflection, 87.5% of student responses indicated that the student felt that the inquiry based-cooperative learning laboratory had a positive effect on their ability to perform the tasks associated with scientific inquiry and experimental design. This was also reflected in exam scores (ANOVA $p < .0001$). It is our belief that this lead to increased genetics education in our major's classes by increasing student ownership through their participation in the process of experimental design, implementing sample collection, and data analysis.

127

Evaluation of a web-based decision aid for people considering the *APOE* genetic test for Alzheimer's risk. *D.T. Zallen[1], M. Ekstract[2], G.I. Holtzman[3], K.Y. Kim[4], S.M. Willis[5].* 1) Science and Technology in Society, Virginia Tech, Blacksburg, VA; 2) Breakneck Turtles Productions, NY; 3) Dept. of Statistics (emeritus), Virginia Tech; 4) Catawba Hospital and Virginia Tech Carilion School of Medicine, Roanoke, VA; 5) Virginia Tech Center for Survey Research.

Some forms of genetic testing enable individuals to estimate their risk for developing diseases that can occur later in their lives. Among these is the test for genes at the *apolipoprotein E (APOE)* locus, where the e4 allele is a marker for increased risk for the common form of Alzheimer's disease (AD2 [MIM 104310]). While Alzheimer's risk information can be useful for some people, for others it can provoke serious emotional distress, have adverse effects on family members, and evoke concerns about possible misuse. In the past, decisions about genetic testing were made in consultation with genetic counselors. Now, tests are offered through physicians' offices and on the internet. There is little time or opportunity for counseling. Many now enter into testing with inadequate information or preparedness. We have developed an online decision-aid prototype designed for people considering *APOE* genetic testing. It contains four components that consumers have identified as the most important when deciding about genetic testing (D.T. Zallen, *To Test or Not to Test,* Rutgers 2008). Among these is a values-clarification component with dramatized vignettes that present the pros and cons of genetic testing. The decision aid has been evaluated using before-and-after surveys to determine its effectiveness in communicating relevant knowledge, in improving understanding of risk, and in eliciting the value components of genetic testing. Suggestions for improvements were also solicited. More than 1,200 participants completed both surveys and provided extensive feedback. Quantitative and qualitative analysis of the survey responses showed considerable satisfaction with this web-based tool as a guide for decision making. Over 90% wrote that they would recommend this online aid to others. Individuals who initially indicated that they were poorly informed reported a substantial gain in insight after using it. There was an expressed preference for using the decision aid as the basis for further discussions with genetic counselors and physicians. Slightly more than half the participants changed their perceived likelihood of testing after using the tool: 35% shifting to greater likelihood and 20% to lesser likelihood. Specific suggestions for improvement have been implemented to enhance overall utility and functionality.

128

Introduction of population based NGS expanded carrier screening in the Netherlands. *K.M. Abbott, M. Viel, M. Veldhuis, M. Plantinga, T. Dijkhuizen, J. Schuurmans, I. van Langen, R. Sinke.* Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands.

With increased international focus on personalized healthcare and preventative medicine, NGS expanded carrier screening (ECS) of severe recessively inherited diseases offers a substantial benefit to existing healthcare options. These tests offer reproductive options not previously available for couples, and may ultimately reduce the number of young children with devastating disorders.ECS testing in the Netherlands currently offers targeted-mutation testing relevant to specific at (higher) risk populations and, as such, is skewed to illnesses more common to the associated populations. By using a targeted-mutation panel, potential early onset disease-relevant variants can be missed.At the University Medical Center Groningen, in the Netherlands, we have developed a whole-gene screenings test of rare, recessive single-gene illnesses. In collaboration with our (Dutch) colleagues we compiled a list of monogenic diseases reported in the clinic that have profound consequences with early-onset and a shortened life span. A final list of 70 genes associated with 50 recessively inherited diseases was established. In contrast to other tests, this is a **whole-gene** sequencing test and it is offered and **analyzed per couple** and not per individual, meaning that couples receive a result based on their collective genetic information.The Dutch population-specific database, Genome of the Netherlands (GoNL), was used to test the appropriateness of our approach. We compared the70 gene-associated variants with the relevant Human Gene Mutation Database (HGMD) 'damaging' variants. Other deleterious variants (premature loss or gain of termination codons, shifts in the reading frame and consensus splice site changes) not currently present in the HGMD were also examined. Based on known population frequencies and our *in silico* analysis of the Dutch GoNL database we expect a positive result for 1 out of 100-150 couples, implying a risk of 1 in 400-600 of conceiving a child with this disease per pregnancy, which is higher than the average risk for Down syndrome.We are currently offering this test through the Dutch healthcare system for couples with a medical indication. It is the first test of its kind to be offered in Europe. In the fall we will begin a pilot study of the general public, offering the test through physicians to enrolled couples in the North of the Netherlands with a desire to conceive.

129

Preconception genome sequencing and patient choice: The psychosocial impact of adverse results. *T.L. Kauffman[1], M. Gilmore[4], P. Himes[4], E. Morris[4], Y. Akkari[3], L. Amendola[2], J. Davis[1], M.O. Dorschner[6], G. Jarvik[2], M. Leo[1], C. McMullen[1], D. Nickerson[6], C. Pak[3], S. Punj[3], J.A. Reiss[1], J. Schneider[1], C.S. Richards[3], D.K. Simpson[4], A.L. Rope[4], P. Robertson[6], B. Wilfond[5], K.A.B. Goddard[1], CSER consortium and NextGen study team.* 1) Center for Health Research, Kaiser Permanente Northwest, Portland, OR; 2) Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA; 3) Oregon Health and Science University, Portland, OR; 4) Kaiser Permanente Northwest, Department of Medical Genetics, Portland, OR; 5) Seattle Children's Research Institute, Seattle, WA; 6) Department of Genome Sciences, University of Washington, Seattle, WA.

IntroductionGenetic screening for reproductive decision making is typically limited to a handful of conditions. As genomic sequencing becomes more cost effective, the option to receive carrier status for hundreds of conditions may be overwhelming to patients. As part of the NHGRI Clinical Sequencing Exploratory Research (CSER) consortium, the NextGen study enrolled women and their male partners seeking preconception genetic screening. Our preparatory focus groups found that patients value choice about selecting the categories of results being shared. However, enabling this choice is challenging because of the large number of options. We developed a strategy to support patient choice, while keeping the task manageable and practical. This presentation highlights the impact of offering broad choices that result in participants learning about specific abnormal genetics results. MethodsWe placed ~800 conditions into five categories related to carrier screening and one category for secondary findings. These categories were developed to improve communication by using terms that are familiar to most people, while reducing the number of options that participants need to consider. All participants receive results in the lifespan-limiting category, but they can select whether to receive results for conditions in the remaining categories; serious, mild, unpredictable outcomes, adult onset, and medically actionable secondary findings. ResultsTo date, 186 participants have enrolled in the study. Most women (89%) chose to receive all categories of results. Of the 18 men, 17 chose all categories. Several of the findings to date have had immediate relevance, including a symptomatic carrier, an x-linked condition in a pregnant woman, and a pathogenic *BRCA2* variant in a male. This presentation will include illustrative case studies that demonstrate that participants who received results in the context of categorical choices did not experience any significant adverse effects following genetic counseling. DiscussionAlthough participants receiving genome sequencing desire choice about the results they receive, they have overwhelmingly chosen to receive all categories of results. When genome sequencing in the preconception context detects findings that are immediately relevant for the participant's health or family planning, the psychosocial impact can be managed with genetic counseling.

130

Economic impact of genome sequencing for preconception carrier screening: the time costs for genetic counseling. *P. Himes[1], F.L. Lynch[1], M.J. Gilmore[1], E.M. Morris[1], J.A. Reiss[1], T.L. Kauffman[1], C. McMullen[1], J.V. Davis[1], M.C. Leo[1], J.L. Schneider[1], K.A.B. Goddard[1], B. Wilfond[2].* 1) Kaiser Permanente Northwest Center for Health Research, Portland, OR; 2) Seattle Children's Research Institute, Treuman Katz Center for Pediatric Bioethics.

Backgound: Genome sequencing (GS) will play an increasing role in clinical practice. As part of the Clinical Sequencing Exploratory Research (CSER) consortium, we are investigating the use of GS in the clinical context of preconception carrier screening. GS can identify gene variants associated with hundreds to thousands of genetic conditions, compared with currently available clinical tests or panels that typically evaluate one or a few conditions. GS may increase and improve reproductive choices for individuals and couples planning a pregnancy; however, the impact on genetic counseling services (GCS) to deliver this expanded information is poorly understood. While other studies have looked at the cost of providing GCS, the time costs required have not been investigated for GS. Time costs are an important factor to consider for complex behavioral services, such as GCS, and there is a risk of underestimating the impact of delivering a new service, such as GS, if the time costs are ignored. Methods: We quantified the time required to prepare for, counsel, and provide follow-up for preconception women and couples regarding results for ~800 pathogenic and likely pathogenic carrier status variants and 112 medically actionable incidental findings using prospective tracking. We then compared how similar genetic counseling for GS is to other services and which aspects of the GCS are most time intensive.Results: Over 60% of study participants were a carrier of at least one condition. The time needed to prepare for the visit ranged from 15 minutes to over two hours, with more preparation time being required for very rare conditions. The counseling visit took an average of 35 minutes (range, 15-50 min.), with visits being longer for multiple or complex results, or when the study participant was pregnant when the results were disclosed. The post-visit follow-up, including documentation in the EMR and coordinating clinical care, took an average of 20 minutes (range, 10-60 min.).Conclusions: Preliminary policy implications of this study if this technology is expanded to a clinical setting will likely include a substantial increase in the time required to deliver GCS, which we estimate would scale to ~1 FTE per million covered lives. This impact could be mitigated by streamlining and standardizing information on rare (and unfamiliar) results in order to reduce preparation time.

131

Phenome-wide association study provides biologic insights into the etiology of age-related macular degeneration. *M. Brilliant[1], J. Mayer[3], J. Liu[1], W. Lee[2], B. Hoch[3], S. Schrodi[1], J. Joyce[1], A. Ikeda[2], S. Hebbring[1,2].* 1) Center for Human Genetics, Marshfield Clinic Research Fndn, Marshfield, WI; 2) Department of Medical Genetics, University of Wisconsin Madison, Madison, WI; 3) Bioinformatics Research Center, Marshfield Clinic Research Fndn, Marshfield, WI.

Age-related macular degeneration (AMD) is the leading cause of blindness in adults. Risk factors include advanced age, gender, ethnicity, smoking, and family history. Early stage AMD is characterized by medium sized extracellular deposits called drusen and loss of endothelial cells from the pillars of choriocapillaris, which serves as the vascular supply for the retinal pigmented epithelium (RPE). Numerous genome-wide association studies (GWASs) have demonstrated that AMD is a genetically complex disease with over 30 independent genetic loci. Conversely, it remains to be elucidated on how these genes increase AMD risk. To better understand the genetic contribution to AMD pathology and etiology, we conducted a comprehensive phenome-wide association study (PheWAS) using ten independent SNPs known to be associated with AMD. We hypothesized that other phenotypes may share a common genetic biology with AMD. We genotyped ten AMD-associated SNPs on 3,887 Marshfield Clinic patients linked to an extensive electronic medical record system. A PheWAS was then conducted by calculating weighted genetic risk scores on the ten SNPs across 4,653 phenotypes defined by diagnostic ICD9 codes. As expected, these ten SNPs were strongly associated with AMD (P=5.5E-22). More importantly, these ten SNPs were also significantly associated with clinical conditions defined as "disorders of the arteries and arterioles" (DAA) (P=2.1E-7). This genetic association was further validated in an independent population. Patients diagnosed with DAA and AMD were further analyzed in approximately 1.4 million Marshfield Clinic patients. In the larger clinic population, DAA was associated with AMD risk (P=1.9E-37, OR_{adj} 1.5) and patients diagnosed with DAA tended to be diagnosed with AMD three years earlier compared to those never diagnosed with DAA (P<0.0001). Immunohistochemistry staining of two AMD candidate genes in endothelial cell lines and clinical DAA tissues demonstrate strong expression. This study demonstrates an expanded application of the PheWAS approach to study complex diseases such as AMD. In conjunction with the known pathology of AMD, these results also provide strong evidence that AMD-predisposing SNPs may increase AMD risk by directly affecting blood vessels of the eye and that these genetic effects are not limited to the eye.

132

Complex diseases are associated with variation in Mendelian genes: A phenome-wide study using Human Phenotype Ontology and a population genotyped on the Exome BeadChip. *L.A. Bastarache[1], J. Mosely[2], T. Edwards[2], R. Carroll[1], H. Mo[1], L. Wiley[1], W. Wei[1], J. Denny[1,2].* 1) Biomedical informatics, Vanderbilt University, Nashville, TN; 2) Department of Medicine, Vanderbilt University Medical Center, Nashville, TN.

Many Mendelian syndromic diseases have been identified and associated with genes. These diseases are caused by exposure to high penetrance genotypes and are often phenotypically extreme, which facilitates case ascertainment and genetic mapping. However, it is known that Mendelian disease genes can harbor mutations that modulate risk for common, complex disease that are similar to the more severe syndrome. We searched for this phenomenon using the Human Phenotype Ontology in over 25,822 European ancestry individuals genotyped on the Exome BeadChip in the Vanderbilt BioVU resource. BioVU is a DNA biorepository linked to an extensive electronic medical records database. We classified individuals as cases or controls for 513 phenotypes defined using PheWAS codes (curated abstractions of ICD-9 diagnosis codes). The Human Phenotype Ontology (HPO) has been mapped to both Orphanet and OMIM. We mapped HPO traits to PheWAS traits to define clusters of PheWAS traits that are subphenotypes of a Mendelian syndromic trait. For example, thyroid dyshormonogenesis caused by the *DUOX2* gene is defined in PheWAS traits for hypothyroidism, Goiter, mild cognitive impairment, etc.We created a syndrome score for each individual for all HPO traits by summing the number of PheWAS traits, weighted by the inverse of the incidence in our population. We regressed the syndrome score on to dominantly coded genotypes in a logistic model adjusted for age and sex using rare (MAF<5% and >0.1%) nonsynonymous variants within the Mendelian gene.Results:Eight gene-syndrome pairs crossed the Bonferroni threshold 1.0e-5 (4,806 tests). Some of these were replications. Three of these associations were related to a mutation on the *JAK2* gene and Polycythemia vera and Essential Thrombocythemia (p-values 6.4e-27 to 1.3e-51). Alpha-1 antitrypsin deficiency was associated with *SERPINA1* (p=2.3e-6). The remaining significant results were novel and included Joubert syndrome and KIF7 (p=4.3e-6) characterized by cognitive impairment, abnormal gait, strabismus, and four other traits; and Keratitis-ichthyosis-deafness syndrome and GJB2 (p=8.1e-6), characterized by Keratoconjunctivitis sicca, Hypohidrosis, Oral leukoplakia, and six other traits; and Bohring-Opiz syndrome and *ASXL1* (p=7.1e-6), defined by Mental retardation, Exophthalmos, Seizures, Hirsutism, and 19 other traits.Conclusion:We replicated known associations and identified several novel loci associated with syndromes defined by Mendelian diseases.

133

PheWAS study using research participants' self -reported data provides insight into Th17/IL-17 pathway. *M.G. Ehm[1], J.L. Aponte[3], S.F. Cook[3], S. Ghosh[1], A. Gupta[1], D.A. Hinds[4], C.G. Lariminie[2], L. Li[3], T. Johnson[2], C. Tian[4], S.C. Zelt[3], D. Rajpal[1], D.M. Waterworth[1].* 1) Target Sciences, GlaxoSmithKline, King of Prussia, PA; 2) Target Sciences, GlaxoSmithKline, Stevenage, UK; 3) PAREXEL International, Research Triangle Park, NC; 4) 23andMe, Inc., Mountain View, CA.

T helper 17 cells (Th17) and their signature cytokine, IL-17, play a role in multiple autoimmune diseases. There are promising drugs targeting this pathway, recently approved and in development for dermatology and immuno-inflammation diseases. We performed a phenome-wide association study (PheWAS) to provide insight into pathway mechanisms and to identify traits not previously shown to be influenced by variants affecting the Th17/IL-17 pathway. The PheWAS was performed for known or putatively functional variants in *IL17A*, *IL17F*, *IL17RB*, *IL23R*, *RORC*, *TRAF3IP* and *TYK2 (*one in each gene). The seven variants were analyzed across 1255 traits consisting of research participants' self-reported phenotypes in the 23andMe database of 700,000 research participants, including more cases of psoriasis, eczema, acne and rosacea, than in previous studies. The threshold for statistical significance was determined by Bonferroni correction for independent traits ($p=0.05/1000=5x10^{-5}$). The results replicated known associations with several autoimmune phenotypes, illustrating that participants' self-reported outcomes can be a surrogate for clinically assessed conditions. There were several statistically significant novel associations, including: (i) association of allergy phenotypes with rs4845604, a variant known to have a regulatory effect on *RORC* expression, which may be indicative of counter regulation of Th1/Th2 and Th17/IL-17 pathways; (ii) association of throat infections with the *TYK2* Ile684Ser variant, validating a previous hypothesis that individuals carrying *TYK2* variants might be at increased risk of serious infection due to partial inhibition of *TYK2* (Diogo *et al* PMID: 25849893); and (iii) an association of dandruff, with *IL23R* Arg38Gly. Suggestive associations (FDR<0.1) included (a) associations of eczema and rosacea with *IL23R* Arg38Gly; (b) associations of stroke and migraine with *RORC* rs4845604; and (c) associations of male pattern baldness and acne with *TRAF3IP* Asp10Asn. The associations of eczema, stroke, migraine, male pattern baldness and acne will be studied in independent datasets. In summary, the 23andMe database enabled a rapid study of dermatological and immuno-inflammation phenotypes in large numbers as well as the study of phenotypes like allergies and infections that would likely not have been considered a priori and haven't been comprehensively studied.

134

Functional Variant PheWAS in 30,000 exomes using EHR-extracted Laboratory Measures in the Geisinger Health System MyCode - Regeneron Genetics Center Collabrative Project DiscovEHR. *A. Verma[1], J. Leader[2], S. Dudek[1], J.R. Wallace[1], C. Dushlaine[3], C. Van Hout[3], L. Haebagger[3], A. Lopex[3], F.E. Dewey[3], O. Gottesman[3], J. Overton[3], J.G. Reid[3], A. Baras[3], A.R. Shuldiner[3], D.J. Carey[4], J.L. Kirchner[2], D.H. Ledbetter[6], M.D. Ritchie[5], S.A. Pendergrass[5].* 1) Center for Systems Genomics, The Pennsylvania State University, State College, PA; 2) Center for Health Research, Geisinger Health System, Danville, PA; 3) Regeneron Genetics Center, Tarrytown NY; 4) Weis Center for Research, Geisinger Health System, Danville, PA; 5) Biomedical and Translational Informatics, Geisinger Health System, Danville, PA.

Phenome-Wide Association studies (PheWAS) using electronic health records (EHRs) have mainly evaluated associations between genetic variants and case/control status derived from international classification of disease (ICD) codes. The rich resource of clinical laboratory measures collected within the EHR can be used for high-throughput PheWAS analyses and discovery. Using the Geisinger Clinic MyCode biorepository, we have developed a novel and sound methodology for extracting a wide range of high-quality laboratory measures. Using this approach as proof-of-principle, we extracted 23 different clinical laboratory tests from more than 50,000 participants of the MyCode biorepository, and calculated the median result value from these laboratory test results for ~30,000 subjects matching samples with whole exome sequencing from the Geisinger Health System – Regeneron Genetics Center DiscovEHR project. Tools for evaluating the functionality of SNPs can be leveraged for identifying SNPs more likely to be functional, thus we identified 1,430 SNPs with MAF > 0.01 that are putative loss of function variants. We evaluated the association between these SNPs and 23 clinical laboratory measures along with 1,706 ICD-9 based case/control diagnoses. We replicated known associations, and identified potentially novel associations, as well as potential pleiotropy. For example, we found an association between the stop-gain *PARVB* SNP rs201332415 and the laboratory measures of aspartate aminotransferase ($p = 3.92 \times 10^{-12}$), alanine aminotransferase ($p = 9.65 \times 10^{-12}$), and platelet count ($p = 1.95 \times 10^{-5}$), as well as ICD-9 derived diagnoses of "571.8 other chronic non-alcoholic liver disease" ($p = 8.67 \times 10^{-6}$) and "794.8 nonspecific abnormal results of liver function study" ($p = 3.57 \times 10^{-5}$), all associations with the same direction of effect and all related measures for liver disease. Associations were also identified between the stop-gain *CCHCR1* SNP rs3130453, and "696.1 other psoriasis" ($p = 4.27 \times 10^{-8}$), "250.01 type 1 diabetes" ($p = 6.98 \times 10^{-7}$), additional ICD-9 diabetes related codes, as well as the diabetes related laboratory measure of hemoglobin A1c ($p = 2.53 \times 10^{-4}$). In our future studies we will apply our clinical laboratory extraction method to access thousands of phenotypes from the EHR, and evaluate SNPs and rare variants within the PheWAS framework for novel hypothesis generation and greater understanding of the impact of genetic architecture on diagnoses and traits.

135

Exome-wide study identifies loci displaying pleiotropic associations with multiple cardiometabolic traits in continental Africans. *F. Tekola-Ayele[1], A. Adeyemo[1], A.R. Bentley[1], A.P. Doumatey[1], J. Zhou[1], G. Chen[1], D. Shriner[1], C. Adebamowo[2], J. Achaempong[3], J. Oli[4], O. Fasanmade[5], T. Johnson[5], A. Amoah[6], K. Agyenim-Boateng[3], B.A. Eghan Jr.[3], D. Ngare[7], C.N. Rotimi[1].* 1) Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 2) School of Medicine, University of Maryland, MD; 3) University of Science and Technology, Kumasi, Ghana; 4) University of Nigeria Teaching Hospital, Enugu, Nigeria; 5) University of Lagos, Lagos, Nigeria; 6) University of Ghana Medical School, Accra, Ghana; 7) Moi University, Eldoret, Kenya.

Clustering of cardiometabolic abnormalities displays high heritability. However, the underlying physiological dysregulations behind these shared abnormalities remain unresolved. Here we aim to identify genetic loci with pleiotropic effects on cardiometabolic traits and inflammation markers in the largest sample (n=4,218) of continental Africans genotyped to-date on the Illumina HumanExome BeadChip v1.1 and the Affymetrix Axiom Exome 319® Array. After filtering-out variants that were non-exonic, with minor allele frequency <0.01, genotype call rate <0.98, and Hardy-Weinberg equilibrium $P <= 10^{-6}$, a total of 32,889 exonic variants in 11,049 genes were retained for analysis. Gene-based analysis was performed with SKAT on 16 quantitative traits (including levels of albumin, creatinine, HDL, LDL, cholesterol, BMI, glucose, insulin, urea, uric acid, triglycerides, total protein, waist circumference, waist to hip ratio, systolic blood pressure, and diastolic blood pressure), each adjusted for age, sex, the first three principal components and genotyping chip. We found that *A2ML1*, *COL4A1*, *PARVB*, *TTN*, and *CDK5RAP2* genes showed Bonferroni-corrected significant associations ($P < 4.53 \times 10^{-6}$) with five or more traits including albumin, HDL, insulin, triglycerides, systolic blood pressure, and uric acid levels. The encoded proteins of these genes have important cardiometabolic roles: A2ML1 is a ligand of the well-known low density lipoprotein receptor-related protein 1 (LRP1) that plays an important role in lipid metabolism; over-expression of PARVB enhances PPARG activity and lipogenesis; COL4A1 is expressed in the vascular tissue. Three of these loci (*COL4A1*, *TTN* and *PARVB*) were previously reported to be associated with coronary artery disease and arterial stiffness (*COL4A1*), hip circumference and lipid traits (*TTN*), and non-alcoholic fatty liver disease and serum triglycerides (*PARVB*) in Europeans and Asians. In all, we identified five loci displaying pleiotropic associations with several cardiometabolic traits providing insight into the pathogenesis of the clustering of metabolic disorders in Africans and other human populations.

136

Trans-ethnic meta-analysis reveals novel loci and effector genes for kidney function in diverse populations. *A. Morris[1,2], A. Mahajan[2], J. Haessler[3], Y. Okada[4], A. Stilp[5], J. Whitfield[6], C. Laurie[5], N. Franceschini[7].* 1) Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom; 2) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 3) Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; 4) Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan; 5) Department of Biostatistics, University of Washington, Seattle, WA; 6) QIMR Berghofer Medical Research Institute, Brisbane, Australia; 7) University of North Carolina, Chapel Hill, NC.

Chronic kidney disease (CKD) is a major public health problem in diverse racial populations. Reduced estimated glomerular filtration rate (eGFR) is a measure of kidney function used to define CKD. Genome-wide association studies (GWAS) have been successful in identifying loci for eGFR, primarily in individuals of European ancestry. However, these loci typically map to large genomic intervals, limiting progress in identifying causal transcripts and understanding the downstream pathogenesis of CKD. To address these challenges, we performed trans-ethnic meta-analysis to: (i) discover novel eGFR loci; and (ii) fine-map eGFR loci by leveraging differences in linkage disequilibrium between diverse populations.We considered 9 GWAS including 71,638 individuals of European, African American, Hispanic, and East Asian ancestry, each imputed up to the 1000 Genomes Project reference panel (March 2012 release). Within each study, association with eGFR was tested under an additive model. We combined association summary statistics across studies using fixed-effects meta-analysis for discovery. We identified 20 loci at genome-wide significance ($p<5.0x10^{-8}$), including two not previously reported: *LRP2* ($p=5.6x10^{-10}$) and *NFATC1* ($p=1.3x10^{-8}$). *LRP2* encodes megalin, a kidney epithelial receptor known to be involved in the uptake of filtered nutrients, hormones and other compounds.We constructed “credible sets” of variants that account for 99% of the posterior probability of driving the association signal at each locus. We resolved fine-mapping to less than 10 variants at 11 loci. At two loci, the credible set included protein altering variants: *GCKR* P446L and *CPS1* N1412T. At the remaining 9 loci, the credible set mapped only to non-coding sequence, suggesting that variants driving these association signals impact eGFR through regulatory mechanisms. For example, at the *RGS14-SLC34A1* locus, the two variants in the credible set overlap promoter histone marks and DNAse hypersensitivity sites in multiple cell types. The lead SNP, rs35716097, is an eQTL for *RGS14* in multiple tissues, highlighting this gene as the likely effector transcript at this locus.Our findings provide evidence that trans-ethnic GWAS across diverse populations at high risk for CKD are useful for discovery of novel loci and for prioritisation of potential causal variants that can be taken forward for experimental validation, thereby enhancing our understanding of the pathophysiology of kidney function.

137

How low can you go: cohort-wide 1x whole genome sequencing in a Greek isolate reveals multiple quantitative trait signals. *A. Gilly[1], L. Southam[1,2], R. Moore[1], A.-E. Farmaki[3], J. Schwartzentruber[1], P. Danecek[1], E. Tsafantakis[4], G. Dedoussis[3], E. Zeggini[1].* 1) Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; 2) Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK; 3) Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece; 4) Anogia Medical Centre, Crete, Greece.

Very low-depth sequencing has been proposed as a potential means to increase sample sizes in next-generation association studies while keeping sequencing costs low. Simulations have shown that the increase in power provided by a greater sample size outweighs the loss of accuracy caused by reduced depth. Isolated populations offer power gains in detecting associations in rare and low-frequency variants. Here, we conduct sequence-based association studies in an isolated cohort from Crete, Greece using very low depth (1x) whole genome sequencing (WGS). We first established a reference pipeline for the analysis of 1x WGS data. Following benchmarking of 15 different pipelines involving state-of-the-art imputation tools (BEAGLE, IMPUTE2, MaCH/Thunder, MVNCall), we find that a two-pass BEAGLE approach using a large reference panel can detect over 80% of true low-frequency (1%<MAF<5%) variants and 100% of common-frequency variants, with an average minor allele concordance of 90% across the allele frequency spectrum. We carried out single-point association of 29,616,352 variants with 42 medically-relevant quantitative traits, including anthropometric (8), lipid (5), glucose and insulin-related (12), cardiometabolic and haematological (12), thyroid-related (2), liver function (2) and bone formation (1) measurements. We find 57 independent genome-wide significant ($p<1x10^{-8}$) signals (binomial $p<1x10^{-80}$) in the HELIC-MANOLIS cohort of 1,239 individuals, 4 of which are shared across two or more traits. Two thirds of these signals arise from variants with frequency <1%; 86% of all associated variants are intronic or outside of genes. Initial analysis indicates a mixture of known and novel loci. For example, we replicate association of the rare R19X variant in *APOC3* with blood triglyceride levels ($\beta=-1.09,\sigma=0.164$, $p=1.01x10^{-10}$). We find it to contribute to a rare variant burden in *APOC3* ($p=3.0x10^{-18}$), which remains genome-wide significant after excluding R19X ($p=6.15x10^{-10}$). This strong burden signal was not recapitulated in an analysis of GWAS data in the same cohort imputed up to a combined reference panel of 4,873 sequenced individuals from the 1000 Genomes Project and UK10K, as well as 250 individuals from the same population with WGS at 4x depth. We present one of the first population-based next-generation association studies based on very low depth WGS, and demonstrate the advantages of this approach over a mixed GWAS and imputation study design.

138

Whole genome sequencing increase the power to detect trait-associated rare variants shifted towards high frequencies in the Sardinian island population. *C. Sidore[1], M. Zoledziewska[1], F. Danjou[1], F. Busonero[1], A. Maschio[1], E. Porcu[1], A. Mulas[1,2], C. Chiang[3], G. Pistis[1], M. Steri[1], S. Naitza[1], M. Pitzalis[1], J. Marcus[4], R. Nagaraja[5], A. Angius[1,6], J. Novembre[4], S. Sanna[1], D. Schlessinger[5], G. Abecasis[7], F. Cucca[1,2].* 1) IRGB CNR, Cagliari, Cagliari, Italy; 2) Università degli Studi di Sassari, Sassari, Italy; 3) Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA, USA; 4) Department of Human Genetics, University of Chicago, IL, USA; 5) Laboratory of Genetics, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA; 6) Center for Advanced Studies, Research, and Development in Sardinia (CRS4), AGCT Program, Parco Scientifico e tecnologico della Sardegna, Pula, Italy; 7) Center for Statistical Genetics, Ann Arbor, University of Michigan, MI, USA.

Large sequencing studies based on next generation technologies profoundly improved the resolution of genome wide association studies. Although such an approach allowed for the detection of a large number of rare variants, enough statistical power is needed to successfully identify their associations with strong statistical support. Families and founder populations, where variants rare elsewhere can occur at moderate frequencies, provide an alternative to large meta-analyses and help overcome these limitations. Here we performed a large whole genome sequencing study of 3,514 individuals from the Sardinian population, whose demographic history provides a unique opportunity to study the effect of variants enriched due to isolation or selection. We detected >23M single nucleotide polymorphisms (SNPs) and generated a reference panel for imputation in 6,602 individuals genotyped at ~900K SNPs. This approach allowed us to reach extremely high imputation accuracy even for low frequency variants (r^2 with directly measured genotypes=0.90 for variants with MAF 0.5-5%). The effects of isolation are clear from the extent of genetic differentiation with mainland Europeans (allele sharing ratio <0.6 for MAF 1-5%), and in the significantly higher deleteriousness of variants enriched in Sardinians (p=0.02). We then performed GWAS scans on several traits: height, 4 lipid levels (LDL, HDL, TG and TC), 5 inflammatory markers (ESR, hsCRP, Adiponectin, MCP-1, IL-6) and 3 hemoglobin levels (HbA1, HbA2, HbF). Overall, we identified 58 independently associated variants including 18 variants not previously described in prior GWAS. The advantages of analyzing this founder population are particularly evident in two scenarios: a) signals with strong effect and that are extremely rare in Europe (MAF<0.01%) but enriched in Sardinia (MAF-0.5-5%) such as *APOA5* associated with TG, *GHR* associated with height and a long stretch of variants on chromosome 12 region associated with hsCRP and ESR; b) signals rare in Europe (MAF <1%) and common in Sardinia (MAF >5%) such as *CCDN3* associated with HbA2 and *KCNQ1* associated with height. Overall, these results demonstrate the benefits of our sequencing-based approach for the discovery of the effects on phenotypes of rare variants enriched in the informative population of Sardinia.

139

Phased Annotation of Protein-Coding Variants Across 60,706 Human Exomes. *A.J. Hill[1,2], B. Cummings[2,3,4], K.J. Karczewski[2,3], M. Lek[2,3], D.G. MacArthur[2,3], Exome Aggregation Consortium (ExAC).* 1) Genome Sciences, University of Washington, Seattle, WA; 2) Broad Institute, Cambridge, MA; 3) Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 4) Harvard Medical School, Boston, MA.

Nearly all variant interpretation pipelines currently ignore local phasing of variants. In order to demonstrate the importance of phasing on variant interpretation, we have identified multi-nucleotide polymorphisms (MNPs) that change interpretation of protein coding variation in 60,706 human exomes as part of the Exome Aggregation Consortium. We found approximately 6,000 MNPs (average 23 per sample) where analysis of the underlying SNPs without phasing will result in incorrect functional interpretation. These include cases where the effect of a loss of function (LOF) variant is eliminated by an adjacent SNP (Rescued LOF) or where underlying synonymous or missense variants result in LOF MNPs (Gained LOF). Using these annotations, we identified rescued LOF MNPs in *MLH1* and *FANCA*, where LOF variants are associated with autosomal dominant Lynch syndrome risk and autosomal recessive Fanconi Anemia, respectively. We also identified one sample as a carrier for a gained LOF MNP in *MUTYH*, where homozygous or compound heterozygous LOF variants have been tightly linked to MUTYH-associated polyposis. Additionally our analysis revealed 10 MNPs that have previously been reported as disease causing mutations in HGMD as deletion-insertions. One such MNP created a stop gain in *COH1*, which has previously been identified in a Cohen Syndrome patient, rendering the individuals carriers of an autosomal recessive disease. These variants would be missed by virtually all currently available next generation sequencing pipelines. These results represent the first large-scale analysis of the functional impact of MNPs, and illustrate the importance of considering local phase information in variant calling and clinical annotation.

140

Meta-analysis of more than 2,100 trios reveals novel genes for intellectual disability. *C. Gilissen[1], S.H. Lelieveld[1], M.R.F. Reijnders[1], R. Pfiundt[1], H. Yntema[1], P. de Vries[1], B.A. de Vries[1], T. Kleefstra[1], M. Nelen[1,2], J.A. Veltman[1,2], H.G. Brunner[1], C. Vissers[1].* 1) Human Genetics, Radboud University Nijmegen Medical Center, Nijmegen, Netherlands; 2) Department of Clinical Genetics, Maastricht University Medical Centre, Maastricht, Netherlands.

Trio-based whole exome sequencing (WES) studies of large cohorts have shown the power of identifying novel disease genes for sporadic disorders including various neurodevelopmental disorders such as autism spectrum disorder, intellectual disability and epilepsy. As part of routine genetic diagnostics at the Radboud university medical center, we sequenced the exomes of 820 patients with intellectual disability and their unaffected parents and identified 1,184 de novo mutations. Mutations in this cohort are significantly enriched for loss-of-function variants (P = 2.019e-12) as well as recurrent gene mutations (P < 1e10-6). Based on gene specific mutation rates we performed a statistical analysis for recurrent de novo gene mutations and identified 28 genes to be significantly enriched for either loss-of-function or functional de novo mutations, of which 6 were not previously implicated in ID. To increase statistical power, we combined our results with those of four published trio-based WES studies of ID cohorts. In the combined cohort of 2,104 trios, we performed a meta-analysis and we identified an additional 48 genes of which 12 were novel, giving rise to a total of 15 unique novel ID genes. Notably, for three novel genes we identified only recurrent missense mutations rather than loss-of-function mutations. In support of our results, the same analyses on a published cohort of control trios yielded no significant results. Interestingly, assessment of the function of these genes in pathways associated with intellectual disability, showed that four of the novel genes have a role in synaptic Wnt-signaling and two genes in TGFβ signaling. Assessment of the clinical phenotypes of patients with de novo mutations in these 15 novel genes shows clear phenotypic overlap between individual cases with the same gene affected, thereby corroborating our findings. Moreover, analysis for tolerance to population variation in these 15 novel genes shows the identified genes to be similarly intolerant as >650 well-established genes for intellectual disability. In summary, we identified 15 novel ID genes by performing a meta-analysis on WES data of >2,100 intellectual disability trios, highlighting the potential of unbiased statistical analyses of large trio-based sequencing studies to identify novel genes for sporadic diseases.

141

Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool. *C. Douville[1], D.L. Masica[1], P.D. Stenson[3], D.N. Cooper[3], D. Gygax[4], R. Kim[4], M. Ryan[4], R. Karchin[1,2].* 1) Biomedical Engineering, Johns Hopkins University, Baltimore, MD; 2) Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD; 3) Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK; 4) In Silico Solutions, Fairfax, VA, USA.

The average human exome contains over four hundred naturally occurring insertion/deletion variants. Approximately half of these variants disrupt the translational reading frame. Many researchers have assumed that insertion/deletions are predominantly pathogenic, but if that were true, we would expect a high prevalence of inherited disease in the general population. Bioinformatics methods that predict whether or not insertion/deletion variants are pathogenic are important for high-throughput mutation analysis pipelines, and several novel methods have been proposed. Here, we introduce a new Random-Forest based bioinformatic method—the Variant Effect Scoring Tool for Insertions and Deletions (VEST-indel). VEST-indel has high balanced accuracy when applied to both in-frame and frameshift insertion/deletions (0.883 and 0.900, respectively). We also show how several existing methods to predict pathogenic insertion/deletions can be combined into a meta-predictor, which is particularly useful for in-frame variants, representing a significant methodological advance over other methods. We did not find any evidence to support the view that selective pressure to maintain deleterious variants at higher than expected frequencies can explain the low specificity of current methods that predict pathogenic frameshifts.

142

Deep genetic connection between cancer and developmental diseases. *H. Qi[1,4], C. Dong[2,3], K. Wang[2,3], Y. Shen[4,5].* 1) Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA; 2) Biostatistics Division, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA; 3) Zilkha Neurogenetic Institute, University of Southern California, Los Angeles, CA 90089, USA; 4) Departments of Systems Biology and Biomedical Informatics, Columbia University Medical Center, New York, NY, USA; 5) JP Sulzberger Columbia Genome Center, Columbia University Medical Center, New York, NY, USA.

Cancer and developmental diseases share disrupted cellular processes like cell proliferation, growth, and differentiation. Underpinning these processes are common molecular pathways and genes. Indeed, recent large-scale genomic studies of cancer and developmental diseases, such as autism, intellectual disability, epilepsy, and developmental delay, revealed a substantial number of genes implicated in both classes of diseases, with recurrent somatic mutations in cancer and highly penetrant germline *de novo* mutations in developmental diseases. In this study, we aim to quantify the genetic connection between cancer and developmental diseases. We compiled a large set of *de novo* mutations from about 5500 developmental disease cases in recent published studies and quantify the deep connection from three aspects. There is a significant enrichment of germline *de novo* mutations in developmental disease patients among high-confidence cancer driver genes (enrichment 2.6x; p=0.0002), especially the ones that are regulated during development (enrichment 3.8x; p=1.5e-20). We estimated that these cancer drivers made up to about a third of risk genes contributing to developmental diseases. The implicated pathways include every level of gene regulation, such as genome organization, chromatin modification, transcription, pre-mRNA processing, post-translational modification, and signaling cascades. We further inferred somatic missense mutation hotspots in cancer using a hidden Markov Model, and observed that missense variants in developmental diseases were often located in these hotspots, which indicates these genes (such as *PTPN11*, *CTCF*, *PPP2R1A*) and pathways are disrupted through similar molecular mode of action between cancer and developmental diseases. By leveraging the vast amount of somatic mutations observed in cancer patients, we can improve our ability to identify causal mutations or variants in patients with developmental diseases.

143

Characterizing Ribosome Readthrough in Humans. *J. Moore, Z. Weng.* Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA.

Ribosome readthrough has been reported in several model organisms, but has not been extensively studied in humans. Purely computational predictions of readthrough events in humans have been unsuccessful in identifying more than a handful of readthrough candidates. However, studies have demonstrated it is possible to predict ribosome readthrough using ribosome profiling data. Our goal was to identify ribosome readthrough events in humans from publically available ribosome profiling data in several cell and tissue types. We first developed an extensive computational pipeline to identify candidate ribosome readthrough genes. We considered the "leakage" rate across stop codons as well as the phasing of reads in the 3' UTR. We used our pipeline to analyze data from several cell and tissue types including 72 Yoruban lymphoblast cell lines (LCLs), normal and tumorigenic brain tissue, and fibroblasts. We predicted over 40 high confidence candidate readthrough genes, with candidates in every cell types. Several of our predictions, including *AQP4* and *BRI3BP,* support previous readthrough predictions based on conservation and *in vitro* experiments. When we then further analyzed readthrough candidates from LCLs, we found the genes were enriched in ontology terms related to extracellular processes and the cell membrane. Readthrough candidates were also significantly enriched with the UGA stop codon, previously associated with readthrough. Of particular interest was *CNNM4*, a gene with a UGA stop codon and a weak selenocysteine element in its 3' UTR. The SECIS-like element could lead to the observed high level of readthrough which results in the addition of an AKT kinase phosphorylation site to the protein's C-terminus. Similarly, we were able to identify functional protein motifs in the C-terminal extension of a quarter of the readthrough genes suggesting readthrough may have important biological roles.We then analyzed common genetic variants in the proximal 3' UTRs of candidate readthrough genes. We did not find any variants directly associated or in linkage disequilibrium with disease associated variants. However these genes were enriched in common variants in their proximal 3' UTRs compared to non-readthrough protein coding genes.Our data suggests that ribosome readthrough may have an important biological role in humans and should be further characterized to understand its full biological impact. .

144

A comprehensive methodology for assessing variant-specific gene dysfunction in the context of non-disease-associated genomes. *M.J. Silver[1], J.L. Larson[1], A.J. Silver[1], C. Borroto[1], B. Spurrier[1], R.M. Lim[1], N. Delaney[1], L.M. Silver[1,2].* 1) GenePeeks, Inc., New York, NY 10014 and Cambridge, MA 02142; 2) Department of Molecular Biology and the Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton NJ, 08544.

Background: Carrier testing is unique among medical diagnostics in that recessive disease liability is only predicted to occur in persons other than those actually being tested. The utility of carrier testing is compromised by multiple factors. First is the failure of a binary classification system (with "positive" and "negative" endpoints) to illuminate allele- or genotype-specific differences in predicted phenotypes. Second is the inability of clinical testing panels to include the majority of deleterious variants which have not been ascertained in clinical or functional studies. Third is a reliance on clinical classifications that sometimes contravene biological plausibility.**Methods:** We developed a dynamic algorithm to generate Variant-specific Gene Dysfunction (VGD) scores along a continuous scale from zero to complete effect on gene function. Our computational model uses a neural network-inspired approach to integrate quantitative and categorical data from multiple information sources. This tunable and extendable network provides the framework for balancing oppositional assessments of gene dysfunction. Currently included in the model are mutation class, clinical classifications of pathogenicity, four computational models of evolutionary and structural constraints on variant proteins, and population-specific allele and genotype frequency data from more than 60,000 individuals without pediatric disease in the Exome Aggregation Consortium (ExAC) dataset. After completion of the independent variant scoring phase, we use clinically classified variants as "standard candles" to validate our method and calibrate scores on a gene-by-gene basis.**Results:** We calculated VGD scores for over 250,000 variants located in genes associated with recessive disease. The median clinically naive VGD score of variants correlated with disease is above 0.98, while benign variants have a median clinically naive VGD score of less than 0.01.**Conclusions:** The primary goal of the methodology presented here is to estimate a variant's impact on the functionality of a single copy of a recessive disease gene in the context of a diploid organism. By focusing on the haploid-level molecular phenotype rather than a patient's disease state, we overcome the unrealistic imperative of a binary classification system and better estimate the risk of conceiving a child with disease.

145

A Massively Parallel Pipeline to Clone DNA Variants and Examine Molecular Phenotypes of Human Disease Mutations. *H. Yu.* Biological Statistics and Computational Biology, Cornell University, Ithaca, NY.

Understanding the functional relevance of DNA variants is essential for all exome and genome sequencing projects. However, current mutagenesis cloning protocols require Sanger sequencing, and thus are prohibitively costly and labor-intensive. We describe a massively-parallel site-directed mutagenesis approach, "Clone-seq", leveraging next-generation sequencing to rapidly and cost-effectively generate a large number of mutant alleles. Our innovative Clone-seq pipeline can generate verified clones for >3000 mutations using one lane of 1×100 bp Illumina HiSeq sequencing with a >10-fold overall cost reduction. Clone-seq is entirely different from previously described random mutagenesis approaches. In Clone-seq, each mutant clone has a separate stock and different clones can therefore be used separately for completely independent downstream assays; we can generate targeted mutations on *thousands of genes*. In random mutagenesis, a pool of sequences containing different mutations for *one gene* is generated. Therefore, it is not possible to separate one mutant sequence from another and the whole pool can only be used for the same assay(s) together. Using Clone-seq, we further develop a comparative interactome-scanning pipeline integrating high-throughput GFP, yeast two-hybrid (Y2H), and mass spectrometry assays to systematically evaluate the functional impact of mutations on protein stability and interactions. We use this pipeline to show that disease mutations on protein-protein interaction interfaces are significantly more likely than those away from interfaces to disrupt corresponding interactions. We also find that mutation pairs with similar molecular phenotypes in terms of both protein stability and interactions are significantly more likely to cause the same disease than those with different molecular phenotypes, validating the *in vivo* biological relevance of our high-throughput GFP and Y2H assays and indicating that both assays can be used to determine candidate disease mutations in the future. The general scheme of our experimental pipeline can be readily expanded to other types of interactome-mapping methods to comprehensively evaluate the functional relevance of all DNA variants, including those in non-coding regions.

146

Gene Discovery in Mendelian Diseases. *D. Vuzman[1,2], N.Y. Frank[1], N. Stitziel[1,3], S. Chopra[1], S.R. Sunyaev[1,2], R.L. Maas[1], Brigham Genomic Medicine Program (BGMP).* 1) Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 2) Broad Institute of MIT in Harvard, Cambridge, MA; 3) Cardiovascular Division, Washington University School of Medicine, St. Louis, MO.

Brigham Genomic Medicine Program (BGMP) is an interdisciplinary collaboration among clinical geneticists, computational biologists and physicians throughout BWH to identify the genetic basis of Mendelian disorders of uncertain etiology. These disorders comprise ~7,000 diseases that in the aggregate contribute significantly to disease burden, both at Brigham and elsewhere. Mendelian diseases also provide unique opportunities to understand disease etiology in individual patients, as well as in more common forms of the same diseases. In turn, direct knowledge of the causal mutation may resolve long-standing, expensive diagnostic dilemmas and provide newfound knowledge that advances our understanding of disease pathobiology. BGMP has developed a unique genomic analysis pipeline that won the international 2012 CLARITY Challenge, and we have now successfully used genomic sequencing to identify novel disease-causing genes in BWH patients. The findings from these cases include high impact discoveries of disease-associated genes, such as *PIEZO2*, where we discovered that gain-of-function missense mutations in the mechanically activated ion channel of *PIEZO2* cause a subtype of Distal Arthrogryposis. In another case, we used sequence analysis to diagnose and discover the genetic etiology of disease in a 34 year old male with a history of precocious polyarthritis that had been previously attributed to rheumatoid arthritis, in the *WISP3* gene. We have also identified highly promising disease-gene candidates for *C3, LOX*, and *XRCC5* which are currently undergoing validation. Gain- and loss-of-function mutations in these genes are associated with high impact genetic disorders, including idiopathic acrocyanosis, cardiovascular disease, and types of lung cancer.

147

An important role for rare loss-of-function variants in spermatogenic failure. *R. George[1], J. Hughes[1], L. Brown[1], L. Lin[2], D. Koboldt[2], R. Fulton[2], R. Oates[3], S. Silber[4], R. Wilson[2], D. Page[1,5,6].* 1) Whitehead Institute, Cambridge, MA; 2) McDonnell Genome Institute, Washington University, St. Louis, MO; 3) Department of Urology, Boston University School of Medicine, Boston, MA; 4) Infertility Center of St. Louis, St. Luke's Hospital, St. Louis, MO; 5) Howard Hughes Medical Institute, Chevy Chase, MD; 6) Department of Biology, Massachusetts Institute of Technology, Cambridge, MA.

Severe spermatogenic failure is the production of very little to no sperm and affects roughly 1% of the male population. About 30% of cases can be explained by cytogenetic abnormalities of the sex chromosomes and microdeletions of the Y chromosome. However, for the remaining 70% of cases, the underlying cause remains unknown. We hypothesized that a large fraction of unexplained cases are caused by rare deleterious variants scattered across many independent loci. To test this hypothesis, we sequenced the sex chromosomes and 500 candidate autosomal genes in 284 men with unexplained non-obstructive azoospermia and 289 men with normal sperm counts. Men with non-obstructive azoospermia have a significant excess of rare loss-of-function variants in genes related to spermatogenesis compared to our controls (OR = 3.58, $p = 3.9 \times 10^{-04}$) or to an independent cohort of unphenotyped individuals (OR = 2.99, $p = 8.9 \times 10^{-06}$). Further, our results suggest that different sets of genes are disrupted in different clinical subtypes of non-obstructive azoospermia. In particular, meiotic genes have more rare loss-of-function variants in men with testicular maturation arrest than in men with Sertoli-cell-only syndrome ($p = 0.011$). Collectively, we estimate that rare loss-of-function variants in spermatogenesis genes explain up to 7% of unexplained cases of non-obstructive azoospermia and are present in ~1 in 2,000 men who are otherwise healthy.

148

Complex mitotic-origin aneuploidy in human embryos: genetic risk factors and fertility consequences. *R.C. McCoy[1], Z. Demko[2], A. Ryan[2], M. Banjevic[2], M. Hill[2], S. Sigurjonsson[2], M. Rabinowitz[2], H.B. Fraser[1], D.A. Petrov[1].* 1) Department of Biology, Stanford University, Stanford, CA; 2) Natera, Inc., San Carlos, CA.

Aneuploidy—extra or missing chromosomes compared to a balanced 46-chromosome complement—affects three-quarters of human embryos at the cleavage stage. Aneuploid embryos rarely survive to term. We applied 24-chromosome SNP-microarray-based preimplantation genetic screening (PGS) to 46,439 embryos from 6366 *in vitro* fertilization (IVF) cycles, assigning copy number variations to specific parental homologs and achieving a high-resolution view of embryonic aneuploidy. Using distinct chromosomal signatures, we classified aneuploidies of maternal meiotic, paternal meiotic, and mitotic (postzygotic) origin without embryo disaggregation, facilitating separate investigation of the cytogenetic mechanisms underlying their formation and the consequences for early development. Our data revealed an extreme diversity of aneuploidies in day-3 blastomere biopsies. A common form of catastrophic, putative mitotic-origin aneuploidy involved multiple forms of chromosome loss (maternal monosomy, paternal monosomy, and nullisomy), presumably due to centrosome and mitotic spindle aberrations leading to chromosomal chaos. These complex aneuploidies were comparatively rare in day-5 embryo biopsies, suggesting strong selection following zygotic genome activation at the 4-8 cell (day-3) stage. In line with this hypothesis, we found that patients referred for PGS due to previous IVF failure had significantly higher rates of these mitotic-origin aneuploidies than patients referred for other reasons. This finding also suggested that uncharacterized environmental and genetic factors influence aneuploidy risk. Seeking to uncover these factors, we performed a genome-wide association study of embryonic aneuploidy. We identified a strong association between putative mitotic-origin aneuploidies and common (~30% global minor allele frequency) maternal genetic variants on chromosome 4. The associated region spans over 600 Kb and encompasses the gene *PLK4*, a strong causal candidate given its well-described role in mediating centriole duplication and its ability to induce chromosome mis-segregation upon minor dysregulation. Intriguingly, the associated region also displays signatures of a selective sweep in ancient humans, preceding out-of-Africa migrations. Together, these findings shed light on novel characteristics of aneuploidies affecting preimplantation embryos, uncover parental genetic variants influencing their occurrence, and reveal important consequences for human fertility.

149

Expanding non-invasive prenatal testing to include microdeletions and segmental aneuploidy: cause for concern? *T. Sahoo, M.N Strecker, N. Dzidic, S. Commander, M.K Travis, C. Doherty, K. Hovanes.* Combimatrix, San Diego, CA.

INTRODUCTION: Evaluation of circulating cell-free fetal DNA by massively parallel shotgun or targeted sequencing has emerged as a powerful tool in screening for fetal aneuploidies. More recently, major providers of this technology have expanded their test offerings to include screening for common microdeletion syndromes. Recent literature suggests a cautious approach to the interpretation of non-invasive prenatal testing (NIPT) results based on higher-than-previously reported false positive rates when compared to invasive testing, as well as concerns regarding the potential for over-representation of the positive predictive value for specific aneuploidies.**METHODS**: We evaluated the outcome and concordance of invasive prenatal diagnostic testing (by fetal karyotype and or prenatal microarray) for 181 consecutive cases referred to our laboratory following NIPT.**RESULTS**: For all cases where both NIPT and invasive testing results were available (173/181), the overall true positive (TP) and false positive (FP) rates were estimated at 78% (135/173) and 22% (38/173), respectively. The chromosome-specific TP and FP rates, respectively, were 84% and 9% for Trisomy 21 (and with 7% partially concordant), 72% and 24% for trisomy 18 (with 4% partially concordant), 60% and 40% for trisomy 13, and 53% and 42% for sex chromosome aneuploidies (with 5% partially concordant). For cases that were predicted by NIPT to harbor specific microdeletions (N=16), the outcomes were particularly remarkable for high FP rates. There was a 60% FP rate for 22q11.2 deletion (DiGeorge syndrome), a 50% FP rate for 1p36 deletion, and an 80% FP rate for 5p deletion. For single cases with a deletion or duplication predicted by NIPT, the results of invasive testing were normal for a predicted 4p deletion and a 21q deletion but were confirmed for a NIPT predicted 9p duplication and 15q deletion.**CONCLUSIONS**: Overall TP and FP rates for the 16 cases with predicted microdeletion or duplication were 37.5% and 62.5%, respectively. These results suggest that claims regarding the reliability of NIPT for accurately predicting segmental imbalances may be premature and that such results should be evaluated with profound caution. It is imperative that both segmental aneuploidy cases and predicted whole chromosome aneuploidies continue to be followed-up with invasive prenatal diagnostic testing, and that clinical decisions not be undertaken purely based upon NIPT results.

150

NLRP2: a paternally-imprinted gene implicated in innate immunity and blastocyst development has a major gene effect on endometriosis. *K. Ward, R. Chettier, P. Farrington, H. Albertsen.* Juneau Biosciences, LLC, Salt Lake City, UT.

Endometriosis is a complex condition that affects 5-10% of women. Family studies have confirmed the heritability of endometriosis and GWAS studies have implicated several chromosomal regions, GWAS studies focus on SNPs typically with MAF >3%; but in contrast, exome sequencing (ES) can detect private and rare variants which may have larger functional effects. In this study, we sequenced the exomes of a multiplex family and unrelated individuals with endometriosis to detect variants likely to predispose women to develop endometriosis. METHODS: ES was performed on 8 related women with endometriosis using the Ion Proton platform (LifeScience Technologies). Pair-wise relatedness ranged from 1 meiosis to 5 meiosis apart. DNA variants were determined and confirmed using the GATK (Genome Analysis Toolkit). Prediction of protein function was evaluated using the Polyphen 2 database. We examined damaging rare variants (ExAc MAF >0 but <0.01, and polyphen2 >0.447). RESULTS: ES was accomplished with 100X mean coverage and detected 2246 rare, probably-damaging variants in 8 related subjects. We found a single gene (harboring two mutations in cis) shared by 8 out of the 9 affected women. The one affected woman who does not carry these variants probably has a disease phenocopy caused by a different mechanism (as frequently seen in pedigrees with common, complex disorders). The variants are located in a paternally imprinted gene NLRP2 (NACHT, LRR and PYD domains-containing protein 2). The variants were always transmitted by a mother in the index family. Two other research subjects with endometriosis share the same finding, they are third degree relatives, but they show no detectable (> 5MB) identity-by-descent to the index family. Subsequently, we tested for coding variants (MAF<0.005) in NLRP2 gene in an independent set of 268 unrelated patients with endometriosis. 9 variants were observed in 14 affected women (5.2%), a 5-fold increase of these variants compared to allele frequencies in the ExAc data (p= 1.6e-06, OR=5.2[3.0-9.0]) which is expected to include the population rate of affected women and male “carriers”. The NLRP2 gene is expressed in reproductive tissues, but there are no prior studies suggesting a role in endometriosis. CONCLUSION: We identified the first endometriosis predisposition gene showing a major gene effect. If confirmed, studying NLRP2 function and variation may increase our understanding of the pathogenesis of endometriosis.

151

Investigation of DNA variants responsible for Pre-eclampsia. *R. McGinnis[1], F. Dudbridge[2], D. Lawlor[3], J. Kemp[3], C. Franklin[1], N. Williams[1] On behalf of the InterPregGen Consortium.* 1) Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom; 3) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom.

Pre-eclampsia (PE) is a potentially fatal disorder characterized by hypertension and proteinuria after week 20 of pregnancy. Although no single-locus PE associations have been replicated or achieved genome-wide statistical significance in PE mothers or offspring ("fetal cases"), we evaluated evidence for shared DNA variants responsible for PE and other diseases by applying three methods to GWAS datasets. The polygenic score analysis (PGSA) and genome-wide complex trait analysis (GCTA) methods require individual genotype data which we sourced from Wellcome Trust Case-Control Consortium 1 (WTCCC1) for 6 disease phenotypes tested for correlation with our PE GWAS data. LD score analysis (LDSC) requires only summary GWAS results and thus could evaluate variant sharing between PE and more target diseases with publicly available summary data from much larger meta-analysis sample sizes. For PGSA, "discovery" GWAS SNPs below specific pvalue thresholds were identified in our PE meta-analyses of 3830 maternal or 2650 fetal PE cases and 47000 European controls and were tested in each WTCCC1 phenotype (~2000 UK cases, ~3000 UK controls). For GCTA, individual GWAS SNP genotypes from InterPregGen (1900 maternal or 1000 fetal UK cases, 5500 UK controls) were tested for correlation with each WTCCC1 phenotype. Both PGSA and GCTA found significant evidence ($p<0.001$) for shared DNA variants responsible for the WTCCC1 hypertension (HT) phenotype and maternal PE despite strict exclusion of pre-pregnancy HT from our maternal cases. No significant evidence of variant sharing was observed for maternal or fetal PE and other WTCCC1 phenotypes by PGSA or GCTA. For LDSC, comparison of our PE summary data with that of 11 other phenotypes found statistically significant variant sharing between maternal PE and gestational HT ($p<5\times10^{-7}$), coronary artery disease ($p<0.00015$), and type2 diabetes ($p<0.00016$). These initial results suggest overlap in women for risk of PE and other cardiometabolic diseases. We are refining our PGSA, GCTA and LDSC investigations and continuing to genotype large replication datasets to follow up statistical peaks from our PE GWAS meta-analyses. We will provide an update of both our variant sharing results and our efforts to identify genome-wide significant ($p<5\times10^{-8}$) single-locus association with maternal or fetal PE. Funding: EU FP7 grant 282540;Wellcome Trust (WT090355/A/09/Z, WT090355/B/09/Z,WT088806,WT087997MA 102215/2/13/2);MRC (102215/2/13/2,MC_UU_12013/5).

152

Ancestry Matched Genome-wide Association Study Identifies Variants Associated with Spontaneous Preterm Birth. *M. Sirota[1], J. Toung*[2], W. Sikora-Wohfeld[2], C.R. Gignoux[2], C.D. Bustamante[2], G. Shaw[2], H. O'Brodovich[2], D. Stevenson[2], A.J. Butte[1].* 1) Institute for Computational Health Sciences, UCSF, San Francisco, CA; 2) Stanford University, Stanford, CA.

Preterm birth (PTB), or the delivery of an infant prior to 37 weeks of gestation, is a significant determinant of infant morbidity and mortality. Globally 15 million babies are born preterm, every year with a frequency approximating 10%. Causes of preterm birth remain largely mysterious and thus not amenable to intervention to reduce the rate of this catastrophic event for children and their families. Although twin studies estimate that genetics account for approximately 27 to 36% of the variation in preterm birth, to date no causal mutations have been identified through unbiased genome-wide genetic screens. In this study, we performed a large genome-wide association study of over 1300 preterm birth babies and over 12,000 ancestry matched controls from the Health and Retirement Study. To account for differences in population substructure between our cases and controls, we used 1,815 individuals from the Phase 3 of the 1000 Genomes Project as an anchor point to match our cases and controls. We tested over 2 million SNPs for association to the spontaneous preterm birth phenotype with sex and the first ten principal components of genetic ancestry as covariates in our model across five populations including African (AFR), the Americas (AMR), European (EUR), South Asian (SAS) and East Asian (EAS). We identified 96 loci associated with spontaneous preterm birth at a genome-wide level of significance including 7 exonic hits in the AFR and EUR populations. The most significant exonic hit corresponded to rs1937839 (*P*=1.04E-16), a synonymous SNP in *AKR1C3*. *AKR1C3*, also known as Prostaglandin F Synthase (PGFS), and acts as a key steroidogenic enzyme that metabolizes various steroid hormones toward sex hormones. In particular, studies on the biochemical pathways involved in human labor has found *AKR1C3* expression to be significantly correlated with gestational age and tissue type. Furthermore, reactions catalyzed by AKR1C3 stimulate smooth muscle contraction during parturition. Other exonic loci found to be significantly associated to spontaneous preterm birth include rs10417778, a synonymous SNP in *PSG4* (*P*=9.98E-13) and rs33923703, a nonsynonymous SNP (Met-to-Val) in *AURKA* (*P*=1.22E-8). In this work, we were able to leverage publicly available data in order to identify a large control group as well as match our cases and controls based on their ancestry lineage in order to carry out population specific case-control analysis identifying variants associated with PTB.

153

Large scale genome-wide association study for birth weight identifies 13 novel loci and reveals genetic links with a variety of adult metabolic and anthropometric traits. *M. Horikoshi[1,2], F.R. Day[3], J.R.B. Perry[3], J-J. Hottenga[4,5], R. Li-Gao[6], M.N. Kooijman[7,8,9], R. Beaumont[10], N.M. Warrington[11,12], N.J. Timpson[13], Early Growth Genetics (EGG) Consortium.* 1) The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 2) Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK; 3) MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK; 4) Netherlands Twin Register, Department of Biological Psychology, VU University, Amsterdam, the Netherlands; 5) EMGO Institute for Health and Care Research, VU University and VU University Medical Center, Amsterdam, the Netherlands; 6) Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands; 7) The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, the Netherlands; 8) Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, the Netherlands; 9) Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, the Netherlands; 10) Institute of Biomedical and Clinical Science, University of Exeter Medical School, Royal Devon and Exeter Hospital, Exeter, UK; 11) The University of Queensland Diamantina Institute, The University of Queensland, Translational Research Institute, Brisbane, Australia; 12) School of Women's and Infants' Health, The University of Western Australia, Perth, Australia; 13) MRC Integrative Epidemiology Unit at the University of Bristol, University of Bristol, UK.

Lower birth weight (BW) is associated with adult metabolic diseases including type 2 diabetes (T2D) and cardiovascular disease, but the underlying causes are poorly understood. We previously reported 7 loci associated with BW, 5 of which were also associated with adult traits (T2D, height and blood pressure (BP)). Here we report expanded analyses with genome-wide association studies of 75,924 European (EUR) and 9,960 non-EUR singletons from 36 studies and imputation up to the phase 1 integrated 1000 Genomes Project reference panel. We aimed to (i) discover novel BW loci through increased sample size and improved coverage of variants with minor allele frequency (MAF)<5%, and (ii) explore the shared genetic overlap between BW and adult traits. We combined association statistics between sex-specific BW Z-scores and each of 10.3M SNPs across studies via fixed-effects meta-analysis. We detected 13 novel loci at genome-wide significance ($P<5x10^{-8}$): near *EPAS1* ($P=2.0x10^{-17}$), *ESR1* ($P=1.9x10^{-14}$), *YKT6* ($P=2.9x10^{-12}$), *CLDN7* ($P=5.9x10^{-12}$), *ZBTB7B* ($P=5.4x10^{-10}$), *HHEX/IDE* ($P=5.7x10^{-9}$), *SREBF2* ($P=8.1x10^{-9}$), *C20orf203* ($P=1.1x10^{-8}$), *FOXA2* ($P=1.4x10^{-8}$), *TBX20* ($P=2.3x10^{-8}$), *CCND1* ($P=3.3x10^{-8}$), and *WNT4-ZBTB40* ($P=3.8x10^{-8}$) from sex-combined analysis, and near *INTS7* ($P_{female}=3.6x10^{-8}$, $P_{male}=0.042$) from sex-specific analysis. The lead SNPs at all loci were common except at *YKT6* (MAF EUR: 1%, African-American: 0.2%). There was no evidence of heterogeneity in allelic effects between studies at lead SNPs at any of the loci (Cochran's Q $P>0.05$), highlighting common BW signals are shared across ancestry groups. Conditional analyses confirmed that multiple novel BW signals overlapped those associated with adult metabolic and anthropometric traits: *HHEX/IDE* in T2D, *FOXA2* in hyperglycemia, *ZBTB7B* in waist-to-hip ratio and *INTS7* in height. Using LD score regression, we estimated the genome-wide genetic correlation (R_G) between birth weight and publicly available GWAS of other traits and found strong correlations with height ($R_G=0.43$, $P=7x10^{-43}$), T2D ($R_G=-0.35$, $P=4x10^{-8}$), systolic BP ($R_G=-0.30$, $P=1x10^{-8}$) and coronary artery disease (CAD, $R_G=-0.32$, $P=2x10^{-8}$). In summary, we extended the number of BW associated loci from 7 to 20, and our analysis provided strong genome-wide genetic links between lower BW and higher risk of T2D, high BP and CAD in later life, which partially explain the complex relationships between genetic variation, early growth and adult metabolic disease.

154

Assessing Genetic Counsellors' Use of Shared Decision-Making. *P.H. Birch[1], A.V. Port-Thompson[1], S. Adam[1], F. Legaré[2], J.M. Friedman[1].* 1) Dept Med Genetics, Univ British Columbia, Vancouver, BC, Canada; 2) Family Medicine and Emergency Medicine Dept., Université Laval, QC, Canada.

Shared decision-making (SDM) is a clinical practice model whereby clinicians and patients consider the evidence favouring or opposing all options. They explore patient preferences together to reach a decision. Genetic counsellors (GCs) are generally not trained in this technique but it has been suggested that SDM is widely used in genetic counselling. Our goal was to measure the extent that SDM is practiced in prenatal counselling by GCs. 30 pregnant women at increased risk to have a child with a birth defect owing to the presence of "soft markers" on ultrasound examination, elevated maternal serum screening or family history, and who met with one of 10 experienced GCs, agreed to participate. Genetic counselling sessions lasted 49±16 (mean±SD) minutes and were audio-taped, transcribed verbatim, and scored independently by 3 researchers using OPTION, a validated tool widely used to measure SDM in clinical encounters [Elwyn G, et al., Health Expect 2005]. The OPTION tool scores clinical sessions on criteria such as whether the clinician draws attention to the identified problem, discusses available options and their pros and cons, and explores patient concerns, expectations, and understanding. Patients' decisional conflict and state anxiety were measured before and after the counselling session using self-administered validated tools. OPTION scores ranged from 28 to 70 (out of 100), with a mean score of 46 ± 9 (mean±SD). Two consistently low scoring areas were eliciting patients' preferred involvement in decision-making and assessing their preferred approach to receiving information. Mean pre-encounter decisional conflict scores were 42±10 (possible range is 16 to 80), and mean state anxiety scores were 44±13 (possible range is 20 to 80), both indicating that the patients were anxious and conflicted about their decision before the clinical encounter. Two weeks after the encounter, decisional conflict and anxiety scores had dropped to 26±9 and 35±15, respectively. This is the first study to assess SDM in genetic counselling sessions. OPTION scores were high compared to those observed among physicians in previous studies [Couet N, et al., Health Expect 2013], indicating that GCs often use SDM. Formal training of GCs in SDM could enhance their use of this method, especially in particular areas where its use was inconsistent, and might provide additional strategies for GCs to use in clinical encounters.

155

Mutations inDCHS1Cause Mitral Valve Prolapse. *D. Milan[1], R. Durst[3], K. Sauls[2], M. Talkowski[3], J.J. Schott[6], x. Jeunemaitre[4], A. Hagege[5], R.A. Levine[1], R.A. Norris[2], S. Slaugenhaupt[3].* 1) Cardiovascular Research Center, Massachusetts General Hospital, Charlestown, MA; 2) Department of Regenerative Medicine and Cell Biology, Medical University of South Carolina, Charleston SC; 3) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 4) INSERM, UMR-970, Paris Cardiovascular Research Center, 75015 Paris, France; 5) Hôpitaux de Paris, Cardiologie Département, Hôpital Européen Georges Pompidou, Paris, France; 6) Inserm U1087; institut du thorax; University Hospital, Nantes, France.

Mitral valve prolapse (MVP) is a common cardiac valve disease that affects nearly 1 in 40 individuals. It can manifest as mitral regurgitation and is the leading indication for mitral valve surgery. Despite a clear heritable component, the genetic etiology leading to non-syndromic MVP has remained elusive. Four affected individuals from a large multigenerational family segregating non-syndromic MVP underwent capture sequencing of the linked interval on chromosome 11. We report a missense mutation in the DCHS1 gene, the human homologue of the Drosophila cell polarity gene dachsous (ds) that segregates with MVP in the family. Morpholino knockdown of the zebrafish homolog dachsous1b resulted in a cardiac atrioventricular canal defect that could be rescued by wild-type human DCHS1, but not by DCHS1 mRNA with the familial mutation. Further genetic studies identified two additional families in which a second deleterious DCHS1 mutation segregates with MVP. Both DCHS1 mutations reduce protein stability as demonstrated in zebrafish, cultured cells, and, notably, in mitral valve interstitial cells (MVICs) obtained during mitral valve repair surgery of a proband. Dchs1 +/- mice had prolapse of thickened mitral leaflets, which could be traced back to developmental errors in valve morphogenesis. DCHS1deficiency in MVP patient MVICs as well as in Dchs1+/- mouse MVICs result in altered migration and cellular patterning, supporting these processes as etiological underpinnings for the disease. Understanding the role of DCHS1 in mitral valve development and MVP pathogenesis holds potential for therapeutic insights for this very common disease.

156

***ROBO-SLIT* Mutations Predispose Individuals to Bicuspid Aortic Valve with Ascending Aortic Aneurysm.** *R.A. Gould[1,2], H. Aziz[1,2], A. Kumar[3], C. Preuss[4], C. Woods[5], N. Sobreira[6], H. Ling[6], S.A. Mohamed[7], A. Franco-Cereceda[8], G. Andelfinger[4,10], A.S. McCallion[5], P. Eriksson[9], L.V. Laer[3], B.L. Loeys[3,11], H.C. Dietz[1,2,12], MIBAVA Foundation Leducq.* 1) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore , Maryland, USA; 2) Howard Hughes Medical Institute, Baltimore, Maryland, USA; 3) Center for Medical Genetics, Faculty of Medicine and Health Sciences, Antwerp University Hospital and University of Antwerp, Antwerp, Belgium; 4) Cardiovascular Genetics, Department of Pediatrics, Centre Hospitalier Universitaire Sainte-Justine Research Centre, Université de Montréal, Montreal, Quebec, Canada; 5) Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 6) Center for Inherited Disease Research, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 7) Department of Cardiac and Thoracic Vascular Surgery, University of Luebeck, Luebeck, Germany; 8) Cardiothoracic Surgery Unit, Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; 9) Atherosclerosis Research Unit, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, Stockholm, Sweden; 10) Department of Pediatrics, Université de Montréal, Montreal, Quebec, Canada; 11) Department of Human Genetics, Radboud University Medical Centre, Nijmegen, Netherlands; 12) Department of Pediatrics, Division of Pediatric Cardiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.

Bicuspid aortic valve (BAV) is the most common congenital heart defect, affecting up to 2% of people; a sizeable subset of individuals show predisposition for ascending aortic aneurysm (AscAA) and tear. BAV/AscAA segregates as an autosomal dominant trait with reduced penetrance and overt male predominance. The etiology of this condition is largely unknown, with mutations in *NOTCH1* accounting for less than 1% of cases. No strong signals have been detected using parametric linkage or GWAS, suggesting extreme locus heterogeneity and/or a complex genetic architecture. We performed whole exome sequencing (WES) on 181 patients, including a subset with familial segregation, to identify genes predisposing to BAV/AscAA. A heterozygous obligate splice site mutation in *ROBO4* (c.2056+1G>T) was seen in all 7 affected individuals in a multigenerational kindred, with complete penetrance in males (6/6) but not females (1/3). Functional analyses demonstrated a stable transcript lacking constitutively utilized exon 13, predicting the formation of a transmembrane protein that binds ligands (SLITs) but lacks signaling potential. We show that *ROBO4* is normally expressed in endocardial cushion and proximal aortic endothelium during relevant stages of development, and that targeted silencing results in enhanced endothelial cell migration with loss of tight junctions and barrier function. Patient aortic tissue shows deep infiltration of *ROBO4*-expressing cells into the aortic media with attendant upregulation of a-smooth muscle actin and collagen production, which strongly suggests pathogenic endothelial-to-mesenchymal transition. Overall, WES revealed significant enrichment for rare (MAF<0.01%) and predicted deleterious variation in the *ROBO-SLIT* axis, with particular enrichment in *ROBO2* (with 4 highly deleterious variants that were entirely novel), but also relevance in *ROBO3* (3 putative pathogenic variants) and *ROBO4* (2 variants). All genes encoding Slit ligands (1-3) showed rare and predicted deleterious variation including a novel mutation in *SLIT1* that segregated with disease in a small kindred and an exceedingly rare (MAF<0.005%) variant in *SLIT3* that was recurrent among unrelated BAV/AscAA probands. Taken together these data define a novel pathogenic mechanism for a common human disease phenotype with major public health burden; a new emphasis on endothelial cell biology may reveal unanticipated therapeutic strategies.

157

Genetic, developmental and paracrine interactions in the complex pathogenesis of heritable aneurysm conditions. *E. Gallo MacFarlane[1], J.P. Habashi[1, 2], Y. Chen[1], D. Bedja[3], H.C. Dietz[1, 4].* 1) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA; 2) Division of Pediatric Cardiology, Department of Pediatrics, Johns Hopkins University, Baltimore, MD, USA; 3) Department of Molecular and Comparative Pathobiology, Johns Hopkins University, Baltimore, MD, USA; 4) Howard Hughes Medical Institute, Bethesda, MD, USA.

Aortic aneurysm is a critical feature of Loeys-Dietz syndrome (LDS), a connective tissue disorders with many features in common with Marfan syndrome (MFS) and other TGFβ vasculopathies. LDS is caused by heterozygous *loss-of-function* mutations in genes encoding positive effectors of the TGFβ pathway (*TGFBR1*, *TGFBR2*, *SMAD3*, *TGFB2* and *TGFB3*), however, a signature for paradoxically increased TGFβ signaling is found in aortic tissue from both patients and mouse models, as assessed by phosphorylation of Smad2 and expression of TGFβ target genes. This has engendered ambiguity regarding the precise role of TGFβ in aneurysm progression and the wisdom of treatments based on TGFβ antagonism. The spatial distribution of aneurysms in LDS and MFS is distinctly nonrandom, with uniform involvement of aortic segments where vascular smooth muscle cells (VSMCs) of different embryonic origin interface. For example, the aortic root shows interspersion of VSMCs derived from the second heart field (SHF) and cardiac neural crest (CNC). We hypothesized that signaling overdrive in LDS might be caused by TGFβ signaling imbalance and consequent paracrine interactions that are established postnatally. Using lineage tracing and FACS, we observed that wild-type (WT) SHF- and CNC-VSMCs have equivalent signaling potential in the acute-phase response to TGFβ. This signaling response is also fully preserved in CNC-VSMCs harboring a heterozygous LDS-associated missense mutation in *Tgfbr1*, but entirely abrogated in corresponding LDS SHF-VSMCs. Loss of suppressive TGFβ signaling in LDS SHF-VSMCs leads to upregulation of the angiotensin II (AngII) type 1 (AT1) receptor, unique sensitization to AngII-mediated extracellular signal-regulated kinase (ERK) activation (not seen in LDS CNC-VSMCs or in WT VSMC of either lineage), and ultimately increased expression of TGFβ1 ligand that can overdrive neighboring LDS CNC-VSMCs, which retain full TGFβ signaling potential, in a paracrine manner. In keeping with this model, we find clear overlap between the CNC lineage mark and gain of TGFβ signaling in LDS mice using *in situ*-based assays. These data reconcile the low signaling/high signaling paradox, the spatial distribution of aneurysms and prevention of aneurysm in LDS mice treated with AT1 receptor blockers. They also suggest that potent global or CNC-specific TGFβ antagonism has therapeutic promise in TGFβ vasculopathies, a concept under evaluation using pharmacologic and genetic provocations.

158

Discovery of novel genes underlying congenital heart defects driven by analysis of 4,593 exomes from affected families. *A. Sifrim[1], M-P. Hitz[1], S. Al Turki[1,3], A. Wilsdon[2], J. McRae[1], T. Singh[1], B. Thienpont[5], J. Breckpot[5], K. Setchfield[2], F. Bu'Lock[13], A.K. Manickaraj[4], A.V. Postma[6], S. Omer[3], J. Bentham[11], S. Bhattacharya[12], C. Cosgrove[12], H. Watkins[12], H. Abdul-Khaliq[15], H-H. Kramer[17], O. Tokan[16], U. Bauer[18], P. Daubeney[14], R. Abu-Sulaiman[3], K. Devriendt[5], S. Mital[4], B. Keavney[7], J. Goodship[8], S. Klaassen[9,10], D. Brook[2], M.E. Hurles[1], UK10K Consortium, Deciphering Developmental Disorders Study.* 1) Wellcome Trust Sanger Institute, Cambridge, Cambridgeshire, United Kingdom; 2) School of Life Sciences, University of Nottingham, Nottingham, United Kingdom; 3) Department of Cardiac Sciences , King Abdulaziz Cardiac Center, Riyadh, Saudi Arabia; 4) Division of Cardiology, Department of Pediatrics, Hospital for Sick Children, University of Toronto, Toronto, Canada; 5) Centre for Human Genetics, Katholieke Universiteit Leuven, Leuven, Belgium; 6) Heart Center, Academic Medical Center, Amsterdam, the Netherlands; 7) Institute of Cardiovascular Sciences, University of Manchester, Manchester, UK; 8) Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK; 9) Experimental and Clinical Research Center (ECRC), Charite' Medical Faculty and Max-Delbruck-Center for Molecular Medicine, Berlin, Germany; 10) Competence Network for Congenital Heart Defects, Germany; 11) Department of Cardiovascular Medicine and Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 12) Radcliffe Department of Medicine and Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 13) East Midlands Congenital Heart Centre, University Hospitals of Leicester NHS Trust, Leicester, United Kingdom; 14) Division of Paediatric Cardiology, Royal Brompton Hospital; Reader in Paediatric Cardiology at Imperial College, London, United Kingdom; 15) Department of Paediatric Cardiology, Saarland University, Homburg, Germany; 16) Department of Pediatric Cardiology, Children's Hospital, Friedrich Alexander University Erlangen, Erlangen, Germany; 17) Department of Congenital Heart Disease and Pediatric Cardiology, UKSH Kiel, Kiel, Germany; 18) Deutsches Herzzentrum Berlin, Competence Network for Congenital Heart Defects, Berlin, Germany.

Congenital heart defects (CHDs) are the most common form of birth defect, occurring in 0.8% of live births. However, the majority of CHDs remains of unknown genetic etiology, leading to poor clinical diagnostic yields. By exome sequencing 1827 CHD probands (1383 parent-child trios), covering a broad spectrum of syndromic/non-syndromic cardiac abnormalities, we present the largest CHD rare variant sequencing study to date. Using an unbiased genotype-driven approach across a large cohort of probands we are able to confirm known and putative CHD genes, as well as discover 13 novel candidate genes (with an estimated FDR=0.05, genes currently under validation at the time of abstract submission). By combining de novo mutation analysis, recessive inheritance analysis, rare inherited variant burden testing (using 11225 inhouse control exomes), a CNV meta-analysis and a network-based pathway analysis we are able to further delineate the genetic etiology of CHDs. Additionally, we are able to stratify our cohort into syndromic and non-syndromic subcohorts in order to understand differences in their respective underlying genetic architectures. By applying and integrating this broad scope of analyses on a large, and morphologically wide, cohort of CHDs we hope to provide insights into the complexities of the disease and pave the way for the next generation of CHD studies.

159

A rare missense variant in the B-type natriuretic peptide gene *NPPB* is associated with increased risk for incident heart failure. *P. Salo[1,2], A. Havulinna[3], P. Jousilahti[3], V. Salomaa[3], M. Perola[1,2,4].* 1) Genomics and Biomarkers Unit, Natl Inst Health & Welfare, Helsinki, Finland; 2) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland; 3) Department of Health, Natl Inst Health & Welfare, Helsinki, Finland; 4) Estonian Genome Center of the University of Tartu, Estonia.

Myocardiocytes release atrial and B-type natriuretic peptides (ANP & BNP) into circulation when the heart muscle stretches excessively. Coded by *NPPA* and *NPPB* on 1p36, ANP and BNP are produced as prohormones (proANP & proBNP) which are cleaved into N-terminal parts and the active hormones. They regulate cardiovascular homeostasis, mainly by increasing natrium excretion via kidneys lowering blood volume and pressure. N-terminal parts of proANP and proBNP (MR-pro-ANP and NT-pro-BNP) as well as the hormones themselves are assayed for diagnostics, while recombinant BNP and inhibitors of neutral endopeptidase, the main enzyme metabolizing AND and BNP, are being developed to treat heart failure.We performed a genome-wide association study of plasma BNP, MR-pro-ANP and NT-pro-BNP concentrations in 5,800 Finns to identify genetic loci related to natriuretic peptide biology and, by extension, to heart failure. We identified a novel locus at 12q21 as associatied with NT-pro-BNP (rs11105298, P=4 x 10^{-11}) near *ATPB2* and refined previously reported cis-associations of variants near *NPPA* and *NPPB*. Furthermore, we identified the rare allele of the Arg72His missense variant rs61761991 within the N-terminal part of BNP to be associated with reduced measured NT-pro-BNP concentration (P=2 x 10^{-80}). As plasma natriuretic peptides are quantified using immunoassays, this association may be caused by a reduced affinity of the antibody for proBNP for carriers of the rare allele, possibly resulting in missed diagnosis of heart failure for carriers with misleadingly low measured plasma NT-pro-BNP. We thus tested the association of rs61761991 with incident heart failure in a cohort of 14,541 Finns with 570 cases of incident heart failure. Surprisingly, we found the rare allele to be associated with increased risk for incident heart failure (rs61761991 T allele, hazard ratio=1.48, 95% CI =1.10-1.99, P= 0.00947).The frequency of the T allele of rs61761991 is 2.90% in Finns but it is extremely rare or absent in other European populations. Previous reports show that compared to other Europeans, Finns are enriched in deleterious rare and low-frequency variants, particularly in the 1% to 5% frequency range, supporting the potential pathogenicity of rs61761991. As the N-terminal fragment of BNP is currently thought to be mostly inactive, the identification of a deleterious variant within it is unexpected and may advance our understanding of a central cardiovascular regulatory system.

160

Meta-analysis of exome chip variants identifies rare and common variants associated with electrocardiographic left ventricular mass. *G. Kosova[1,2], N. Verweij[3], P. van der Harst[3], C. Newton-Cheh[1,2] on behalf of the CHARGE Consortium EX-EKG working group.* 1) Center for Human Genetic Research and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT; 3) University of Groningen, University Medical Center Groningen, Department of Cardiology.

Increased electrocardiographic left ventricular mass (eLVM) is a characteristic of and risk factor for heart failure, as well as mortality in the general population. Prior genetic studies of eLVM focused primarily on common variants, most of which are non-coding. The role of low-frequency (MAF 0.01-0.05) and rare (MAF <0.01) coding variants underlying eLVM remain unexplored. In order to better understand the genetic determinants of eLVM, we tested the association between 246,516 variants on the Infinium HumanExome BeadChip, mostly coding and low-frequency/rare, with 3 EKG measures of eLVM: products of QRS duration and a) 12-lead sum (12LS), b) Sokolow-Lyon (S-L), and c) Cornell (C) voltage. 48,611 individuals of European (n=43,041), African American (n=2,097), Hispanic (n=1,322) and East Asian (n=725) ancestries were genotyped in 11 participating cohorts (AGES, CHS, CROATIA, GS, INTER99, KORA, LIFELINES, MESA, SHIP, WHI and YFS) of the CHARGE EX-EKG consortium. Single variant association tests in each cohort were performed using RAREMETALWORKER, and results were combined using fixed effects meta-analysis in RAREMETAL, using $P<1x10^{-8}$ as a genome-wide significant threshold. We identified variants at 18 independent loci (10 missense, 4 intronic and 4 intergenic) significant for at least one eLVM phenotype, representing 9 novel and 9 previously reported associations. We then evaluated the effects of rare/low-frequency coding variants in aggregate in gene-based tests using SKAT, identifying 4 genes at exome-wide significance ($P<1x10^{-6}$): *TTN* (S-L $P=3.2x10^{-11}$), *DSP* (12LS $P=3.3x10^{-10}$), *KLHL38* (12LS $P=2.4x10^{-7}$) and *NPR1* (12LS $P=7.5x10^{-7}$). *NPR1*, *KLHL38* and *DSP* associations were each explained by a single missense low-frequency variant. The *TTN* SKAT association lost significance ($P=0.99$) after adjusting for the two most significant low-frequency variants, although a common variant remained significant ($P=5.4x10^{-11}$). *TTN* encodes titin, a protein that influences elasticity of cardiomyocytes, and *DSP* encodes desmoplakin, an essential component of intercellular junctions in cardiomyocytes; both are known to harbor rare coding mutations as causes of Mendelian cardiomyopathy. *NPR1* encodes the receptor for atrial and B-type natriuretic peptides; activation of *NPR1* has vasodilatory, natriuretic and anti-fibrotic effects. Taken together, we have identified 9 novel loci harboring common variants and 4 genes harboring rare/low-frequency coding variants associated with eLVM.

161

Investigating the effects of coding variants on QT and JT intervals utilizing data from 95,626 individuals. *N.A. Bihlmeyer[1], J.A. Brody[3], D.E. Arking[2], N. Sotoodehnia[3], CHARGE Consortium EX EKG Working Group.* 1) Predoctoral Training Program in Human Genetics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 2) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, BRB Room 447, 733 N. Broadway St, Baltimore 21205, MD, USA; 3) Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA.

QT interval, measured through a standard electrocardiogram, captures the time it takes for the ventricles in the heart to depolarize and repolarize. JT interval, a measure of ventricular repolarization alone, can be mathematically derived by subtracting the QRS interval from the QT interval. Prolonged QT interval has been linked to higher risk of sudden cardiac death.We performed anexome-wide analysis for both QT and JT intervals, including both common and rare variants from the Illumina ExomeChip using single variant and gene-based statistical models. The gene-based model includes missense, nonsense, splice variants, and indels in a SKAT association test.We perform a meta-analysis of 241,552 variants and 17,574 genes in a sample of 95,626 individuals from 23 cohorts (comprised of 83,884 European ancestry, 9,610 African American, 1,382 Hispanic, and 750 Asian individuals) and identified 10 loci that modulate QT interval and/or JT interval that have not been previously reported in the literature. The novel loci are marked by the genes *PM20D1* [HGNC: 26518], *SLC4A3* [MIM: 106195], *CASR* [MIM: 601199], *SENP2* [MIM: 608261], *SLC12A7* [MIM: 604879], *ZNF37A* [MIM: 616085], *NRAP* [MIM: 602873], *NACA* [MIM: 601234], *KLF12* [MIM: 607531], *GOSR2* [MIM: 604027]. Within these loci, 13 novel single nucleotide variants were discovered, 8 of which are nonsynonymous, 1 splicing, 1 3'-UTR, 2 intronic, and 1 intergenic. Among the SKAT gene-based model results, 6 genes were implicated: *GPR153* [MIM: 614269], *RNF207* [HGNC: 32947], *TEPP* [MIM: 610264], *HVCN1* [MIM: 611227], *KCNH2* [MIM: 152427], and *SLC4A3*. Ongoing analyses include determining whether genetic variants have differential effects on QT and JT intervals. We are also using conditional analyses to determine if previously reported QT interval associated loci from common variant GWAS are explained by nearby coding variants.

162

HUNTing for Susceptibility Genes for Myocardial Infarction with Whole Genome Sequencing. *C. Willer[1,2,3], O. Holmen[4,5], H. Zhang[1], E. Schmidt[1,2], W. Zhou[1,2], J. Chen[1], G. Abecasis[6], M. Boehnke[6], R. Mills[2], H.M. Kang[6], K. Hveem[4,5,7].* 1) Department of Internal Medicine, Division of Cardiology, University of Michigan Medical School, Ann Arbor, Michigan, 48109, United States of America; 2) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, 48109, United States of America; 3) Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan, 48109, United States of America; 4) HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, 7600 Levanger, Norway; 5) St. Olav Hospital, Trondheim University Hospital, Trondheim, Norway; 6) Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, 48109, United States of America; 7) Department of Medicine, Levanger Hospital, Nord-Trøndelag Health Trust, 7600 Levanger, Norway.

We systematically assessed genome-wide variation to identify new genes influencing myocardial infarction, fine map known disease loci and evaluate whether low-frequency variants with large effects exist for this trait. Using a combination of whole genome sequencing of 2,202 individuals (4.5x average depth) and genotyping of ~50,000 individuals, we examined 18.7 million variants, including 112,487 coding variants in participants in the Nord-Trøndelag Health Study (The HUNT Study).An early analysis in our study identified one variant in *TM6SF2* (encoding p.Glu167Lys), residing in a known genome-wide association study locus for lipid traits, that influences total cholesterol levels and is associated with myocardial infarction. Transient *TM6SF2* overexpression or knockdown of *Tm6sf2* in mice altered serum lipid profiles, consistent with the association observed in humans, identifying *TM6SF2* as a functional gene within a locus previously known as *NCAN-CILP2-PBX4* or 19p13. Other preliminary analyses suggests several small deletions that newly show association with MI, including one near the COL4A1 GWAS locus. Whole genome sequence information is now available for 2,202 individuals and follow-up genotyping and imputation is ongoing for ~50,000 individuals. Here we report further detailed analyses of MI and lipid susceptibility loci, including testing of structural variants, enrichment with regulatory regions and prioritization of potential functional variants, gene-based burden tests for coding variants and regulatory-based burden tests for non-coding variants.

163

In Silico Predictive Modelling of CRISPR/Cas9 guide efficiency. *J. Listgarten[1], I. Smith[2], M. Hegde[2], J. Doench[2], N. Fusi[1].* 1) Microsoft Research New England, Microsoft, CAMBRIDGE, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA.

The CRISPR/Cas9 system provides state-of-the art genome editing capabilities. However, several facets of this system are under investigation for further characterization and optimization. One in particular is the choice of guide RNA that directs Cas9 to target DNA--given that one would like to target the protein-coding region of a gene, hundreds of guides satisfy the constraints of the CRISPR/Cas9 Protospacer Adjacent Motif sequence. However, only some of these guides efficiently target DNA to generate gene knockouts. One could laboriously and systematically enumerate all possible guides for all possible genes and thereby derive a dictionary of efficient guides, however, such a process would be costly, time-consuming, and ultimately not practically feasible. Instead, one can (1) enumerate all possible guides over each of some smaller set of genes, and then test these experimentally by measuring the knockout capabilities of each guide, (2) thereby assemble a training data set with which one can "learn", by way of predictive machine learning models, which guides tend to perform well and which do not, (3) use this learned model to generalize the guide efficiency for genes not in the training data set. In particular, by deriving a large set of possible predictive features consisting of both guide and gene characteristics, one can elicit those characteristics that define guide-gene pairs in an abstract manner, enabling generalizing beyond those specific guides and genes, and in particular, for genes which we have never attempted to knock out and therefore have no experimental evidence. Based on such a set of experiments, we present a state-of-the art predictive approach to modeling which RNA guides will effectively perform a gene knockout by way of the CRISPR/Cas9 system. We demonstrate which features are critical for prediction (e.g., nucleotide identity), which are helpful (e.g., thermodynamics), and which are redundant (e.g., microhomology). Finally, we combine our insights of useful features with exploration of different model classes, settling on one model which performs best (gradient-boosted regression trees). Finally, we elucidate which measures should be used for evaluating these models in such a context. A prediction server implementing our final approach is available as a cloud service.

164

A novel algorithm for estimating shared haplotype segments using rare genetic variants. *P. Albers[1], M. McCarthy[1,2,3], G. McVean[1,4].* 1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, Oxfordshire, United Kingdom; 2) Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK; 3) Oxford NIHR Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK; 4) Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom.

Whole genome sequencing revealed that the majority of genetic variants resides in the lower end of the frequency spectrum. Rare alleles, which are shared by few individuals in a sample, are typically population-specific and young in age, having arisen a few generations ago through mutation. A shared rare allele can be indicative that it has been inherited from a common ancestor, along with the surrounding haplotype region. As recombination had less time to break down LD, rare alleles are expected to sit in longer haplotype tracts that are identical by descent (IBD). Detection of such segments is not directly possible, as recombination leaves no detectable breakpoints. Here, we use rare alleles to tag IBD tracts and find regions of extended haplotype sharing (EHS) that can be identical by descent or state, enclosing the IBD region surrounding the focal allele. EHS is delimited by genotypes distal to the focal site that exhibit opposite homozygotes, where a genotype pair is unlikely to share a haplotype, providing a first indication of recombination. We do this for alleles that are shared by $n \geq 2$ individuals in unphased genotype data. First, we use this information to recover the sharing structure in individuals carrying a rare allele, which is informative for recent ancestry and fine-scale population structure. Using 1000 Genomes data (phase 3), we compared alleles at different frequencies; e.g. 0.1% and 0.5%, where median EHS length was 273kb (0.41cM) and 184kb (0.28cM) respectively, and median fraction of individuals sharing rare alleles outside their own continental population was 16% and 25% respectively, but showing larger variations across populations. Second, by considering EHS in all pairs of individuals sharing a focal allele and sharing patterns at nearby loci, we attempt to distinguish chromosomes on which rare alleles reside. This phasing method is based on Hidden Markov Models, where overall state space is reduced to EHS recovered per individual. At frequencies $\leq 0.5\%$ we found that EHS regions cover ~99% of the genome on average across samples. Further, IBD breakpoints are unlikely to be inferred from distances to a focal site or EHS bounds alone, reflecting the random nature of recombination. We attempt to approximate breakpoints after phasing, using identity distributions of shared and unshared haplotypes across EHS overlaps. Algorithms exploiting rare allele sharing may be powerful across a range of applications in statistical genetics.

165

PADRE: Pedigree Aware Distant Relationship Estimation. *J.E. Below[1], J. Staples[2,3], D.J. Whitherspoon[4], L.B. Jorde[4], U.W. C.M.G.[2], D.A. Nickerson[2], C.D. Huff[5].* 1) Human Genetics Center, University of Texas Health Science Center, Houston, TX; 2) Regeneron, Terrytown, NY; 3) Dept of Genome Sciences, University of Washington, Seattle, WA; 4) Dept of Human Genetics, University of Utah, Salt Lake City, Utah; 5) Dept of Epidemiology, MD Anderson Cancer Center, Houston, TX.

Although accurate knowledge of shared ancestry has always been important in genetic studies, in the current frontier of rare variation, genetically reconstructing large pedigrees connected through distant, and possibly cryptic, relatedness is key to designing appropriate and well-powered analytic strategies. Current relationship prediction programs such as Estimation of Recent Shared Ancestry (ERSA) utilize the length and distribution of shared genomic segments to estimate pairwise relationships as distant as ninth-degree relatives. On the other hand, pedigree reconstruction algorithms, such as PRIMUS (Pedigree Reconstruction and Identification of a Maximum Unrelated Set), use estimates of genome-wide IBD proportions to reconstruct actual pedigrees from genetic data. We demonstrate that combining these distant relationship estimates with pedigree information can substantially improve relationship estimation power and accuracy. We have developed a Pedigree Aware Distant Relationship Estimation (PADRE) algorithm that uses relationship likelihoods generated by ERSA to identify the maximum likelihood connection between family networks reconstructed by PRIMUS. To demonstrate the power of PADRE to estimate relationships from SNP genotypes, we performed pedigree simulations and we analyzed real data from 169 individuals within three previously described extended pedigrees. PADRE results in substantial improvement of both the accuracy and power to detect distant relationships compared to ERSA alone- correctly predicting on average 20% more of the fourth- through ninth-degree relationships within one degree. PADRE also correctly predicts 59% of the tenth- through thirteenth-degree simulated relationships within one degree of relatedness, compared to 4% with ERSA alone. The improvement seen in the real pedigrees is consistent with the improvement seen with simulations. We also used PADRE to estimate relationships among the HapMap3 CEU samples, illustrating distant sixth- to eighth-degree relationships connecting established pedigrees. PADRE, which is publicly available, greatly expands the range of relationships that can be estimated using genetic data in pedigrees.

166

Haplotype phasing using cluster graphs. *D.C. Aguiar[1], L.T. Elliott[2], Y.W. Teh[2], B.E. Engelhardt[1].* 1) Computer Science and Center for Statistics and Machine Learning, Princeton University, Princeton, NJ; 2) Statistics, University of Oxford, Oxford, UK.

Haplotype-based analyses, including genotype imputation, inferring demographic history, and multilocus association mapping, suffer from two problems that reduce their efficacy in genomics and medical research. First, haplotypes are often not known precisely: experimental methods to characterize haplotypes are prohibitively expensive, whereas computational approaches either use complex models that are intractable for large samples, or use oversimplified models for computational efficiency. Second, many haplotype-based analyses use the particular haplotype sequences as a proxy for the latent genealogy (e.g., demographic history and association mapping); these analyses would greatly benefit from a representation that captures the rich coalescent ancestry of the sample in a computationally tractable and statistically robust framework. Here, we present a Bayesian nonparametric statistical model and efficient inference algorithms that improves on the Li & Stephens PAC model by adding exchangeability [Li & Stephens 2003] with the computational efficiency of HMM-based models such as SHAPEIT2 [Delaneau *et al.* 2012]. Our method, *phaseME*, improves on existing methods by (a) estimating the haplotype clustering with an expressive model that, at each locus, clusters the haplotypes as a coalescent process; (b) using reference haplotypes with an exchangeable model so the distribution across haplotypes does not depend on haplotype order; (c) inferring a biologically meaningful latent structure, called a *haplotype cluster graph*, that captures the evolutionary relationships between haplotypes; and (d) exploiting the compressed latent structure of a large collection of haplotypes to make haplotype phasing and imputation of unphased genotypes tractable. In particular, phaseME captures complex haplotype structure among reference haplotype sequences by inferring the underlying haplotype cluster graph. Then, new genotypes can be efficiently phased and the full set of reference loci imputed by threading paths through the reference haplotype cluster graph. We apply phaseME to compute complete genome-wide haplotype cluster graphs for the 1000 Genomes Project data, and use the cluster graphs to phase and impute whole genome data. We compare the results from phaseME with leading haplotype phasing and imputation software based on precision and computational resources, and we illustrate the benefits of model expressibility in this tractable framework.

167

Fine-mapping cellular trait QTLs with RASQUAL and ATAC-seq. *N. Kumasaka, A. Knights, D. Gaffney.* Wellcome Trust Sanger Institute, Cambridge, United Kingdom.

When cellular traits are quantified using high-throughput DNA-sequencing quantitative trait loci (QTLs) can be detected using two orthogonal sources of information: population level differences between individuals and allelic differences between cis-haplotypes within individuals. We present RASQUAL (Robust Allele Specific Quantitation and quality controL) that combines both these sources information in an efficient framework for association mapping. RASQUAL substantially improves association detection and causal variant localisation over existing methods across a broad range of next generation sequencing data sets, including RNA-seq, DNaseI-seq and ChIP-seq. In addition, RASQUAL explicitly models a range of technical biases in allele-specific data and can be applied to existing data sets without requiring computationally expensive realignment or generation of personal reference genomes, making it substantially quicker and easier to run than existing approaches. We illustrate how RASQUAL can be used to maximise association detection in small samples by generating the first map of chromatin ccessibility QTLs (caQTLs) in a European population using a recently developed assay for transposase accessible chromatin (ATAC-seq). Despite a modest sample size, RASQUAL identified 1,798 independent caQTLs at FDR10%, uncovering regulatory variants that alter chromatin structure over very long distances, revealing an intruiging potential link between chromatin accessibility and replication timing, and providing powerful information for localising putatively causal disease-associated genetic variants. Our results illustrate the power of joint modelling of population and allele-specific signatures for functional interpretation of noncoding variation in health and disease.

168

A phenotype-aware approach to the prioritization of coding and non-coding rare disease variants. *D. Smedley[1], J. Jacobsen[1], C. Mungall[2], N. Washington[2], S. Kohler[3], S.E. Lewis[2], M. Haendel[4], P.N. Robinson[3].* 1) Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; 3) Institute for Medical and Human Genetics, Charite-Universitatsmedizin, Berlin, Germany; 4) University Library and Department of Medical Informatics and Epidemiology, Oregon Health & Sciences University, Portland, OR, USA. Whole-exome sequencing has revolutionized rare disease research. However, many cases remain unsolved, in part because of the problem of prioritizing the ~100-1000 loss of function, candidate variants that remain after removing those deemed as common or non-pathogenic. We have developed the Exomiser suite of algorithms to tackle this through additional use of patient phenotype data e.g. hiPHIVE assesses each candidate gene by comparing the patient phenotypes to existing phenotypic knowledge from disease and model organism databases. For genes with missing data, a guilt-by-association approach is used based on the phenotypic similarity of near-by genes in a protein-protein association network. Benchmarking on known HGMD mutations added to unaffected 1000 Genomes Project exomes, reveals the causative variant is detected as the top hit in 97% of samples and in 87% when phenotypic knowledge of the known disease-gene association was masked to simulate novel disease gene discovery. Exomiser is being successfully applied to diagnosis and gene discovery in the Undiagnosed Disease Program.However, many cases still remain unsolved after exome analysis and a significant fraction may be due to the presence of non-coding, causative variants. With many projects such as the UK 100,000 Genomes Project adopting whole genome sequencing, the potential to detect such variants exists if scalable and accurate analysis tools are developed. We have extended hiPHIVE to also assess variants in proximal and distal non-coding regions including tissue-specific enhancers. The predicted deleteriousness of these variants is assessed through combining measures of conservation with indicators of regulatory regions such as DNase I hypersensitivity and transcription factor binding sites. Through manual curation of the literature, we have developed a database of genotype and phenotype associations for several hundred regulatory mutations known to cause rare diseases. hiPHIVE was able to detect these regulatory mutations as the top hit in 64% of samples when they were added to whole genomes from the 1000 Genomes Project. By variant category the performance was 55% for promoters, 79% for 5'UTR, 47% for 3'UTR, 100% for microRNA genes and 57% for enhancers. Finally, this extension to non-coding variation paves the way for development of phenotype-aware analysis software for common disease.

169

Subclonal hierarchy inference from somatic mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *N. Niknafs[1,2], V. Beleva-Guthrie[1,2], DQ. Naiman[3], R. Karchin[1,2,4].* 1) Biomedical Engineering Department, Johns Hopkins University, Baltimore, MD; 2) Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD; 3) Applied Math and Statistics Department, Johns Hopkins University, Baltimore, MD; 4) Oncology Department, Johns Hopkins University, Baltimore, MD.

Recent improvements in next-generation sequencing of tumor samples and the ability to identify somatic mutations at low allelic fractions have opened the way for new approaches to model the evolution of individual cancers. The power and utility of these models is increased when tumor samples from multiple sites are sequenced. Temporal ordering of the samples may provide insight into the etiology of both primary and metastatic lesions and rationalizations for tumor recurrence and therapeutic failures. Additional insights may be provided by temporal ordering of evolving subclones- cellular subpopulations with unique mutational profiles. We present a new modular framework based on a rigorous statistical hypothesis test to infer the SubClonal Hierarchy from Somatic Mutations (SCHISM). Our framework decouples the problems of mutation cellularity estimation and temporal ordering, and can thus be flexibly combined with existing tools addressing either of these problems. The SCHISM framework includes tools to interpret hypothesis test results, which inform phylogenetic tree construction, and we introduce the first genetic algorithm designed for this purpose. The utility of our framework is demonstrated in simulations and by application to data from three published multi-region tumor sequencing studies of (murine) small cell lung cancer, acute myeloid leukemia, and chronic lymphocytic leukemia. Using a number of different configurations of tools in SCHISM framework, we were able to identify subclonal phylogenies that were either identical to or inclusive of the phylogenies reconstructed by study authors using manual expert curation. SCHISM can be applied in patients with multiple characterized tumor samples, to study the genomic underpinnings of tumor evolution, and possibly therapeutic resistance.

170

A powerful new method based on mutual information to simultaneously test for additive, dominance and interaction effects. *A.I. Young[1], F. Wauthier[1,2], P. Donnelly[1,2].* 1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, Oxfordshire, United Kingdom; 2) Department of Statistics, University of Oxford, 1 South Parks Road, Oxford.

There is a major open question as to how important dominance and interaction (non-additive) effects are in the genetic architecture of human diseases and traits. The controversy remains unsolved in part through lack of powerful methods for detecting these effects and in part through the lack of suitably sized datasets. The imminent availability of large population based studies, including biobanks, will for the first time offer the sample size necessary to address this question properly. The mutual information between a phenotype and genetic variant is a completely general measure of their dependence, which captures all of the phenotypic information carried by the genetic variant. We develop a new test statistic based on the mutual information which is proportional to the sum of the log likelihood ratio test statistics for additive, dominance and interaction effects. We also introduce a new visualisation tool, a generalization of the traditional Manhattan Plot, which stacks the evidence for additive, dominance and interaction effects respectively at each SNP. In particular, it highlights genomic regions with significant evidence for involvement in interactions. We show theoretically that, under certain conditions, our test is the most powerful for detecting a locus with both additive and non-additive effects. Our model incorporates a random effect to adjust for complex population structure and polygenic additive effects, and we provide a novel algorithm to fit the model which scales linearly with sample size, making it possible to analyse biobank scale datasets. Simulations on real genetic data show that it has increased power over existing methods to detect loci involved in interactions. To assess the extent of genetic interactions in human traits, we apply our method to search for gene-gene and gene-environment interactions in the UK Biobank, which has rich phenotype data on ~150,000 individuals genotyped at ~800,000 SNPs.

171

A new ciliopathy protein complex directing assembly of the IFT machinery is implicated in OFD syndrome and other ciliopathies. *C. Thauvin-Robinet[1,2], AL. Bruel[2], M. Toriyama[3], C. Lee[3], S.P. Taylor[4], I. Duran[4], D.H. Cohn[5], J.M. Tabler[3], K. Drew[3], M.R. Kelley[6], S. Kim[3], T.J. Park[3], D. Braun[7], G. Pierquin[8], A. Biver[9], K. Wagner[10], A. Malfroot[11], I. Panigrahi[12], H.A. Al-Lami[13], Y. Yeung[13], Y.J. Choi[14], L. Faivre[1,2], JB. Rivière[2,15], J. Chen[14], K.J. Liu[13], E.M. Marcotte[3], F. Hildebrandt[7,16], D. Krakow[5], P.K. Jackson[6], J.B. Wallingford[3,16].* 1) Clinical genetics centre and Eastern referral centre for developmental anomalies and malformative syndromes , FHU-TRANSLAD, Dijon, France; 2) EA4271GAD Genetics of Developmental Anomalies, FHU-TRANSLAD, Medecine Faculty, Burgundy University, F-21079, Dijon, France; 3) Department of Molecular Biosciences, Center for Systems and Synthetic Biology, and Institute for Cellular and Molecular Biology, University of Texas at Austin, Texas, USA; 4) Departments of Orthopaedic Surgery, Human Genetics and Obstetrics and Gynecology, David Geffen School of Medicine at UCLA, Los Angeles, California, USA; 5) Department of Molecular Cell and Developmental Biology, University of California at Los Angeles, California, 90095, USA; 6) Stanford University School of Medicine, Baxter Laboratory, Department of Microbiology & Immunology, Stanford, California, 94305, USA; 7) Department of Medicine, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA; 8) Clinical genetics centre, University Hospital Center, Luik, Belgian; 9) Pediatric unit, Hospital Center, Luxemburg, Luxemburg; 10) Cardiological Pediatric unit, Hospital Center, Luxemburg, Luxemburg; 11) Clinic of Pediatric Respiratory Diseases, Infectious Diseases, Travel Clinic and Cystic Fibrosis Clinic at the Universitair, Ziekenhuis UZ Brussel, Belgium; 12) Department of Pediatrics Advanced, Pediatric Centre Pigmer, Chandigarh, India; 13) Dept. of Craniofacial and Stem Cell Biology, Dental Institute, King's College London; 14) Departments of Pathology and Dermatology, Stony Brook University, Stony Brook, New-York 11794, USA; 15) Laboratory of Molecular Genetics, FHU-TRANSLAD, PTB, CHU Dijon, F-21079 Dijon, France; 16) Howard Hughes Medical Institute, Chevy Chase, Maryland, USA.

Oral-facial-digital syndromes (OFD) belong to ciliopathies, a broad class of human diseases which share an etiology of defective cilia structure or function. To date, 9 causal ciliary genes have been identified, encoding for proteins implicated in the centriole elongation, the basal body and the transition zone. Cilia is indeed an organelle assembled and maintained by multi-protein machines, such as the BBSome, a multi-protein complex involved in trafficking of ciliary membrane proteins, the Nphp/Mks and B9 complexes assembling the ciliary transition zone which controls access to the cilium and dynein arms that drive ciliary beating. The intraflagellar transport (IFT) system, which links cargos to microtubule motors for transport along ciliary axonemes, is also comprised of two multi-protein complexes, IFT-A and IFT-B. In two OFD patients with cardiac defects, exome sequencing identified a homozygous frameshift mutation (p.Asn132Lysfs*11) in the novel *INTU* gene and compound heterozygoty for frameshift (Leu176Ilefs*21) and missense (p.Asp54Asn), predicted to alter splicing by Human Splice Finder, mutations in the *WDPCP* gene. In addition, in the *INTU* gene, we also identified a well-conserved homozygous missense mutation (p.Ala452Thr) in a child with nephronophthisis and growth retardation, as well as compound heterozygosity for nonsense (p.Glu355*) and missense (p.Glu500Ala) mutations in an affected case with Short-Rib Polydactyly syndrome. Combining also proteomics, *in vivo* cell biology, and mouse genetics studies, we demonstrated that INTU and WDPCP proteins belong to a same protein complex, named CPLANE (for ciliogenesis and planar polarity effectors), which also include the C5orf42 protein, encoded by the gene that we previously identified as majorly responsible for OFD VI syndrome. CPLANE proteins specifically interact with IFT-A machinery, acting at the base of cilia to facilitate association of peripheral IFT-A proteins with the IFT-A core. In the absence of CPLANE function, dissociation of peripheral IFT-A proteins disrupts retrograde transport of IFT-B, but leaves bi-directional traffic of the IFT-A core unaffected. Finally, examination of mutant mice and ciliopathy alleles from human patients reinforce the connection between CPLANE proteins and the IFT-A machinery, demonstrating that CPLANE, growing the list of ciliogenic multi-protein complexes, plays a broad and essential role in ciliogenesis and human ciliopathies.

172

Differential regulation of glucose homeostasis and β-cell mass in Bardet-Biedl Syndrome and Alstrom Syndrome. *S. Lodh, T. L. Hostelley, C. C. Leitch, E. A. O'Hare, N. A. Zaghloul.* School of Medicine, EDN, University of Maryland, Baltimore, MD.

Bardet-Biedl syndrome (BBS) and Alstrom syndrome are rare obesity syndromes caused by dysfunction in genes associated with primary cilia. Though these two ciliopathies are similar in the highly penetrant rates of obesity, they differ significantly in the prevalence of diabetes. BBS cohorts exhibit a lower prevalence of diabetes, while Alstrom patients are more prone to develop diabetes. This variability suggests the possibility of discrepant mediation of glucose regulation by ciliopathy genes, perhaps independent of obesity. One possible mechanism might be differential effects on **β-**cell mass. To address this question, we investigated the production and development of pancreatic β-cells in zebrafish models of BBS and Alstrom. We examined the production of β-cells *in vivo* in *insulin*:mCherry transgenic zebrafish embryos depleted for *bbs1, bbs4* or *alms1.* Upon analysis of the area of β-cell mass and numbers of β-cells at 2 and 5 days post fertilization (dpf), we observed opposing effects between BBS genes and the Alstrom gene. Production of β-cells was significantly reduced with depletion of *alms1* expression at both stages, but in contrast, loss of *bbs1* and *bbs4* resulted in significantly increased β-cell mass throughout development. These effects appeared to be specific to β-cells as other differentiated cell types remained unchanged by 5 dpf in either model. Moreover, reduced *bbs1* and *bbs4* expression impeded the expression of pancreatic progenitor markers, suggesting the possibility of early skewing of endocrine populations in *bbs* morphants specifically towards a β-cell fate. To examine a potential role for these genes in β-cell adaptive capacity, we also assessed expansion in response to high glucose conditions or regeneration in response to β-cell ablation. While *alms1* morphants' β-cell mass did not expand or regenerate as a result of increased apoptosis and significantly reduced proliferation, *bbs1* morphants were able to maintain increased β-cell mass in the presence of glucose and were able to regenerate after ablation similar to controls. Taken together, our findings suggest differential roles for BBS and Alstrom genes in maintaining the β-cell population and offer a potential mechanism for the discrepant prevalence of diabetes across these two obesity ciliopathies.

173

Loss-of-function mutations in *IFIH1* predispose to severe viral respiratory infections in children. *S. Asgari[1], LJ. Schlapbach[2], S. Anchisi[3], C. Hammer[1], I. Bartha[1], PJ. McLaren[1], T. Junier[1], D. Garcin[3], J. Fellay[1].* 1) School of Life Science, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, VD, Switzerland; 2) Paediatric Critical Care Research Group (PCCRG), Mater Research, University of Queensland, Brisbane, Australia; 3) Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, Geneva, Switzerland.

Respiratory viruses are the most common pathogens leading to non-elective admission to Pediatric Intensive Care Unit (PICU). Dramatic inter-individual differences are observed in the severity of these viral infections. This project aims at identifying and functionally characterizing rare genetic variants conferring unusual susceptibility to common viral respiratory infections in the pediatric population.Previously healthy children requiring intensive care support because of a severe viral respiratory infection were prospectively recruited in Swiss and Australian PICU. After exome sequencing, we used a combination of bioinformatic tools for variant calling and annotation and only included in downstream analyses single nucleotide variants (SNVs) and small insertions/deletions (indels) meeting stringent quality criteria. We searched for an enrichment of rare loss-of-function (LoF) variants in the study population. Functional validation was carried out to confirm the relevance of potential causal variants.A total of 242,954 variants, including 1,878 putative LoF variants were identified in 120 study participants. After filtering, 7 rare LoF variants were found at a higher frequency in our study population than in the Exome Aggregation Consortium database (ExAC) and in an in-house set of 485 exomes, and in homozygous form in at least one participant. Among them was a rare splice-site disrupting variant in the Interferon Induced With Helicase C Domain 1 (IFIH1) gene, which encodes a RIG-I-like cytoplasmic sensor of long double-stranded RNA and plays a role in innate immune response against RNA viruses. This variant was found in four of the patients (1 homozygous, 3 heterozygous). RNA sequencing showed that the variant results in exon skipping and premature stop, leading to the deletion of the IFIH1 C-regulatory domain, which is required for viral RNA recognition. In vitro expression of the alternatively spliced transcript showed that the resulting mutant protein is unable to induce interferon-beta in presence or absence or RNA ligand, that it has no ATPase activity and that it is less stable than wild-type IFIH1. We also observed a dominant negative effect in co-transfection experiments.This study demonstrates that a LoF variant in IFIH1 results in primary immunodeficiency against RNA viruses, both in homozygous and in heterozygous form, suggesting a central role for IFIH1 in the establishment of an efficient response against common viral respiratory infections.

174

A novel (epi)genotype-specific and histotype-targeted tumor surveillance protocol in Beckwith-Wiedemann Syndrome based on cancer data meta-analysis. *A. Mussa[1], M. Gregnanin[1], C. Molinatto[1], G. Baldassarre[1], S. Russo[2], A. Riccio[3], G.B. Ferrero[1].* 1) Department of Public Health and Pediatric Sciences, University of Torino, Torino, Italy; 2) Laboratory of Cytogenetics and Molecular Genetics, Istituto Auxologico Italiano, Milan, Italy; 3) DiSTABiF, Second University of Naples and Institute of Genetics and Biophysics “A. Buzzati-Traverso” - CNR, Naples, Italy.

Objective: Beckwith-Wiedemann syndrome (BWS), the most common overgrowth cancer predisposition disorder, entails a 8% risk of cancer development. BWS implies the need for specific tumor surveillance, based on abdominal ulstrasound (US) for Wilms’ tumor diagnosis and serum a-fetoprotein (aFP) measurement for hepatoblastoma detection. The objective of this study was to compare tumor risk and histotypes in the 4 main BWS molecular subgroups: IC1-GoM (Imprinting Center 1 hypermethylation), IC2-LoM (Imprinting Center 2 hypomethylation), UPD (chromosome 11p15 paternal uniparental disomy), *CDKN1C* mutations. **Methods:** Cancer data from published BWS cohorts (Mussa, 2015, Ibrahim, 2014, Brioude, 2013, Cooper, 2005, Bliek, 2004, Weksberg, 2001, Gaston, 2001, Romanelli, 2010) were meta-analyzed. **Results:** 1,009 IC2-LoM, 412 UPD, 157 IC1-GoM, 76 *CDKN1C* mutated patients were pooled: 128 developed malignancies (7.7%). Cancer cases were 23 (2.3%) in IC2-LoM, 61 (14.9%) in UPD, 39 (24.8%) in IC1-GoM and 5 (6.6.%) in *CDKN1C* patients ($p<0.001$). Wilms’ tumor was almost exclusively observed in IC1-GoM and UPD cases ($p<0.001$), representing 95% and 50 % of malignancies in these subgroups. Only 1 case of Wilms’ tumor was observed in IC2-LoM patients. Hepatoblastoma was associated with UPD ($p<0.001$) in 4.4% of UPD cases *vs* 0.6% in the other subgroups. Adrenal carcinoma was found only in 1.7% of UPD cases ($p<0.001$). Neuroblastoma was associated with *CDKN1C* mutations ($p=0.002$) representing 80% of *CDKN1C* tumors. **Conclusions:** Data on >1,600 BWS cases allow a revision of tumor surveillance protocol, allowing to propose differentiated screenings based on specific (epi)genotype and targeted to specific histotypes. Clearly, IC1-GoM and UPD patient benefit from renal US, due to the high risk of Wilms’ tumor. aFP monitoring is worthwhile in UPD patients, given the high hepatoblastoma risk. The novel association between UPD and adrenal carcinoma implies the possibility of a specific screening based on adrenal US, clinical and hormonal evaluation for adrenal hyperfunction. A 5% prevalence of neuroblastomas in *CDKN1C* mutation patients may justify screening based on urinary catecholamines, associated with abdominal US. Based on these data, both US and aFP appear questionable in IC2-LoM cases, given the low tumor risk and the variegate histotype spectrum hampering the application of screening strategies: likely clinical follow-up and *ad-hoc* tests may be a reasonable alternative for these patients.

175

A recurrent mosaic mutation of *SMO* (Smoothened) causes Curry-Jones syndrome. *S.R.F Twigg[1], Y. Zhou[1], K.A. Miller[1], S.J. McGowan[1], J. Taylor[2], J. Craft[2], J.C. Taylor[2], A.F. Brady[3], J. Clayton-Smith[4], C.L. Clericuzio[5], C.J. Curry[6], W.B. Dobyns[7], D.K. Grange[8], D. Horn[9], R.B. Hufnagel[10], M.C. Jones[11], I.K. Temple[12], A.O.M. Wilkie[1,13].* 1) Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK; 2) Wellcome Trust Centre for Human Genetics, Roosevelt Drive, University of Oxford, Oxford, UK; 3) North West Thames Regional Genetics Service, Kennedy-Galton Centre, Northwick Park Hospital, Harrow, UK; 4) Manchester Centre for Genomic Medicine, St Mary's Hospital, University of Manchester, Manchester, UK; 5) Division of Genetics/Dysmorphology, Department of Pediatrics, University of New Mexico, Albuquerque, New Mexico, USA; 6) Genetic Medicine, UCFS Fresno, California, USA; 7) Center for Integrative Brain Research, Seattle Children's Hospital, Seattle, USA; 8) Department of Pediatrics, Division of Genetics and Genomic Medicine, Washington University School of Medicine, St. Louis, MO, USA; 9) Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, Berlin, Germany; 10) Division of Human Genetics, Department of Pediatrics, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, Ohio, USA; 11) Department of Pediatrics, University of California, San Diego, California, USA; 12) Human Genetics and Genomic Medicine, Faculty of Medicine, University of Southampton, UK; 13) Craniofacial Unit, Oxford University Hospitals NHS Trust, Oxford, UK.

Curry-Jones syndrome (CJS [MIM 601707]) is characterized by unicoronal craniosynostosis, corpus callosum agenesis, polysyndactyly, and abnormalities of skin, gut and eyes. Medulloblastoma has been described in a single case. The combination of asymmetry of clinical features, patchy skin manifestations and limb anomalies suggested that this could be a mosaic condition, possibly involving sonic hedgehog signaling. We used exome sequencing to analyze DNA obtained from blood or skin from 4 CJS individuals. In a single sample from affected skin, an apparently heterozygous nonsynonymous variant (44/83 reads) in *SMO* (c.1234C>T; p.Leu412Phe), encoding Smoothened, a G protein-coupled receptor that transduces hedgehog signaling, was identified. Comparison with exome data from eyelid skin DNA of the same patient, showed the variant at much lower frequency (5%; 2/40 reads). Further scrutiny of the exome sequence revealed the identical variant in a second case at 4% (6/167 reads). Following systematic collection of further samples (including from affected archival paraffin-embedded tissues) and analysis by deep sequencing, we found that 6/8 cases are mosaic for the identical p.Leu412Phe substitution, with varying amounts of the mutant allele (0-57%) in different tissues, demonstrating that this mutation, arising post-zygotically, accounts for most cases of CJS.Somatic mutations of *SMO* that result in constitutive activation have been described in several tumours, including medulloblastoma, ameloblastoma and basal cell carcinoma. Strikingly, the most common *SMO* mutation leads to p.Leu412Phe (COSMIC); this substitution has been shown to activate Smoothened in the absence of hedgehog signaling, providing an explanation for the development of neoplasms in CJS. Our discovery of the causative mutation in CJS raises therapeutic possibilities using recently generated Smoothened inhibitors.Review of MRI scans in CJS reveals a variety of previously undescribed brain abnormalities, including dilated ventricles and polymicrogyria, emphasizing that mosaic *SMO* mutations have developmental effects as well as cancer-related ones. We are currently investigating samples obtained from patients with relevant isolated skin and brain malformations to determine whether the clinical spectrum of *SMO* mutations can be expanded.In summary, our work uncovers the major genetic cause of CJS and illustrates strategies for gene discovery in the context of low-level tissue-specific somatic mosaicism.

176

The Genetics of Emphysema: Mutations in telomere genes uncover a distinct genetic etiology and common mechanism for pathogenesis. *S.E. Stanley[1,2], C.D. Applegate[1,3], M. Armanios[1,3].* 1) Oncology, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Medical Scientist Training Program, Johns Hopkins University School of Medicine, Baltimore, MD; 3) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD.

Pulmonary emphysema, a type of chronic obstructive pulmonary disease (COPD), is a leading cause of mortality worldwide. Severe forms of emphysema cluster in families where alpha-1 antitrypsin deficiency has been excluded, suggesting that yet-unexplained genetic causes play a role. We recently identified heterozygous germline mutations in *TERT*, the telomerase reverse transcriptase, in two cohorts of smokers with severe COPD. The frequency of *TERT* mutation carriers rivaled that of alpha-1 antitrypsin deficiency caused by homozygous mutations in *SERPINA1.* Here we report three additional telomere genes, *RTEL1, TR,* and *DKC1*, in a new series of cases with severe emphysema. A distinct set of comorbidities clustered in this group, including osteoporosis, myelodysplastic syndrome, and varying degrees of interstitial lung disease. All of the assessable cases reported a family history of pulmonary fibrosis that manifested in a predictable pattern: emphysema only developed in smokers, while never-smokers developed pulmonary fibrosis. The co-occurrence of emphysema and fibrosis within families suggests a common etiology, and exposes a unique gene-environment interaction between smoking and lung disease phenotype in telomere gene mutation carriers. Recognizing telomere-mediated emphysema cases has critical implications for patient care, especially in the lung transplant setting. Because short telomeres cause stem cell senescence in the lung, these genetic findings point to stem cell aging as a contributor to emphysema pathogenesis in a disease subset.

177

High Frequency of VACTERL Association in Fanconi Anemia. *B.P. Alter, S.A. Savage, N. Giri.* Clinical Genetics Branch, DCEG, National Cancer Institute, Rockville, MD.

Objective: To determine the frequency of VACTERL association in patients with Fanconi anemia (FA). VACTERL stands for anomalies of vertebrae, anal atresia, congenital heart disease, trachea-esophageal fistula, esophageal (or duodenal) atresia, renal, and limb anomalies. Anomalies in at least three of these categories are required to be classified as VACTERL. The presence of a VACTERL phenotype among cases with FA is considered to be about 5%; the frequency of FA among patients with a VACTERL phenotype is unknown. **Methods:** We examined 54 patients with FA among those in the National Cancer Institute Inherited Bone Marrow Failure Syndrome Cohort for features of VACTERL. Our assessment included imaging studies (radiology and ultrasound) when applicable. We correlated the results with genotypes when known. **Results:** Eighteen of the 54 (33%) patients had 3 or more VACTERL features. These were in 6/30 with *FANCA* (MIM 607139), 4/9 with *FANCC* (MIM 613899), 0/1 with *FANCD1/BRCA2* (MIM 605724/600185), 2/3 with *FANCD2* (MIM 613984), 0/2 with *FANCF* (MIM 613897), 3/3 with *FANCI* (MIM 611360), 2/2 with *FANCJ* (MIM 609054), and 1 of 4 with gene unknown. Vertebral anomalies were frequently identified by radiography, renal structural anomalies by ultrasound, and cardiac by symptoms in infancy. Only limb (thumb and/or radius) birth defects were clinically obvious. Thus the apparent presence of VACTERL association in FA is much higher than the estimated 5% ($p<0.0001$). Four of the 18 (22%) with VACTERL features (4/54 total, or 7%, similar to the expected 5%) had been classified as VACTERL prior to the diagnosis of FA. There was no association of the presence or absence of VACTERL with development of cancer, stem cell transplant, or survival. **Discussion:** A much larger proportion of patients with FA than predicted had features consistent with the VACTERL association, comprising one-third of those in the NCI cohort. Imaging studies were important for vertebral and renal anomalies, and explain the high frequency in our cohort. Identification of any components of the VACTERL association should lead to imaging studies, and to consideration of the diagnosis of FA, particularly if the patient has both radial ray and renal anomalies. .

178

***De novo* deletions and truncating mutations in *USP9X* cause a recognizable ID syndrome with multiple congenital abnormalities in females.** *M.R.F. Reijnders[1,10], V. Zachariadis[2,10], B. Latour[1,10], G.M. Mancini[3], C.M.A. van Ravenswaaij-Arts[4], H.E. Veenstra[4], B.M. Anderlid[2], S. Wood[5], A.S. Brooks[3], H. Malmgren[2], M. Vreeburg[6], V.R. Sutton[7], Z. Stark[8], J. Gecz[9], L. Jolly[9], C. Gilissen[1], R. Pfundt[1], T. Kleefstra[1], R. Roepman[1,11], A. Nordgren[2,11], H.G. Brunner[1,11].* 1) Department of Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands; 2) Department of Molecular Medicine and Surgery and Centre for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden; 3) Department of Clinical Genetics, Erasmus MC, Sophia Children's Hospital, Rotterdam, The Netherlands; 4) University of Groningen, University Medical Center of Groningen, Department of Genetics, Groningen, The Netherlands; 5) The Eskitis Institute for Drug Discovery, Griffith University, Brisbane, Australia; 6) Department of Clinical Genetics, Maastricht University Medical Center, Maastricht, The Netherlands; 7) Department of molecular and human genetics, Baylor College of medicine, Houston, Texas, USA; 8) Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Melbourne, Australia; 9) Neurogenetics, School of Paediatrics and Reproductive Health and the Robinson Research Institute, the University of Adelaide, Adelaide, Australia; 10) These first authors contributed equally; 11) These senior authors contributed equally.

Over a hundred genes have been reported to cause recessive X-linked intellectual disability (XLID). In contrast, the number of identified X-linked genes in which *de novo* mutations cause ID in females is still limited. Interpretation of the X-linked variants is complicated due to variable X-inactivation, which tends to lead to differences in phenotype between males and females. An example is *USP9X*, in which missense mutations were reported to cause recessive XLID and epilepsy in males, while female carriers were unaffected. We report here nine females with *de novo* deletions and truncating mutations in *USP9X* with a distinct phenotype including ID, short stature and multiple congenital abnormalities, comprising choanal atresia, hearing loss, hypomastia, heart abnormalities, post-axial polydactyly, hip dysplasia, anal abnormalities and structural brain abnormalities. Thus far, no truncating mutations had been observed in male patients, indicating that they might be lethal in males. *USP9X* encodes a highly conserved deubiquitinating enzyme that has been implicated in a variety of different biological processes such as embryogenesis, stem cell- and neural development and oncogenesis. The features observed in the female patients overlap those of known ciliopathy syndromes, which raised the possibility that USP9X might have a ciliary function. Further evidence comes from interrogation of an affinity proteomics dataset of ciliary protein modules that yielded USP9X as the most abundant deubiquitinating enzyme. We initiated functional studies in primary skin fibroblasts of the affected females with *USP9X* mutations, to assess whether USP9X mutations disrupted ciliary architecture and intraflagellar transport. We visualized the structure of the cilium by immunofluorescence and found that endogenous USP9X localizes to the ciliary axoneme. We next assessed mTORC1 activity using phophorylation levels of S6 ribosomal protein, one of the known interactors of USP9X, as a readout . We observed aberrant mTORC1 activation under starvation conditions in the affected females´ fibroblasts when compared to age and gender matched controls. This study defines a novel X-linked syndrome with distinctive and clinically recognizable features that is caused by *de novo* deletions and truncating mutations in *USP9X* in affected females, which may be attributable to disrupted ciliary function.

179

Genome-wide association study of olanzapine pharmacokinetics. *K.L. Bigos[1,2], R.M. Haynes[1], D. Chen[1], D.R. Weinberger[1,2].* 1) Lieber Institute for Brain Development, Baltimore, MD; 2) Johns Hopkins School of Medicine, Baltimore, MD.

The response to antipsychotics is highly variable, in part due to the wide variability in the pharmacokinetics (PK) of these drugs. In the CATIE schizophrenia trial, 64% of patients discontinued treatment with olanzapine due to lack of efficacy and/or side effects. This study aimed to identify genetic predictors of olanzapine clearance and determine whether they predict response to olanzapine. Patients with schizophrenia were treated with olanzapine as part of the CATIE trial. Plasma samples and dosing information were collected during study visits, and olanzapine plasma concentrations were measured using HPLC. The olanzapine pharmacokinetic model was developed using nonlinear mixed-effects modeling techniques. Drug clearance was modeled with each single nucleotide polymorphism (SNP) serially using an additive genetic model using the publically available CATIE genome-wide association (GWA) data. The olanzapine PK GWA was completed for a Caucasian cohort (CAUC, 157 patients, 560 plasma concentrations) and an African American cohort (AA, 73 patients, 261 plasma concentrations), separately. The top SNP associated with olanzapine clearance in the CAUC cohort is in the gene *CSMD1* (rs17413343, linear regression p=6.35e-16, r^2=0.22, minor allele frequency (MAF)=0.06). The top SNP associated with olanzapine clearance in the AA cohort is 20kb upstream of *TMCO4* (rs16822923, linear regression p=5.79e-20, r^2=0.69, MAF=0.18). Other top SNPs included families of enzymes and transporters known to be involved in pharmacokinetics including *CYP4F12* in the CAUC cohort (rs7253210, p=3.0e-7, r^2=0.16, MAF=0.067) and *SLC10A7* in the AA cohort (3'UTR SNP rs1057560, p=5.6e-16, r^2=0.60, MAF=0.096). Olanzapine clearance is associated with symptoms of schizophrenia (PANSS total score in CATIE phase 1); patients with higher clearance (lower plasma concentrations) have more symptoms (linear regression, CAUC cohort: n=101, p=0.0053, r^2=0.081; AA cohort: n=46, p=0.032, r^2=0.10). *CSMD1* fast-metabolizing genotype is associated with more symptoms of schizophrenia (linear regression p=0.026, r^2=0.049) in the CAUC cohort. The fast metabolizing genotype of rs16822923 is associated with more positive symptoms of schizophrenia (linear regression p=0.026, r^2=0.11) in the AA cohort. SNPs and polygenetic scores associated with olanzapine clearance and clinical response in this study may be useful in the future to optimize the selection and dosing of olanzapine in patients with schizophrenia.

180

Concurrent direct human leukocyte antigen (HLA) genotyping and genome-wide association study (GWAS) reveal major genetic determinants of anti-thyroid drug-induced agranulocytosis. *P. Chen[1,2,3,4], S. Shih[2,5], P. Wang[6], C. Fann[7], W. Yang[2,3,4,5], T. Chang[2,5].* 1) Department of Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan, Taiwan; 2) Division of Endocrinology and Metabolism, Department of Internal Medicine, National Taiwan University Hospital, Taipei 100, Taiwan; 3) Graduate Institute of Medical Genomics and Proteomics, College of Medicine, National Taiwan University, Taipei 100, Taiwan; 4) Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei 100, Taiwan; 5) Department of Medicine, College of Medicine, National Taiwan University, Taipei 100, Taiwan; 6) Department of Internal Medicine, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung 833, Taiwan; 7) Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan.

Graves' Disease (GD, MIM 27500) is the leading cause of hyperthyroidism affecting 1.0-1.6% of the population. Anti-thyroid drugs (ATDs, including methimazole, carbimazole and propylthiouracil) are relatively simple molecules known as thionamides, which have been cornerstones of GD treatment across the globe. A**T**Ds **i**nduced-**a**granulocytosis, namely **t**hionamide **i**nduced **a**granulocytosis, (**TiA**, defined as an absolute granulocyte count < 500 per cubic millimeter while taking ATDs), is the most feared adverse effect of ATDs and can occur in 0.1-0.37% of GD patients receiving these medications. Genetic predisposition of TiA was previously unknown. Here we conducted a two-stage association study on two separate subject sets (in total 42 agranulocytosis cases and 1,208 Graves' disease controls; all ethnic Chinese Han in Taiwan), using two independent methodologies. The first method was direct human leukocyte antigen (HLA) genotyping covering six classical HLA loci (*HLA-A*, *-B*, *-C*, *-DPB1*, *-DQB1* and *-DRB1*). The other method was SNP-based genome-wide association study. We demonstrated *HLA-B*38:02* (Armitage trend $P_{combined} = 6.75 \times 10^{-32}$) and *HLA-DRB1*08:03* ($P_{combined} = 1.83 \times 10^{-9}$) as independent susceptibility loci. The genome-wide association study identified the same signals. Estimated odds ratios for these two loci comparing effective allele carriers to non-carriers were 21.48 (95% confidence interval = 11.13-41.48) and 6.13 (95% confidence interval = 3.28-11.46), respectively. Carrying both *HLA-B*38:02* and *HLA-DRB1*08:03* increased odds ratio to 48.41 ($P_{combined} = 3.32 \times 10^{-21}$, 95% confidence interval = 21.66-108.22). We performed three-dimensional structure modeling to propose possible binding mechanism. These two HLA alleles are not uncommon in Asians but are rare in Caucasians, which implies that there might be additional HLA alleles responsible for TiA in different populations. It is intriguing that both class I and class II HLA loci can contribute to the same idiosyncratic drug adverse effect. Our results could be useful for anti-thyroid-induced agranulocytosis and potentially for agranulocytosis caused by other chemicals.

181

EuDAC: Genome-wide association study of drug-induced agranulocytosis in Europe. *M. Wadelius[1], N. Eriksson[2], L. Ibañez[3], E. Bondon-Guitton[4], R. Kreutz[5], A. Carvajal[6], M. Lucena[7], E. Sancho Ponce[8], J. Martin[9], T. Axelsson[10], Q-Y. Yue[11], P.K. Magnusson[12], P. Hallberg[1] on the behalf of EuDAC.* 1) Department of Medical Sciences, Clinical Pharmacology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden; 2) Uppsala Clinical Research Center and Department of Medical Sciences, Uppsala University, Uppsala, Sweden; 3) Fundació Institut Català de Farmacologia, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain; 4) Service de Pharmacologie Médicale et Clinique, Centre Hospitalier Universitaire, Faculté de Médecine de l'Université de Toulouse, Toulouse, France; 5) Charité - University Medicine, Institute of Clinical Pharmacology and Toxicology, Berlin, Germany; 6) Centro de Estudios sobre la Seguridad de los Medicamentos, Universidad de Valladolid, Valladolid, Spain; 7) Farmacologia Clinica, IBIMA, H Universitario Virgen de la Victoria, Universidad de Málaga, Málaga, Spain; 8) Capio Hospital General de Cataluña HGC, Sant Cugat del Vallès, Spain; 9) Instituto de Parasitologìa y Biomedicina López Neyra Avda, Armilla, Granada, Spain; 10) Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden; 11) Medical Products Agency, Uppsala, Sweden; 12) Swedish Twin Registry, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

Background: Agranulocytosis is a rare condition that can be caused by a variety of drugs. Due to the severity of the reaction, it would be valuable to be able to predict patients at risk. This is the aim of the European Drug-induced Agranulocytosis Consortium (EuDAC) that is led by the Swedish adverse drug reaction biobank Swedegene (www.swedegene.se). **Method**: A genome-wide association study (GWAS) was performed on 234 cases of drug-induced agranulocytosis from Sweden, Spain, Germany and France and 5170 population controls. Cases and controls were genotyped using different Illumina arrays: HumanOmniExpress-12v1, HumanOmni1-Quad 1M and HumanOmniExpress 700K. After quality control, the merged genotyped set contained 596,010 single nucleotide polymorphisms (SNPs). Principal component analysis (PCA) showed that the cases (96 males and 138 females) could be appropriately matched with the controls. After phasing and imputation with SHAPEIT and IMPUTE2, the dataset contained 9,380,034 SNPs (ref 1000 Genomes). SNP2HLA was used to impute HLA amino acids and classical HLA alleles on chromosome 6. We set the genome-wide significance p-value threshold to $p < 8.39$ e-08 to correct for multiple testing. **Results:** Most of the cases were associated with antithyroid agents, sulfasalazine, antibiotics and non-steroidal anti-inflammatory drugs (NSAIDs). The analysis demonstrated genome-wide significant associations with SNPs that tag the HLA-B and HLA-C loci. When agranulocytosis was treated as a single phenotype, the strongest signals came from HLA-C*02:02 (OR [95% CI] = 2.56 [1.87, 3.52], p = 4.99 e-10) and HLA-B*27:05 (odds ratio (OR) [95% confidence interval (CI)] = 2.61 [1.94, 3.52], p = 3.29 e-10). Further analysis showed that the effect was mainly driven by agranulocytosis caused by antithyroid agents. To avoid confounding by indication, the association was tested using Swedish cases with antithyroid agents (n=25) and Swedish controls matched for hyperthyroidism (n=74), which increased the odds ratio for HLA-C*02:02 to 8.33 [2.33, 29.78], p = 1.12 e-03.The p-value did not reach genome-wide significance, probably due to the small sample size. **Conclusion:** We found an association between drug-induced agranulocytosis and an HLA-B – HLA-C haplotype. There was no sign of confounding by indication when using matched controls. We are proceeding with analyses stratified by drugs, and are planning to replicate the findings. Collaborators with replication cohorts are invited.

182

The impact of genetics on drug efficacy and implications for future research. *M.R. Nelson[1], T. Johnson[2], L. Warren[3], A.R. Hughes[3], S.L. Chissoe[4], C-F. Xu[2], D.M. Waterworth[1].* 1) Target Sciences, GSK, King of Prussia, PA; 2) Target Sciences, GSK, Stevenage, UK; 3) Translational Sciences, PAREXEL International, RTP, NC (work performed while employed by GSK); 4) Target Sciences, GSK, Seattle, WA.

Lack of efficacy is the most common cause of attrition in late phase drug development. Many have proposed that germline genetics could be widely used to drive stratified drug development by identifying and enrolling patients most likely to respond. We critically reviewed the genome–wide association study (GWAS) and candidate gene literature to identify the drugs with convincing evidence for genetic influence on efficacy. Of 63 drug efficacy GWAS covering 36 unique drugs or drug classes, we found that 11 studies (17%) corresponding to 5 drug classes (14%) identified one or more variants that affect drug efficacy. Combining these results with several other well validated examples from candidate gene studies, we have identified eleven drugs or drug classes for which there are 19 robust genetic associations predictive for drug efficacy. The genes and pharmacogenetic mechanisms underlying the associations are mostly well understood and can be divided into three main categories: genes involved in drug exposure (37%), genes encoding the drug target (21%), and disease-related genes (42%). Based on the observed discovery rates and effect sizes, we estimate that approximate 9% (95% confidence interval of 2.7–18.6%) of drugs may have a germline genetic influence on their efficacy that is large enough to potentially influence treatment options. Because we cannot predict which drugs will have their efficacy influenced by clinically useful germline variants, we argue for early, routine, and cumulative screening for genetic efficacy predictors, as an exploratory part of clinical trial analysis. This can be done in both hypothesis–driven and hypothesis–free approaches to increase the probability of finding true predictors. Such a strategy would result in the identification of any clinically relevant predictors that may exist at the earliest possible time, allowing them to be integrated into subsequent clinical development and into an assessment of overall patient benefit–risk for the drug. Such discoveries could also provide mechanistic insights into drug disposition and patient–specific factors that influence response, and can therefore indirectly support related drug discovery and development efforts.

183

Regulatory variants other than *VKORC1* -1639 G>A may explain the effect on warfarin dose. *M. Cavalli[1], N. Eriksson[2], G. Pan[1], H. Nord[1], S.J. Connolly[3], M.D. Ezekowitz[4], S. Yusuf[3], L. Wallentin[2], M. Wadelius[5], C. Wadelius[1].* 1) Department of Immunology, Genetics and Pathology and Science for Life Laboratory, Uppsala University, Sweden; 2) Uppsala Clinical Research Center and Department of Medical Sciences, Uppsala University, Uppsala. Sweden; 3) Population Health Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, ON, Canada; 4) Sidney Kimmel Medical College, Lankenau Medical Center, Thomas Jefferson University, Villanova, PA, USA; 5) Department of Medical Sciences and Science for Life Laboratory, Uppsala University, Uppsala, Sweden.

Background: Warfarin is the most commonly used oral anticoagulant for thrombotic disorders and atrial fibrillation. The required dose is highly influenced by genetic variation of *CYP2C9* and *VKORC1*. *CYP2C9* variants are coding, but *VKORC1* variants are non-coding and thought to act through regulation of gene expression. Due to high linkage disequilibrium (LD) in the *VKORC1* region, it has not been resolved which variant that mediates the effect. ENCODE data can be used to search for functional variants in regulatory transcription factor (TF) binding sites. We aimed to determine which variants that regulate *VKORC1*. **Methods:** A genome-wide association study (GWAS) was conducted in 982 warfarin treated patients from the RE-LY genomics study. Genotyping was performed with the Illumina Human610-quad chip at Uppsala SciLife SNP&SEQ Technology platform. We sequenced the liver cell line HepG2 to locate all heterozygous positions in LD r^2>0.8 with the GWAS top hits, and used ENCODE data to find variants bound in an allele-specific (AS) way by TFs (AS-SNPs). The functional effect of allele-specific SNPs was evaluated using luciferase assays and over-expression of candidate TFs. **Results:** *VKORC1* -1639 G>A (rs9923231) was as expected among the top GWAS hits in RE-LY. However, rs9923231 was not located in a regulatory element according to ENCODE, nor was there any difference in transcriptional activity between the two alleles in luciferase assays. According to ENCODE data, there was one AS-SNP on the same haplotype as rs9923231 in HepG2, rs56314408. A second SNP, rs2032915, was located in the same liver enhancer 20 bp apart from rs56314408. The C alleles of rs56314408 and rs2032915 showed higher transcriptional activity in luciferase assays, which was further increased after over-expression of the TFs YY1 and USF1. Both functional candidates were significantly associated with warfarin dose in the RE-LY GWAS (rs56314408 p=2.9e-67 and rs2032915 p=3.9e-70). **Conclusions:** The conventionally analyzed *VKORC1* -1639 G>A (rs9923231) is not located in a regulatory element in the liver, and has no evidence of being a causative variant. We propose that rs9923231 predicts warfarin dose adequately only in populations where it is in high LD with the causative variants rs56314408 and rs2032915.

184

Assessing the Clinical Impact of Ethnicity-Specific Pharmacogenetic Allele Variation in over 100,000 Patients with Biobank-linked Electronic Medical Records. *N. Gonzaludo[1], T.J. Hoffmann[2], D.K. Ranatunga[3], C. Schaefer[3], N. Risch[2,3], P.Y. Kwok[2].* 1) Department of Bioengineering & Therapeutic Sciences, University of California, San Francisco, San Francisco, CA; 2) Institute for Human Genetics, University of California, San Francisco, San Francisco, CA; 3) Kaiser Permanente Northern California Division of Research, Oakland, CA.

Pharmacogenetic information can be extremely useful in optimizing patient therapy and avoiding adverse clinical outcomes, potentially reducing the cost burden of hospitalizations and treatment of adverse drug events. As part of the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH), we analyzed 102,979 members of the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort in one of the largest, most ethnically diverse pharmacogene characterization studies to date. Leveraging data derived from a biobank linked to a comprehensive electronic medical record (EMR), we assessed cohort metabolizer statuses for 7 drug-gene interactions (DGIs) for which there is moderate to strong evidence suggesting the use of pharmacogenetic information to guide therapy. Genetic data were translated to star allele diplotypes and clinical implementation guidelines were utilized to derive metabolizer status phenotypes for each drug. We quantified the large variation in ethnicity observed for star allele and metabolizer status phenotype frequencies, and found that 89% of the cohort had at least one actionable allele for the 7 DGIs in this study (90% of Non-Hispanic White, 76% of African American, 81% of Latino, 93% of Asian, 88% of Other/Uncertain). 13% total were considered to be at high risk for an unfavorable drug response. We then leveraged over a decade's worth of pharmacy data in RPGEH to retrospectively assess the clinical relevance of this information. 66% of the cohort had been exposed to at least one of 33 drugs with pharmacogenetic-based prescribing guidelines. For the 7 DGIs in our study, we found that 17,747 individuals had been prescribed a drug for which they had an actionable or high-risk metabolizer status phenotype. That is, had pharmacogenetic information been available at point-of-care for 17% of GERA, these individuals would have received a more personalized, optimal drug or dose than the standard drug regimen. Our study demonstrates the potential utility and clinical impact of using biobank-linked EMRs to derive pharmacogenetic information, as well as highlights the high frequency and high ethnic variability of pharmacogenetic variants.

185

Efficacy of whole genome sequencing over a lifetime: medically actionable genomic mutations in 300 patients. *M. He[1,2,3], M. Brilliant[1,3].* 1) Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI; 2) Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI; 3) Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI.

Statement of purpose: Advances in genomic medicine have the potential to change the way we treat diseases, but translating these advances into patient care stems from our ability to discover disease and/or drug associated medically actionable genomic mutations. Next-generation sequencing (NGS) technologies are increasingly used to find disease and/or drug associated genes. Integrating functional characterization of identified mutations with thorough genome interpretation and clinical data can provide compelling evidence implicating new disease/drug-contributing mutations in phenotypically well-characterized patients. The rise of large-scale data in NGS will contribute to better treatment paradigms, leading to improvements in diagnosis and targeted medications that may ultimately lead to an overall cost-savings in health care. Our objective was to detect medically actionable genomic mutations using patients' whole genome sequencing data and comprehensive clinical data.**Methods used**: To investigate medically actionable genomic variants for potentially use in disease diagnosis and personalized treatments, we classified medically actionable pathogenic genomic mutations in 192 genes in 300 deceased patients with nearly complete long-term medical records. The genes include 56 genes recommended by the American College of Medical Genetics and Genomics (ACMG), 60 additional "actionable" genes (Amendola et al., 2015), and non-overlapping genes from the 84 "very important pharmacogenes" (VIPs) defined by the Pharmacogenomics Research Network (PGRN). To infer biological insights from massive amounts of NGS data and comprehensive clinical data in a short period of time, we developed an in-house analysis pipeline within a software framework called SeqHBase (He et al., 2015) to quickly identify disease/drug-contributing genetic variants/genes.**Summary of results**: Among the 300 participants, 17/300 (5.67%) had a pathogenic variant annotated by ClinVar while 2 (0.67%) had likely pathogenic variants in the 56 ACMG genes; 32/300 (10.67%) had a pathogenic variant while 6 (2.00%) had likely pathogenic variants in the 116 (56 ACMG + 60 "actionable") genes; 297/300 (99.00%) had at least one pathogenic variant while 299 (99.67%) had at least one drug response variant in the 84 VIPs. This work shows an estimate of medically actionable genomic mutations, which can be potentially used for clinical diagnosis and personalized treatments, expected from whole genome sequencing.

186

Genetic variation in *STAT4* predicts response to interferon-α therapy for HBeAg-positive chronic hepatitis B. *D. Jiang[1,2,3,4,5,6], X. Wu[7], J. Qian[2,3], X.P. Ma[2], J. Yang[2,3], Z. Li[8], R. Wang[9], L. Sun[9], F. Liu[2,3,4,5,6,10], P. Zhang[2-5], X. Zhu[7], J. Wu[7], K. Chen[7], L. Zheng[1,6], D. Lu[1,2], L. Yu[1,11], Y. Liu[7], J. Xu[1,2,3,4,5,10].* 1) NorthShore University HealthSystem, The University of Chicago, Evanston, IL; 2) State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, P.R. China; 3) Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, P.R. China; 4) Center for Genetic Epidemiology, School of Life Sciences, Fudan University, Shanghai, P.R. China; 5) Center for Genetic Translational Medicine and Prevention, Fudan University, Shanghai, P.R. China; 6) Center for Cancer Genomics, Wake Forest School of Medicine, Winston-Salem, NC, USA; 7) National Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, Beijing, P.R. China; 8) Department of Infectious Disease, Affiliated Youan Hospital, Capital University of Medical Science. Beijing, P. R. China; 9) Xiamen Amoytop Biotech Co., LTD., Xiamen, Fujian Province, P. R. China; 10) Fudan Institute of Urology, Huashan Hospital, Fudan University, Shanghai, P.R. China; 11) Institute of Biomedical Science, Fudan University, Shanghai, P.R. China.

Purpose: Interferon (IFN)-α is the first-line therapy for HBeAg-positive chronic hepatitis B (CHB) patients, but produces response in only 30-40% of patients. Considerable efforts have been made to optimize IFNα treatment through pretreatment patient selection based on outcome predictors. However the use of both host and virus related variables that predict the response to therapy have not proved to be clinically useful. Our recent genome-wide association study identified a genetic variant rs7574865 in *STAT4*, a key component of the IFN-JAK-STAT pathway, to be associated with the risk of CHB and hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC). We aimed to assess whether this variant is a predictor of response to IFNα treatment of HBeAg-positive CHB patients. **Method:** We studied 466 HBeAg-positive CHB patients who received either IFNα-2b (N=224) or pegylated (PEG)-IFNα-2a (N=242) therapy for 48 weeks and followed for additional 24 weeks. Treatment efficacy was assessed using the rate of sustained virologic response (SVR), which is defined as HBeAg seroconversion along with HBV DNA level <2000 copies/mL at week 72. Genotyping of rs7574865 in *STAT4* was performed using the Sequenom MassArray platform. Associations between SVR and genotypes of rs7574865 were tested using both univariate and multivariate logistic regression analyses. **Results:** Compared to rs7574865 GT/TT genotype, the GG genotype (a risk factor of CHB and HBV-related HCC) was significantly associated with a reduced SVR rate in both the patients who received IFNα-2b therapy (21.1% vs. 37.2%, P=0.01), and those who received PEG-IFNα-2a therapy (18.0% vs. 41.2%, $P=9.74\times10^{-5}$). In combined analysis of the 466 patients, rs7574865 GG genotype was associated with a ~50% decreased SVR rate (19.3% vs. 39.1%, $P=4.15\times10^{-6}$). A multivariable logistic regression model including rs7574865 and clinical variables showed that rs7574865 was the most significant factor for prediction of SVR in the 224 patients with IFNα-2b therapy (P=0.02) and the 242 patients with PEG-IFNα-2a therapy ($P=8.88\times10^{-5}$) as well as the overall 466 patients ($P=7.62\times10^{-6}$). **Conclusions:** The present study revealed that *STAT4* rs7574865 is a strong predictor for IFNα therapy response in HBeAg-positive CHB patients. Assessment of *STAT4* rs7574865 genotype of CHB patients prior to IFNα treatment has the potential to significantly improve the efficacy of IFNα treatment for CHB patients.

187

The Search for Mendelian Genes (MG) Using Whole Exome Sequencing (WES) - Lessons Learned from analysis of >5,000 cases. *N. Sobreira[1], S. Jhangiani[2], F. Schiettecatte[3], C. Boehm[1], K. Doheny[1], T. Gambin[2], Z. Akdemir[2], D. Muzny[2], R. Gibbs[2], E. Boerwinkle[4], W. Wiszniewski[2], J. Lupski[2], A. Hamosh[1], D. Valle[1].* 1) Johns Hopkins University School of Medicine Baltimore, MD; 2) Baylor College of Medicine Houston, TX; 3) FS Consulting, LLC Salem, MA; 4) The University of Texas Health Science Center at Houston Houston , TX.

The number of genes responsible for Mendelian phenotypes has increased dramatically over the last 5 years, in part due to the wide-spread implementation of WES. The Baylor-Hopkins Center for Mendelian Genomics (BHCMG) is one of the 3 Centers for Mendelian Genomics funded by NHGRI/NHLBI with the goal of discovering Mendelian genes (MG). As of April 2015 (3.5 years into the project), BHCMG has collected 7,146, sequenced 5,788 and analyzed 4,972 samples. More than 3,700 unrelated probands and their phenotypic descriptions have been submitted to BHCMG through PhenoDB (www.mendeliangenomics.org) where the phenotypic description and exome sequencing data are stored in a searchable format. We continue to add functionality to PhenoDB and are currently testing 3 different algorithms to match probands in the database on a phenotypic basis. We have completed the analysis of at least 907 unrelated families with the identification of 189 novel and 168 known MG. For 153 of the known MG, we have gained a better appreciation of the associated phenotypic spectrum (i.e. phenotypic expansion). Interestingly, at least 5 probands have a blended phenotype resulting from the co-occurrence of two Mendelian phenotypes. Proof of causality for a variant and/or gene has been one of our main challenges and often discovery of variants in the same gene in multiple unrelated probands with similar phenotype is the strongest evidence. To assist in finding individuals with the same phenotype and/or variants(s) in the same gene(s) outside the project we developed GeneMatcher (www.genematcher.org). As of June 1st, GeneMatcher contains 2,178 genes, from 486 submitters from 38 countries and have made 307 matches. To optimize data sharing we are also part of the Matchmaker Exchange and together with PhenomeCentral and DECIPHER have developed an API that allows data sharing across databases. Our experience has emphasized: i) the value of rigorous phenotyping; ii) the utility of archiving phenotypes in a searchable relational database (PhenoDB); iii) the importance of data sharing for MG discovery (GeneMatcher); iv) increasing appreciation of the phenotypic spectrum for rare disorders (phenotypic expansion); and, v) blended phenotypes sometimes account for atypical phenotypes. MG discovery leads us to a better understanding of biological systems, enables precise diagnosis and counseling and is a step forward in the path towards informed treatment and the goals of precision medicine.

188

Discovery of novel dominant and recessive causes of severe developmental disorders, in coding and non-coding sequences. *M. Hurles, The Deciphering Developmental Disorders (DDD) Study.* Wellcome Trust Sanger Inst, Cambridge, United Kingdom.

To delineate the genetic architecture of severe undiagnosed developmental disorders in UK children we have deeply phenotyped 13,958 affected children and their parents through a nationwide network of clinical geneticists, and recruited the families into a genetic research study entitled the Deciphering Developmental Disorders (DDD) study. Using a trio-based exome sequencing strategy, we have analysed 4,295 families to date. Incorporating newly discovered genes in tandem with improved detection of post-zygotic mutations and cryptic structural variants has enabled us to increase the diagnostic yield in this cohort to ~35%. We have applied computational analysis of structured phenotypic data (Human Phenotype Ontology) in our diverse cohort, in combination with genotype-based enrichment tests, to discover ~20 novel dominant (validation ongoing) and 4 novel recessive disorders. These novel analytical strategies and new disorders will be described in detail. In addition, in this cohort we have identified 450 de novo mutations in the most highly conserved 2% of non-coding sequences, including several recurrently mutated putatively regulatory sequences. We will describe analyses of these non-coding mutations specifically with regard to their impact on the gain or loss of binding of known transcription factors.

189

Defining Variation Sensitive Regions in Genes Associated with Disease. *A.N. Abou Tayoun[1,2,4], S.H. Al Turki[1,4], M.S. Lebo[2,3], H.L. Rehm[2,3], S.S. Amr[2,3].* 1) Harvard Medical School Genetics Training Program, Cambridge, MA; 2) Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA; 3) Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, MA; 4) Co-first authors.

The classification of genetic variants is a major challenge facing the widespread adoption of comprehensive clinical genomic sequencing and the field of personalized medicine in general. Because most variants needing assessment do not have functional, genetic or population data to support clinical classification, there is an urgent need for novel approaches towards prioritizing candidate disease-causing variants. Although systematic evaluation of gene-disease associations can largely eliminate unnecessary interpretation of variants in genes with weak disease association, a large number of variants in clinically valid genes can still pose a major interpretation challenge, especially for diseases with substantial genetic and phenotypic heterogeneity such as hearing loss and cardiomyopathy.To improve variant interpretation, we leveraged the Exome Aggregation Consortium (ExAC) dataset (n=~60,000) and our clinical database, which consists of around 10,000 clinically curated variants in 163 genes identified in over 10,000 probands mostly with cardiomypathies, Noonan spectrum disorders or hearing loss. Comparing the two datasets, we performed systematic evaluation of domain (n=919 unique domains) and exon (n=4328 unique exons) level disease association. We statistically identify regions that are most sensitive to functional variation in the general population and also most commonly impacted in symptomatic individuals.Our data show that a significant number of exons and domains in genes strongly associated with disease can be defined as disease sensitive or tolerant leading to re-classification of at least 20% (495 out of 2458) variants of uncertain clinical significance in the 163 genes. This approach leverages domain functional annotation and associated disease in each gene to prioritize candidate disease variants, increasing the sensitivity and specificity of variant assessment within these genes.

190

Individual Clinical Variation, Beyond Monogenic Disease: the aggregation of pathogenic variant alleles in a personal genome. *J.E. Posey[1], T. Harel[1], J.A. Rosenfeld[1], P. Liu[1,2], Z. Niu[1,2], F. Xia[1,2], R.E. Person[1,2], M. Walkiewicz[1,2], D.M. Muzny[2,3], C.M. Eng[1,2], E. Boerwinkle[1,3,4], A.L. Beaudet[1,2], S.E. Plon[1,3,5,6,7,8], R.A. Gibbs[1,2,3], Y. Yang[1,2], J.R. Lupski[1,3,5,6].* 1) Dept of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX; 2) Baylor Miraca Genetics Laboratories, Baylor College of Medicine, Houston, TX; 3) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 4) Human Genetics Center, University of Texas Health Science Center, Houston, TX; 5) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 6) Department of Pediatrics, Texas Children's Hospital, Houston, TX; 7) Texas Children's Cancer Center, Houston, TX; 8) Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX.

The unbiased genomic approach of whole exome sequencing (WES) provides enhanced clinical utility compared to locus-specific genetic analyses - it is not limited by clinical ascertainment of a particular class of disorders, nor a singular unifying clinical diagnosis. To determine the frequency and molecular features of those individuals for whom more than one molecular diagnosis related to phenotype was reported, we analyzed 1538 diagnosed cases representing 27.0% (1538/5700) of sequential WES referrals to a clinical diagnostic lab. In 5.5% (84/1538) of cases, the molecular diagnosis involved two (N=78) or three (N=6) disease loci. CNVs contributed to 3 cases, and *de novo* variants contributed to 36.3% (57/157) of diagnoses, including 52.9% (48/85) of autosomal dominant and 50% (9/18) of X-linked diagnoses. Of the 72 cases with parental samples available, a combination of inherited and *de novo* events were responsible for 51.4% (37/72), and two separate *de novo* events occurred in 12.5% (9/72). Whereas some individuals with two independent *de novo* diagnoses had a blended phenotype consisting of a compilation of two distinct conditions, such as Aarskog-Scott syndrome with lissencephaly, or macrocephaly and autism with epileptic encephalopathy, a few patients with *de novo* diagnoses exhibited a modified phenotype involving variants at two related disease loci, such as X-linked and autosomal dominant intellectual disability syndromes. The relative frequency of *de novo* events in cases with multiple diagnoses supports the Clan Genomics hypothesis that posits a substantial role for recently arisen, private variants in human disease. These findings underscore the genetic complexity of undiagnosed disease and illustrate that the diagnostic range of WES extends beyond monogenic disease to detection of pathogenic alleles in multiple unrelated disease loci. Enabling CNV detection in WES may lead to an even higher rate of dual molecular diagnoses. Widespread clinical implementation of WES as a molecular diagnostic assay may spark a paradigm shift in medicine with the concept of genetic disease potentially evolving from monogenic, sometimes digenic disorders to more complex phenotypes driven by an aggregation of individual variation at multiple loci within a personal genome.

191

99 Lives Cat Genome Sequencing Initiative – discovery of feline models for human diseases - every life counts. *L.A. Lyons[1], E.K. Creighton[1], H.C. Beale[2], M-C.W. Lee[2], B. Gandolfi[1], 99 Lives Cat Consortium.* 1) Veterinary Medicine & Surgery, College of Veterinary Medicine, University of Missouri - Columbia, Columbia, MO; 2) Maverix Biomics, Inc. San Mateo, CA 94402 USA.

The 99 Lives cat whole genome sequencing (WGS) initiative is a research community-based effort for sequencing the genomes of > 99 cats to: 1) identify normal and abnormal genetic variation, 2) identify causative variants for specific health concerns, 3) support conservation efforts of wild felids, 4) improve the cat genome assembly, and 6) allow veterinarians to provide individual genome sequencing for state of the art health care (similar to the developing standard of care for humans). Currently, illumina HiSeq 100+ bp paired end sequencing reads is conducted from two PCR-free libraries per cat of 350 bp and 550 bp for at least 15X - 30X coverage. Maverix Biomics aligns the reads to the cat genome Felis_catus-6.2 and perfoms variant calling using FreeBayes or PLATYPUS. The cat reads and SNPs are overlaid onto a UCSC-type browser for viewing ease. Data tables provide the identified SNPs and their effects in specific genes, regions, chromosomes, or the entire genome for individual or different groups of cats. Maverix will provide a cumulative analysis of the variant calls for all cats after large groups or specific milestones (50, 60, 80 and 99) of new cat genomes have been added to the database. Currently, the analysis of 54 cats is available from over a dozen collaborators including 50 domestic cats and two species of wild felids. Contributors are expected to provide the basic signalment of the cats when possible, such as gender, breed, place of origin, and coat color. Other phenotypes are available via collaboration. The project has already identified variants for progressive retinal atropies in *AIPL1* and *IQCB1*. A variant in *COLQ* in Devon Rex myopathic cats suggests the first feline model for congenital myasthenic syndrome. Polycystic kidney disease, a well-known Persian cat disease, has now been discovered in the Pallas Cat (*Otocolobus manul*). A new gene model for Ehlors-Danlos syndrome in *CCDC80* and a novel blindness gene may support the identification of undiagnosed human patients. Causal variants for two forms of inherited lymphomas are being confirmed, as well as variants for feline forms of spondylocostal dysostosis and dwarfism. Analyses of trios, duos and singleton cases have all lead to variant discoveries for cat models of human disease. WGS of several cat breeds with hypertrophic cardiomyopathies and other cardiac diseases are ongoing. Analyzed sequences and variants are being transferred to appropriate public genomic databases.

192

Post-zygotic Point Mutations Are an Underrecognized Source of Novel Genomic Variation. *R. Acuna-Hidalgo[1], M.P. Kwint[1], T. Bo[2], M. van de Vorst[1], M. Pinelli[3], J.A. Veltman[1,4], A. Hoischen[1], L.E.L.M. Vissers[1], C. Gilissen[1].* 1) Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Institute of Neuroscience, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, the Netherlands; 2) State Key Laboratory of Medical Genetics, Central South University, 110 Xiangya Road, Changsha, Hunan 410078, China; 3) Telethon Institute of Genetics and Medicine, Pozzuoli, 80078 Naples, Italy; 4) Department of Clinical Genetics, Maastricht University Medical Centre, Universiteitssingel 50, 6229 ER Maastricht, the Netherlands.

De novo mutations are recognized both as an important source of genetic variation and as a prominent cause of sporadic disease in humans. Mutations identified as *de novo* are generally assumed to have occurred during gametogenesis and, consequently, to be present as germline events in an individual. Because Sanger sequencing does not provide the sensitivity to reliably distinguish somatic from germline mutations, the proportion of *de novo* mutations that occur somatically rather than in the germline remains largely unknown. To determine the contribution of post-zygotic events to *de novo* mutations, we analyzed a set of 107 *de novo* mutations in 50 parent-offspring trios. Using four different sequencing techniques, we found that 7 (6.5%) of these presumed germline *de novo* mutations were in fact present as mosaic mutations in the blood of the offspring and were therefore likely to have occurred post-zygotically. Furthermore, genome-wide analysis of "*de novo*" variants in the proband led to the identification of 4 out of 4,081 variants that were also detectable in the blood of one of the parents, implying parental mosaicism as the origin of these variants. Thus, our results show that an important fraction of *de novo* mutations presumed to be germline in fact occurred either post-zygotically in the offspring or were inherited as a consequence of low-level mosaicism in one of the parents.

193

The hunt for rare disease diagnosis: utilization of social media, model organisms, and pathway analysis in pediatric exome sequencing. *A.I. Nesbitt[1], E. Denenberg[1], Z. Yu[1], S.W. Baker[1], K.B. Pechter[1], E. Dechene[1], H. Dubbs[4], E. Bedoukian[2], A. Wilkens[2], L. Medne[2], X. Ortiz-Gonzale[3,5], E. Zackai[2,3], I. Krantz[2,3], M. Deardorff[2,3], A. Santani[1,3].* 1) Division of Genomic Diagnostics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; 2) Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; 3) Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; 4) Department of Pediatrics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; 5) Division of Neurology, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.

Unlike single gene testing, exome sequencing allows for the concurrent analysis of multiple protein-coding regions. While exome sequencing facilitates identification of genes not yet associated with human disease, the paucity of clinically-relevant information about these genes limits the diagnostic utility of this test. Here we describe assessing novel, candidate genes utilizing information from Mendelian databases, published materials, model organisms, and social media. Whole exome sequencing data was analyzed with an in-house bioinformatics pipeline and filtered using Cartagenia software. Cases for which a pathogenic variant in a known Mendelian gene was not identified were further reviewed for deleterious variants in novel candidate genes using an integrated approach using the resources described above. Case 1: A proband with developmental delay, cardiovascular abnormalities and facial dysmorphisms was found to have a molecular change similar to another proband with a dedicated social media site. Both probands' phenotypes overlapped, and the social media site aided in the diagnosis of an uncharacterized intellectual disability syndrome. Case 2: Exome analysis was performed in a proband with axonal motor neuropathy, ophthalmoparesis and severe global developmental delay. Animal model analysis was performed using the Jackson Laboratory database. We will present findings associated with genes involved in motor neuropathy in mouse models. Case 3: A proband presenting with Noonan-like features: developmental delay, short stature, macrocephaly and persistent anterior fontanelle, were interrogated for rare variants in the Noonan-associated RAS/MAPK pathway. These results highlight the utility of exome sequencing in identifying novel, candidate genes. However, additional research assessing the functional impact of these variants on the protein is needed before making any changes to the clinical management of the patient. With several case studies, we illustrate how strategic utilization of traditional and non-traditional information resources in the analysis of clinical exome sequencing data aids in the diagnosis of rare genetic diseases. These data emphasize the significance of needing additional resources to perform research for functional studies and assessing gene-disease validity along with the importance of collaborative resources to perform follow-up studies.

194

Improving Diagnosis And Furthering Gene Discovery Through Recruitment Of Clinical Whole Exome Sequencing Cases Into Research. *Z.H. Coban Akdemir[1], M.K. Eldomery[1], T. Gambin[1], T. Harel[1], A. Stray-Pedersen[8], S. Penney[1], J.A. Rosenfeld[1], S.N. Jhangiani[2], D.M. Muzny[2], F. Xia[1,3], Y. Yang[1,3], C.M. Eng[1,3], S.E. Plon[1,3,7], V.R. Sutton[1,6], A.L Beaudet[1,3], E. Boerwinkle[2,4], R.A. Gibbs[1,2,3], J.R. Lupski[1,2,5,6].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; 3) Baylor Miraca Genetics Laboratories, Baylor College of Medicine, Houston, TX 77030, USA; 4) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; 5) Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA; 6) Texas Children's Hospital, Houston, TX 77030, USA; 7) Texas Children's Cancer Center, Texas Children's Hospital, Houston, TX 77030, USA; 8) Norwegian National Unit for Newborn Screening, Women and Children's Division, Oslo University Hospital, 0424 Oslo, Norway.

Beginning in July 2013, Baylor College of Medicine (BCM) added a new source of recruitment for subjects into the Baylor-Hopkins Center for Mendelian Genomics (BH-CMG): individuals undergoing clinical whole exome sequencing (WES) at the Baylor Miraca Genetics Laboratories (BMGL) for whom the clinical test failed to identify variants causative of the clinically observed abnormal phenotypes. Published analyses of the first 3,684 clinical exomes from the BMGL (Yang, et al. 2013; Yang, et al. 2014) show a molecular diagnostic rate of 25.2%, primarily in known disease genes, leaving ~75% of these cases unexplained and available for research recruitment and potential novel disease gene discovery. Since the beginning of this project (April 2014), we have enrolled nearly 380 subjects from approximately 142 families and have a queue of over 1600 additional families waiting for enrollment. Pilot project data of unsolved clinical exomes to BH-CMG includes 64 trios, 1 quartet and 1 family comprised of 3 affected siblings. We developed a workflow to identify likely causal variants in the affected child/children in 3 branches: 1) *de novo* variants; 2) homozygous/hemizygous SNVs or CNVs; and 3) compound heterozygous variants. This workflow yielded solutions for 31.8% of these cases. Solutions included discovery of novel disease genes, such as *PURA* (OMIM *600473), *MIPEP* (OMIM *602241) and *TANGO2*. Additionally, solutions for 8 cases were found in disease genes, including *de novo* variants in *ZBTB20* (OMIM *606025) and *NALCN* (OMIM *611549) that were published between the times of clinical and research analyses. Moreover, analysis of shared rare variants among 3 affected brothers revealed a previously reported *PIK3CD* (OMIM *602839) pathogenic variant transmitted from their father, who may have been a mosaic carrier. The pipeline also yields many leads for additional disease gene discovery, with 10 strong novel candidate genes currently undergoing additional analyses to determine pathogenicity. In addition, utilization of the GeneMatcher tool - a part of the Matchmaker Exchange Project- further facilitates our novel gene discovery efforts through the identification of additional families with a similar phenotype. In sum, this collaboration between clinical and research testing not only contributes to finding answers for families searching for a diagnosis, but also has become a fruitful and efficient method for furthering our understanding of human genetic disease.

195

Rare variants are a large source of heritability for gene expression patterns. *R. Hernandez[1,2,3], D. Vasco[1], L. Uricchio[1,4], C. Ye[2,5], N. Zaitlen[6,2,3].* 1) Bioeng. & Therapeutic Sci, UCSF, San Francisco, CA; 2) Institute for Human Genetics, UCSF, San Francisco, CA; 3) Institute for Quantitative Biosciences, UCSF, San Francisco, CA; 4) Department of Biology, Stanford University, Stanford, CA; 5) Epidemiology & Biostatistics, UCSF, San Francisco, CA; 6) Department of Medicine Lung Biology Center, UCSF, San Francisco, CA.

Understanding the genetic architecture of complex traits is a central challenge in human genetics. There currently exists a large disparity between heritability estimates from family-based studies and large-scale genome-wide association studies (GWAS), which has been sensationalized as the "missing heritability problem". Among the possible explanations for this disparity are rare variants of large effect that are not tagged by genotyping platforms. However, recent population genetic models suggest that the conditions under which rare variants are expected to substantially contribute to heritability may be fairly limited. To better understand the heritability of complex phenotypes, we investigated the role of *cis* alleles in gene expression levels across European and African individuals using RNA and whole genome sequencing data from the GEUVADIS and 1000 Genomes Projects. In particular, we investigate whether rare variants are likely to be a source of missing heritability in expression across genes. Using variance-component methods, we partitioned the heritability of expression levels explained by *cis* variants for each gene in the genome across several frequency bins from rare (≤1%) to common (>10%). We performed extensive simulations to validate our heritability estimation procedure. We find that when pooling all variants in *cis* (within 500kb of a gene), heritability estimates are on average h_c^2=17.6% (with 4.7% of genes having h_c^2>50%). Using variance-component methods, we find that in *cis*, rare variants (MAF ≤ 1%) contribute significantly more heritability than common variants (MAF > 10%) across genes (p_{MWU}=1.1×10^{-6}). In particular, 35.6% of h_c^2 across genes is contributed by rare variants, while common variants contribute 22.3%. This observation suggests that rare variants play a substantial role in the heritability of gene expression patterns, which is inconsistent with neutral evolutionary forces operating on the *cis* regulatory architecture of most genes. We discuss our results in the light of recent population genetic models of quantitative traits, and highlight the importance of understanding how natural selection can shape the genetic architecture of gene expression in humans. We conclude by discussing implications for studying a variety of complex phenotypes in humans.

196

Haplotypes of common SNPs explain a large fraction of the missing heritability of complex traits. *G Bhatia[1,2], A Gusev[1,2], P Loh[1,2], BJ Vilhjálmsson[1,2,3], S Ripke[4], S Purcell[2,5,6,7], E Stahl[2,8,9], M Daly[2,5], TR de Candia[10], MC O'Donovan[11], SH Lee[12], N Wray[12], BM Neale[2,5], MC Keller[13], NA Zaitlen[14], B Pasaniuc[15], J Yang[12,16], AL Price[1,2,17], Schizophrenia Working Group of the Psychiatric Genomics Consortium.* 1) Epidemiology, Harvard School of Public Health, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 3) Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark; 4) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 5) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 6) Department of Psychiatry, Mt. Sinai Hospital, NY, USA; 7) Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, NY, USA; 8) Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 9) Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 10) Department of Psychology and Neuroscience, University of Colorado, Boulder, Boulder, CO, United States; 11) MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK; 12) The Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia; 13) Institute for Behavioral Genetics, University of Colorado, Boulder, Boulder, CO, USA; 14) Lung Biology Center, School of Medicine, University of California, San Francisco, San Francisco, CA, USA; 15) Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, CA, USA; 17) The University of Queensland Diamantina Institute, The Translation Research Institute, Brisbane, Queensland, Australia; 18) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

While genome-wide significant associations generally explain only a small proportion of the narrow-sense heritability of complex disease (h^2), recent work has shown that more heritability is explained by all genotyped SNPs (h_g^2). However, much of the heritability is still missing ($h_g^2 < h^2$). For example, for schizophrenia, h^2 is estimated at 0.7-0.8 but h_g^2 is estimated at ~0.3. Efforts at increasing coverage through accurately imputed variants have yielded only small increases in the heritability explained, and inclusion of poorly imputed variants can lead to assay artifacts for case-control traits. We propose to estimate the heritability explained by a set of haplotype variants (haploSNPs) constructed directly from the study sample (h_{hap}^2). Our method constructs haploSNPs by comparing all pairs of computationally phased haploid chromosomes. A shared segment begins at a SNP at which the chromosomes match and extends until a mismatch that violates a four-gamete test. Identical shared segments are merged and the haploSNP values (0, 1 or 2 copies per individual) are used to estimate genetic relationships between individuals. These genetic relationships are used in a linear mixed model to estimate the heritability explained by this set of haploSNPs. Using data from the UK10K project, we show that haploSNPs constructed from common SNPs do a substantially better job of tagging unobserved variants (average best tag r^2 = 0.54) than genotyped or well-imputed SNPs (average best tag r^2 = 0.28 and 0.34, respectively). This improved tagging is highly statistically significant after accounting for the larger number of haploSNPs ($P < 10^{-15}$) and translates into substantial gains in heritability explained relative to genotyped SNPs. In a large schizophrenia data set (PGC2-SCZ), haploSNPs with MAF > 0.1% explained substantially more phenotypic variance (h_{hap}^2 = 0.64 (s.e. 0.10)) than genotyped SNPs alone (h_g^2 = 0.32 (s.e 0.03)). These estimates were based on cross-cohort comparisons, ensuring that cohort-specific assay artifacts did not contribute to our estimates. In a large multiple sclerosis data set (WTCCC2-MS), we observed an even larger difference between h_{hap}^2 and h_g^2, though data from other cohorts will be required to validate this result. Overall, our results suggest that haploSNPs can explain a large fraction of the missing heritability of complex disease suggesting a substantial role for untyped rare variants in the genetic architecture of these traits.

197

Estimating the respective contributions of human and viral genetic variation to HIV control. *I. Bartha[1], P. McLaren[1], Ch. Brumme[2], R. Harrigan[2], A. Telenti[3], J. Fellay[1].* 1) École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; 2) BC Centre for Excellence in HIV/AIDS, Vancouver, Canada; 3) J. Craig Venter Institute, La Jolla, USA.

Viral load setpoint is a major correlate of HIV disease progression. Genome-wide association studies have identified common human polymorphisms that together explain no more than 15% of its phenotypic variance. Here we present a joint assessment of the respective contributions of human and viral variation to setpoint. Human genotype data across the Major Histocompatibility Complex (MHC) region, full-length consensus HIV sequences and setpoint viral load results were available for 541 treatment naïve HIV-1 infected individuals. Heritability (h^2) estimation was carried out with GCTA using three kernel matrices representing: 1) the human relatedness across the MHC, 2) the viral phylogeny, and 3) the sample-specific noise. The human kernel was estimated from 27 common polymorphisms selected by LASSO. Phylogenetic trees were inferred from the viral sequences using RAxML. The viral kernel was derived from the phylogenetic trees by taking the branch length of the shared ancestry. Estimating the host heritability of viral load using the host kernel alone yielded a median estimate of h^2=8% with an interquartile range (IQR) of 1% across 15 bootstrap replicates of the samples. The estimates of the viral heritability drawn from 30 bootstrapped viral trees had a median of 29% (IQR=10%). Combining the host and viral relatedness matrices showed a comparable viral heritability of 26% (IQR=9%) but a decreased host contribution of 4% (IQR=0%). This is the first estimate of the combined and respective contributions of the host and the viral genomes to the observed variability of HIV viral load. We showed that both the pathogen and host genomes have detectable impacts on the clinical outcome of infection, which are however not independent. It was shown before that a large portion of the viral sequence is under selection pressure by the host genotype. Therefore the fraction of phenotypic variance that is attributable to the viral phylogeny but not to the genotype of the current host represents the effects of previous adaptations of the virus to the previous individuals in the transmission chain. This demonstrates that the fraction of phenotypic variance attributable to the genetics of current host does not explain all the human genetic control on HIV viral load at the population level.

198

Rare and Low-frequency coding variants contribute independently to human stature variation. *M.C. Medina Gomez[1,2], E. Marouli[3], M. Graff[4], K.S. Lo[5], K. Lu[6], C. Schurmann[6], H.M. Highland[4,7], N. Heard-Costa[8], C.M. Lindgren[9], D. Liu[10], I.B. Borecki[11], J.N. Hirschhorn[12,13,14], R.J.F. Loos[6], T.M. Frayling[15], F. Rivadeneira[1,2], G. Lettre[5,16], P. Deloukas[3,17] On behalf of the BBMRI, the GOT2D, the CHARGE, and the GIANT Consortia.* 1) Department Internal Medicine, Erasmus MC University, Rotterdam, Netherlands; 2) Department of Epidemiology, Erasmus MC University, Rotterdam, The Netherlands; 3) TWilliam Harvey Research Institute, Barts and The London School of Medicine and Dentistry Queen Mary University of London Charterhouse Square, London, UK; 4) Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; 5) Montreal Heart Institute,Université de Montréal, Montréal,Québec , Canada; 6) The Genetics of Obesity and Related Metabolic Traits Program, The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 7) Human Genetics Center, University of Texas Health Science Center, Houston, TX, USA; 8) Department of Neurology, Boston University School of Medicine, Boston, MA,USA; 9) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 10) Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA; 11) Department of Genetics Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA; 12) Divisions of Endocrinology and Genetics and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA; 13) Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA; 14) Department of Genetics, Harvard Medical School, Boston, MA, USA; 15) Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK; 16) Faculty of Medicine, Université de Montréal, Montreal, Québec, Canada; 17) The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

Large-scale GWAS have discovered 697 common genetic variants within 423 loci associated with human stature, explaining altogether ~20% of adult height heritability. Most of these variants map to non-coding regions. To investigate the contribution of rare (<1% MAF) and low frequency (1-5% MAF) coding variants, we examined exome array genotype data in >450,000 individuals (~380,000 of European ancestry) from 148 studies. We performed single variant (SV) and gene-based association analyses using $p<2x10^{-7}$ and $p<2.5x10^{-6}$ as significant thresholds, respectively. To address the independent contribution of common and rare/low-frequency alleles in shaping height variation, we performed conditional analyses in the European set using RareMETAL. Focusing on non-synonymous and splice site variants with MAF<5%, we used SKAT and VT to perform gene-based tests. In the SV analysis, we identified 1,232 significant variants, of which 147 were either rare (49) or low-frequency (98). Of these 147 variants, 102 clustered within 500kb of known height loci and 45 were new (17 rare and 28 low frequency). We observed the largest effect sizes (~1.9 cm) for rare missense variants in *CRISPLD2* (rs148934412, MAF=0.07%; $P=2.9x10^{-14}$), *IHH* (rs142036701, MAF=0.07%; $P=1.3x10^{-14}$), which are ~10 fold larger than those typically observed for common height SNPs. We also identified rare missense variants in numerous genes part of growth-related pathways, including *ACAN, PTH1R, IHH* and *FBN2* known to harbor mutations causing abnormal skeletal growth. In the conditional analysis, we identified 563 independent significant signals (of which 32 are rare and 56 low-frequency). At least 76 of the new variants did not overlap with previously reported height loci. The gene-based analysis identified 97 genes associated with height, 53 of which map near known height loci. Most gene-based associations were driven by single coding variants. Yet, we identified 11 genes with statistical evidence for multiple variants located at known (e.g. *FLNB*, *CCDC3* and *B4GALNT3*) and novel loci (e.g. *OSGIN1*, *NOX4*, *UGGT2*). While successful, gene-based methodology is highly dependent on type of statistical tests and functional annotation of variants considered. We conclude that rare and low-frequency coding variants among known and novel loci contribute independently to human stature variation, with effects sizes larger than observed for common variants. These coding variants implicate biologically relevant genes more precisely.

199

Imputation of rare variants from the new Haplotype Reference Consortium identifies associations missed by 1000 Genomes. *A.R. Wood[1], R. Beaumont[1], D. Hernandez[2,3], M. Nalls[2], J.R. Gibbs[2], S. Bandinelli[4,5], M.N. Weedon[1], A. Murray[1], A. Singleton[2], D. Melzer[1], L. Ferrucci[6], T.M. Frayling[1], Haplotype Reference Consortium.* 1) Institute of Biomedical and Clinical Sciences, University of Exeter Medical School, Exeter, United Kingdom; 2) Laboratory of Neurogenetics, National Institute of Aging, Bethesda, Maryland 20892, USA; 3) Department of Molecular Neuroscience and Reta Lila Laboratories, Institute of Neurology, UCL, London WC1N IPJ, UK; 4) Tuscany Regional Health Agency, Florence, Italy, I.O.T. and Department of Medical and Surgical Critical Care, University of Florence, Florence, Italy; 5) Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy; 6) Longitudinal Studies Section, Clinical Research Branch, Gerontology Research Center, National Institute on Aging, Baltimore, Maryland 21225, USA.

BackgroundThe phenotypic consequences of rare variation in the 99% of the genome that is non-coding is still largely unexplored. Whole genome sequencing remains prohibitively expensive to perform on a GWAS scale. The ability to impute rarer variation from the Haplotype Reference Consortium (HRC) (including >65,000 haplotypes from 20 whole-genome sequencing studies) offers the chance to perform GWAS of common traits using a much larger set of these variants. AimWe aimed to assess the ability of imputation from the HRC to detect novel signals of association with 93 common traits using 1210 individuals from the InCHIANTI study. These 93 traits included circulating biomarkers of relevance to human health, including vitamins, ions and, interleukins, and for which previous association signals are often detectable in samples of this size at genome-wide significance. MethodsWe estimated the haplotypes of 1,210 InCHIANTI individuals using SHAPEIT-v2. Using the remote "UM Imputationserver" we imputed 39,210,718 and 47,045,346 variants from the HRC and 1000 Genomes (phase 3 version 5), respectively, using Minimac-v3. We performed GWAS of 93 traits using imputed variants with imputation quality 'rsqr' >0.5, and compared our results based on the two datasets. ResultsImputation from the HRC and 1000 Genomes resulted in 15,501,447 and 13,238,838 variants with imputation quality >0.5, respectively. Of the well-imputed variants from the HRC, 2,253,813 were low frequency (1% ≤ MAF < 5%) and 7,809,194 were rare (<1% MAF) compared to 2,327,674 and 4,174.466 respectively for 1000 Genomes. We detected 43 association signals across 35 traits at $P<5x10^{-08}$ and 17 at $P<5x10^{-10}$. Of these 43 association signals 9 index variants were well-imputed rare variants (imputation rsqr >0.5) captured less well by 1000 Genomes (imputation rsqr <0.5). These rare variant signals included associations between rs117231518 and vitamin D (MAF=0.01; $P=3x10^{-08}$), rs150956780 and lactic dehydrogenase (MAF=0.006; $P=6x10^{-16}$), and rs568653750 and FT4 (MAF=0.002; $P=2x10^{-08}$), although only lactic dehydrogenase reached $P<5x10^{-10}$ (correction for ~100 traits tested) and replication is needed. In conclusion imputation from the HRC provides an opportunity to study the phenotypic role of a much larger proportion of rarer variation than previously possible and can identify novel putative associations.

200

Imputing Genotypes of the Haplotype Reference Consortium into the Haplotypes of a Large Case Control Study of Age-related Macular Degeneration. *L.G. Fritsche, S. Das, G.R. Abecasis, International AMD Genomics Consortium.* Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI.

Purpose: Age-related macular degeneration (AMD) is one of the leading causes of blindness in elderly Americans. Common genetic variants in more than 20 genes are known to modify disease risk. In addition, targeted sequencing studies of complement genes identified less common variants with larger risk effects. Here we set out to systematically elucidate the role of low frequency and rare variants in AMD. Methods: We used > 60,000 haplotypes of the Hapotype Reference Consortium (HRC; release 1; ~39 million autosomal variants) to impute into the phased genotypes of >50,000 samples, among others 16,144 advanced AMD cases and 17,832 controls of European decent (software: SHAPEIT and Minimac3). Our genotyping platform, a combination of genome-wide tagging SNPs, exonic variants, and custom content, was enriched for non-synonymous coding changes found in large sequencing studies. Results: Of the ~165,000 polymorphic, autosomal, non-synonymous coding changes of our genotyping platform, the HRC reference panel had ~70% overlapping variants [> 98% overlapping variants with minor allele frequency (MAF) > 0.05%]. Initial results showed that > 100,000 additional non-synonymous coding changes with MAF ≥ 0.05% could be well imputed with the HRC reference panel. Most interestingly, all of the four known rarer AMD risk variants (CFH:p.R1210C, CFI:p.G119R, C9:p.P167S, and C3:p.K155Q; 0.16% ≤ MAF ≤ 1.2%) that we genotyped could have been imputed with high quality (empirical R-square > 0.88) and were observed with association P values < $7x10^{-10}$ in our case control study. Conclusion: Our initial results suggested that genotyping and subsequent imputation with the first release of the HRC reference panel allow a comprehensive analysis of variants with MAF ≥ 0.05%. The observed imputability of confirmed AMD-associated variants of the lower frequency spectrum indicated that the discovery of such risk variants might not be restricted to targeted sequencing studies or dependent on the content of the selected genotyping platform. Imputation of the full data set is ongoing and future analyses that will be presented at the conference will include genome-wide single variant and gene-based burden tests.

201

Whole-genome sequencing and genotype imputation across 35,000 individuals further defines the genetic architecture of inflammatory bowel disease. *C.A. Anderson on behalf of the UK IBD Genetics Consortium.* Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom.

Over 200 risk loci have been identified to date for Crohn's disease (CD) and ulcerative colitis (UC), the two common forms of inflammatory bowel disease (IBD). To further explore the role of low-frequency (0.005 < minor allele frequency (MAF) < 0.05) and rare (MAF < 0.005) genetic variants in disease risk we performed low-depth (up to 6X) whole-genome sequencing across 4400 IBD cases (2700 CD, 1700 UC) and compared them to 3700 UK population controls sequenced as part of the UK10K project. Following quality control (QC), 28 million variants were available for association testing, 10 million of which were not seen in the 1000genomes project. To increase power to detect association we performed rare-variant burden tests within protein coding regions and observed the known enrichment of rare protein-coding variants at *NOD2* in CD cases. Within *NOD2*, eleven potentially new risk variants with MAF < 0.1% contributed to the overall burden, but were individually too rare to achieve significance in our single SNP association tests. Next, we combined our sequence data with that from the 1000genomes project and imputed genotypes into an additional set of 12,000 IBD cases (two thirds of which were newly genotyped for this study) and 15,000 UK population controls. In addition to detecting genome-wide significant evidence of association at many established risk loci ($P < 5\times10^{-8}$), we identified five novel loci driven by common genetic variants. IBD candidate genes within these loci include *CLECL1*, which has been previously associated with type-1 diabetes and multiple sclerosis, and *ITGAV*, which is a paralog of a gene (*ITGAL*) within an established IBD risk locus. We identified ten low-frequency variants with $P < 1\times10^{-6}$ by testing for association within established risk loci conditional on known genetic effects. These will be validated together with low-frequency SNPs with $P<1\times10^{-6}$ that lie outside of known IBD regions. In summary, we have performed one of the largest whole genome sequence-based association studies for a complex disease to date. Our results suggest that there are a limited number of low frequency variants with effects larger than those observed in GWAS (odds ratio > 1.5). While our study further extends the allele frequency spectrum tested for association to IBD risk, higher coverage sequencing of tens of thousands of individuals will be needed to fully elucidate the role of truly rare genetic variants in complex disease risk.

202

The next wave of autism gene discovery by targeted sequencing of thousands of patients. *H. Stessman[1], B. Xiong[1], T. Wang[1,2], K. Hoekzema[1], L. Vives[1], N. Janke[1], C. Lee[1], B. Coe[1], R. Bernier[3], E. Eichler[1,4].* 1) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA; 2) State Key Laboratory of Medical Genetics & School of Life Sciences, Central South University, Changsha, China; 3) Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA; 4) Howard Hughes Medical Institute, Seattle, WA.

De novo protein-coding mutations and large copy number variants are estimated to contribute to ~30% of simplex autism, but the pathogenicity of the majority of the several hundred candidate genes to emerge from exome sequencing studies has yet to be definitively established. We applied single-molecule molecular inversion probes (smMIPs) combined with an established statistical framework to resequence 154 candidate genes in 11,000 autism/intellectual disability probands as well as 4,000 unaffected controls. Samples were obtained from an international network of clinicians and researchers (the ASID network where patient recontact is possible); candidate genes were selected based on recurrent mutations and for enrichment in biological networks based on co-expression and protein-protein interactions. We report the identification of two or more likely gene-disrupting (LGD) events among probands in 44 (29%) of these genes with no disrupting events in unaffected siblings or controls. This set confirms the importance of previously described genes (*CHD8*, *ARID1B*, *PTEN*, *TBR1*, *GRIN2B*, *DYRK1A*, *ADNP*, *CHD2*, *SYNGAP1*, *TRIP12*, *SCN2A* and *PAX5*) as well as many novel autism genes that have reached locus-specific significance for LGD burden (e.g., *NAA15*, *POGZ*, *MED13L*, *TAF13*, *CDKL5*, *SRCAP*, *CUL3*, *DDX3X*, *WAC* and *SETD2*). In addition to *de novo* mutations, we are observing evidence of maternal transmission biases for private LGD events in genes such as *RIMS1* and *AHNAK*. Based on targeted resequencing of these 154 candidates, we currently estimate that we have identified a high-impact risk variant in 3-4% of patients using inexpensive smMIP assays now available to readily sequence tens of thousands of additional patients. Phenotypic and familial follow-up on genes with five or more disruptive mutations are ongoing and are revealing novel syndromic and non-syndromic forms of autism/intellectual disability. This study highlights the importance of *de novo* mutations in ASD and draws on the strength of the genotype-first approach to identify new subtypes and resolve the locus heterogeneity of a complex genetic disease.

203

Allele frequency distribution of pathogenic sequence variants in ExAC and the implications for clinical genetic testing. *Y. Kobayashi, S. Yang, A. McMurry, J. Garcia, S. Lincoln, K. Nykamp, S. Topper.* Invitae, San Francisco, CA.

A key criterion used in the clinical interpretation of sequence variants is the allele frequency observed in the general population. In the recently published ACMG guidelines (2015), "frequency greater than expected for a disorder" is considered strong evidence for benign classification. In principle, allele frequency thresholds can be derived for each gene based on disease incidence and penetrance of pathogenic mutations. In practice however, this is difficult since accurate estimates of incidence and penetrance are not available for most genes, and these can vary greatly depending on ethnic population. Moreover, these analyses often ignore the distribution of pathogenic mutations: some genes have only a few relatively common mutations while others have hundreds of very rare or private mutations. Finally, accurate and precise frequency measurements based on large populations have only recently become available. For all of these reasons, clinical laboratories have typically set their benign allele frequency thresholds conservatively high.To study the distribution of disease-causing alleles in population datasets, we identified pathogenic mutations in ClinVar and determined their allele frequencies in the Exome Aggregation Consortium (ExAC) dataset. We then generated frequency distributions of these mutations in a number of well-studied genes, including BRCA1/2 and CFTR, and derived an allele frequency-based model for identifying variants that are unlikely to be pathogenic. In these genes, the frequency threshold at which 99.9% of known pathogenic mutations were accounted for was substantially lower than current thresholds used by most labs, including ours. The remaining 0.1% above this threshold were essentially all known population-specific founder mutations. Furthermore, the cumulative total of pathogenic mutations in a gene was largely consistent with the expected frequency based on disease incidence and penetrance. Incorporation of this data into variant classification schemes gives clinical geneticists greater confidence that many low-frequency variants are benign; as a result, we observed upwards of 30% reduction in variants of uncertain significance when simulated on a clinical cohort.By leveraging the power of large community databases, we are able to better characterize the allele frequency distribution of disease-causing mutations. This knowledge will greatly improve our ability to accurately classify variants and may lead to better patient care.

204

Classifying Variants Detected by Whole Genome Sequencing of a Healthy Population: The Good, the Bad, and the Ugly. *S. Punj[1], Y. Akkari[1], M.O. Dorschner[2, 3], D.A. Nickerson[3], G.P. Jarvik[3,4], L.M. Amendola[4], D.K. Simpson[7], A. Rope[7], J. Reiss[7,8], K. Kennedy[8], D.I. Quigley[9], C. Harding[10], J. Berg[11], T. Kauffman[7], M. Gilmore[7], P. Himes[7], B. Wilfond[5,6], K.A.B. Goddard[12], C.S. Richards[1].* 1) Department of Molecular and Medical Genetics, Knight Diagnostic Laboratories, Oregon Health & Science University, Portland, OR; 2) Pathology, University of Washington, Seattle, WA; 3) Genome Sciences, University of Washington, Seattle, WA; 4) Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA; 5) Department of Pediatrics, Division of Bioethics, University of Washington, Seattle, WA; 6) Truman Katz Center for Pediatric Bioethics, Seattle Children's Hospital, Seattle, WA; 7) Department of Medical Genetics, Kaiser Permanente Northwest, Portland, OR; 8) Obstetrics and Gynecology, Kaiser Permanente Northwest, Portland, OR; 9) Laboratory, Kaiser Permanente Northwest, Portland, OR; 10) Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR; 11) Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, NC; 12) Center for Health Research, Kaiser Permanente Northwest, Portland, OR.

The NextGen Project of the Clinical Sequencing Exploratory Research Consortium (CSER) focuses on carrier screening by whole genome sequencing (WGS) in a pre-conception reproductive population. Based on literature review and other available carrier screening tests, 700 gene/disorder pairs were selected for analysis and classification of variants. Here we report the laboratory results in the first 75 participants. The conditions reported to date include: serious (~40%), mild (~20%), adult onset (~20%), shortened lifespan (~15%), unpredictable outcome (~10%), and medically actionable (<2%) as categorized by our Return of Results Committee (RORC). Well known variants in more prevalent conditions represent 20% of variants reported and include: *HFE, SERPINA1, GJB2, CPT2*, and *MEFV*. Not surprisingly, two couples were found to carry common disease-causing variants in the same gene. However, based on the respective conditions, neither disorder (hereditary hemochromatosis or alpha-1 antitrypsin deficiency) was an indication for prenatal diagnosis. More importantly, neither couple expressed anxiety or concern over their results. Unlike these more common conditions, ~35% of the conditions reported are not included in clinically available carrier screening tests, and in general, represent very rare conditions. All variants were classified using the 2015 ACMG recommendations for variant classification. Approximately half of the variants reported were classified as pathogenic, with the remaining classified as likely pathogenic. We do not report VUS or benign variants. While the majority of variants had been previously identified, novel variants that result in truncation of a gene with a loss of function mechanism (e.g. nonsense, frameshift, and splice variants), represented ~20% of variants reported. Classification of missense variants, which comprised the majority (67%) of variants returned, was more challenging. For example, we initially classified a missense variant in the *SPG7* gene, which causes spastic paraplegia type 7, as pathogenic; however, after finding the same variant in three out of the first 50 participants, we consulted our RORC and reclassified it as a VUS. We will describe how our laboratory uses the current ACMG recommendations for variant classification, which criteria are most often used or unused, and provide some strategies to avoid potential pitfalls. Finally, we will discuss insights into the utility of our WGS approach for carrier screening.

205

Exploring the landscape of pathogenic genetic variation in the ExAC population database: Insights of relevance to variant classification. *S. Gardner (Equal Contribution), W. Song (Equal Contribution), H. Hovhannisyan, W. Chen, A. Natalizio, K. Bogdanova, K. Weymouth, I. Thibodeau, S. Letovsky, A. Willis, N. Nagan.* Integrated Genetics, Laboratory Corporation of America® Holdings, Westborough, MA and Research Triangle Park, NC.

PURPOSE AND METHODS: We piloted efforts to evaluate the feasibility of the Exome Aggregation Consortium (ExAC) database as a control cohort to classify variants across 3 groups of genes, namely, dominant tumor suppressor genes (*BRCA1, BRCA2, MLH1, MSH2, MSH6,* and *PMS2*), dominant cardiac disorder genes (n=9, *FBN1*, sarcomeric and desmosomal genes), and recessive genes (*CFTR, GJB2, HBB*, and *MEFV*). Our approach involved: 1. Comparing the frequency of pathogenic variants in the ExAC to the calculated maximum expected pathogenic allele frequency after factoring for the prevalence, penetrance, allelic and locus heterogeneity of each gene. Variant classification is based on internal database annotations. 2. Comparing the observed carrier frequency and the ethnicity-specific variant distribution between ExAC and published literature for genes with available information. **RESULTS**: 1. 95% of *BRCA1&2*, 100% of *CFTR*, *GJB2,* and *HBB*, and 89% of *MEFV* variants with ExAC frequencies above the calculated maximum expected pathogenic allele frequency were classified as "Normal." 2. In contrast, several cardiac and Lynch syndrome gene variants with ExAC frequencies above the calculated maximum expected pathogenic allele frequency were classified as "VUS," making these candidates worthy for re-classification. 3. ExAC is not overrepresented for pathogenic variants in *MLH1*, *MSH2*, and *MSH6* genes at a frequency above the reported population incidence of Lynch syndrome (1/500). 4. Carrier frequencies of the most frequent pathogenic variants in ExAC were concordant with reported population frequencies for *BRCA1&2* specific AJ-founder mutations (1/756), and *CFTR* p.F508del (1/71). 5. Distribution of the most frequent variants in *CFTR* (p.F508del)*, GJB2* (p.V37I)*,* and *HBB* (p.E7K and p.E7V) were concordant with the reported ethnic prevalence in European, East Asian, and African respectively. **CONCLUSIONS**: The ExAC database is not overrepresented for pathogenic variants in the genes evaluated. Although supportive of ExAC as a control cohort for classifying variants in clinical settings, we recommend that labs evaluate this database mindful of the mutational spectrum, presence of pseudogenes (eg *PMS2*), and heterogeneity of genes analyzed, but not as a sole evidence for variant classification. As evidenced with cardiac and Lynch syndrome gene variants, underlying genetic complexities and lack of published knowledge can pose challenges in deriving meaningful classifications using control datasets.

206

Assessing the clinical validity of genes implicated in hereditary pheochromocytoma/paraganglioma and pancreatic cancer using the ClinGen framework. *R. Ghosh[1], A. Buchanan[2], N.T. Strande[3], E.R. Riggs[2], S.S. Dwight[7], T.P. Sneddon[7], C.L. Martin[2], J.S. Berg[3], M.J. Ferber[4], K. Offit[5], K.L. Nathanson[6], S.E. Plon[1].* 1) Pediatrics-Oncology, Baylor College of Medicine, Houston, TX; 2) Geisinger Health System, Danville, PA; 3) University of North Carolina, Chapel Hill, NC; 4) Mayo Clinic, Rochester, MN; 5) Memorial Sloan Kettering Cancer Center, New York, NY; 6) University of Pennsylvania, Philadelphia, PA; 7) Stanford University School of Medicine, Stanford, CA.

Advances in genomics have led to a marked increase in reported number of gene-disease associations, and an increasing number of genes on clinical genetic testing panels. There is a critical need for developing standards for determining the clinical validity and assessing the strength of evidence of a given gene-disease association. The Clinical Genome Resource (ClinGen) is a NIH-funded program dedicated to creating an open and centralized resource of clinically relevant genes and variants, evaluated using standardized guidelines to optimize their clinical integration. As part of this effort, ClinGen's Gene Curation Working Group has 1) developed a framework to curate publicly available evidence for gene - monogenic disorder associations, and 2) established the following clinical validity classification scheme: Definitive, Strong, Moderate, Limited, Disputed, Evidence Against, and No Evidence. A given gene-disease pair is assigned to one of these classes based on the strength of the available clinical, functional and contradictory evidence, and whether the initial report of disease-gene association has been replicated.ClinGen's Hereditary Cancer Clinical Domain Working Group has used this framework to determine the clinical validity of genes implicated in predisposition to pheochromocytomas /paragangliomas (PCC/PGLs) and pancreatic cancer, tumors for which multi-gene panels are clinically available. PCC/PGLs are rare neuroendocrine tumors that have a high likelihood of resulting from genetic susceptibility. Following the ClinGen gene curation process we identified that only nine of 21 genes included on PCC/PGL multi-gene panels scored as Definitive evidence for an association with disease. One gene, *HIF2A,* had significant Evidence Against an association with PGL based on the population frequency of the single variant associated with PGL. Similarly, only one of 10 genes commonly included on multi-gene pancreatic cancer panels was found to have a Definitive association with pancreatic cancer susceptibility. These results illustrate the importance of establishing the clinical validity of genes included on clinically available hereditary cancer gene panels. This systematic approach to ascertain clinical validity will provide a powerful resource to inform clinicians of the significance of a given test results with regard to the strength of a gene-disease association as well as defining areas for additional research in clinical genetics.

207

Assessment of mendelian disorders among three Bronx, New York populations using ACMG criteria. *G. diSibio[1, 2], K. Upadhyay[1], P. Meyer[1], B. Baskovitch[3], C. Oddoux[2], H. Ostrer[1].* 1) Albert Einstein College of Medicine of Yeshiva University, Bronx, NY; 2) Montefiore Medical Center, Bronx, NY; 3) 2451 Fillingim Street, Mobile, AL.

Introduction:Screening genomes from healthy members of a population against high density SNP panels allows frequency estimates for potentially deleterious variants that can guide design of genetic disease screens and increase physician-patient awareness for reproductive and personal health decisions. We have followed this approach to identify and annotate, using recent ACMG guidelines,variants in Puerto Rican, African American, and Dominican American populations from Bronx, NY. Methods:Genomic DNAs from 192 individuals from each of the above three populations were hybridized to Affymetrix Axiom Exome 319 chips; pooled DNAs from 100 Dominicans were additionally whole exome sequenced. Resultant variants identified by Genotyping Console software were parsed through a bioinformatics pipeline that identified pathogenicity against OMIM/ClinVar databases and then manually curated by $\geq$ 2 reviewers to confirm *in silico* assessments. Resulting variants were further refined to generate a final list of those that 1) had an MAF of between 0.001 and 0.1 (for AR) or 0.02 (for AD), 2) had clinical relevance, and 3) fell into an ACMG-defined pathogenic category. Results:Of the 305,519 variants identified across all three self-identified populations, 1,440 were called “pathogenic” or “possibly pathogenic” *in silico*. Manual curation confirmed these assignments for 338 variants. 16%, 30%, and 16% of variants were unique to Puerto Ricans, African Americans, and Dominicans, respectively; 16% of variants were shared between Puerto Ricans and African Americans; 6% between African Americans and Dominicans; 4% between Dominicans and Puerto Ricans; and 13% among all three populations. 47 variants met the three criteria above. Neurologic, cardiovascular, hematologic, developmental, and metabolic conditions were among those identified, with autosomal recessive (75%), autosomal dominant (10%), AD/AR (6%), AR/digenic (4%), or complex (4%) inheritance. AR MAF’s ranged from 0.009 (Type Ia Congenital Disorder of Glycosylation) to .036 (Type 1 Hemochromatosis). 28% of identified mutations encoded truncations or splice-site disruptions; the remainder fit $\geq$ 2 other ACMG Pathogenic categories. Conclusions:We provide a refined list of pathogenic DNA variants, identified in three urban populations, annotated using ACMG criteria. This work provides a model for developing medically relevant genetic panels for patients and their health care providers within these populations.

208

Frequency of Cardiovascular Secondary Findings on Whole-Exome Sequencing and Utilization in Familial Testing. *R. Tousignant, A.A. Singleton, B. Friedman, K. Retterer, G. Richard, D. Macaya.* GeneDx, Gaithersburg, MD.

Whole exome sequencing (WES) allows for a comprehensive evaluation of the underlying genetic causes of disease and has become a recognized diagnostic tool in clinical practice. WES has the potential to identify variants in genes unrelated to the primary phenotype, known as incidental findings (IFs). In 2013, the American College of Medical Genetics and Genomics (ACMG) recommended the evaluation and reporting of IFs, now termed secondary findings (SFs), in 56 genes associated with well-known and medically actionable disorders. The identification of SFs will affect patient management beyond the primary phenotype and will impact other family members. Half of the genes on this list are associated with inherited cardiovascular (CV) disorders with variable age of onset and incomplete penetrance. The goal of this study was to determine the yield of cardiovascular SFs among probands who opted to receive SFs, and whether these SFs were used to identify other family members at risk of developing CV disorders. WES results from 4290 families tested at GeneDx were analyzed. The number of probands requesting ACMG secondary findings was 3799 (89%). Cardiovascular SFs were identified in 90 probands (2.4%). Subsequent carrier testing for known pathogenic (KP) variants was performed in 22 of these families (24%) and 27 individuals were identified as harboring a KP variant, thus putting them at risk of developing cardiomyopathy, arrhythmia or thoracic aortic aneurysm/dissection. Additionally, we reported KP variants in 10/3799 (~0.3%) probands in 7 genes associated with medically actionable or severe cardiac phenotypes that are not on the ACMG list (ANK2, CACNA1C, ENG, HRAS, KCNE1, KCNE2 and TTN). The identification of KP variants in these genes highlights the need for continued discussion and evaluation of genes not currently on the ACMG list. These preliminary findings suggest that SFs associated with CV disorders currently prompt targeted testing in only one-quarter of families and it rarely extends beyond the siblings of a proband. In at least two families, siblings were tested for the variants associated with the proband’s primary phenotype but not for the cardiovascular SF. This underscores the need for a systematic review of the utilization of SFs by ordering providers to assess not only the uptake of targeted familial testing but also whether the identification of SFs prompts appropriate clinical follow-up and specialist referrals.

209

Homozygous and compound heterozygous mutations in *FBN1*, unusual situations in molecular diagnosis of Marfan syndrome. *P. Arnaud[1,2], N. Hanna[1,2], M. Aubart[2], B. Leheup[3], S. Dupuis-Girod[4], M.-A. Delrue[5], D. Lacombe[5], O. Milleron[6], M. Langeois[6], M. Spentchian[6], L. Gouya[6], G. Jondeau[2,6], C. Boileau[1,2].* 1) Département de Génétique, Hôpital Bichat, AP-HP, PARIS, France; 2) INSERM UMR_S 1148, Laboratory for Vascular Translational Science, Hôpital Bichat, PARIS, France; 3) Service de Génétique Clinique, CHU Nancy, Hôpital de Brabois, VANDOEUVRE-LES-NANCY, France; 4) Service de Génétique Clinique, Hospices Civils de Lyon, Hôpital Femme-Mère-Enfant, Groupe Hospitalier Est, BRON, France; 5) Service de Génétique Médicale, CHU de Bordeaux-GH Pellegrin, BORDEAUX, France; 6) Centre National Maladies Rares, Syndrome de Marfan et apparentés, Hôpital Bichat, PARIS, France.

Marfan syndrome (MFS, [MIM#154700]) is an inherited autosomal dominant disorder, with an incidence of 1 in 5000. This disease affects different systems, including cardiovascular, ocular and skeletal. Cardiovascular manifestations with aortic aneurysm or dissection are the most serious life-threatening complications of the syndrome. Heterozygous mutations in the *FBN1* gene, encoding fibrillin-1, are the main cause of MFS. These mutations are located throughout the gene without phenotypic association, with the exception of mutations in severe neonatal MFS cases that cluster in exons 24 to 32. Four cases of homozygosity and three cases of compound heterozygosity in the *FBN1* gene, all associated with severe clinical signs are found in the literature. Here, we report 8 new cases of homozygous and compound heterozygous mutations in 8 French families. Patients were part of the 2400 consecutive probands referred nationwide to our Centre for molecular diagnosis of MFS. Systematic bidirectional sequencing of the 65 exons of the *FBN1* gene was performed and biparental origin of mutations was confirmed when possible. Three probands carrying homozygous mutations (c.2513T>C – p.Leu838Ser, c.6998A>G – p.Asp2333Gly and c.7999G>A – p.Glu2667Lys) were identified. They belong to 3 consanguineous families of Northern African ancestry. Taken together, our results and published results represent 7 different missense homozygous mutations that strikingly cluster at the 3' end of the *FBN1* gene (between exons 57 and 63). In parallel, 21 probands carried two mutations and unequivocal compound heterozygosity could be ascertained by family studies in 5 of them. Complete clinical features as listed in the revised Ghent nosology for MFS were available for the 8 French probands reported here (mean age at discovery: 30 y.o; ages ranging from 8 to 53 y.o). All displayed classic manifestations of the syndrome. None presented extremely severe manifestations of MFS in any system compared to carriers of only one mutated *FBN1* allele. This observation is not in keeping with the very severe clinical features reported in the literature for 4 homozygous and 3 compound heterozygous probands. Therefore, there is a large spectrum of severity of the disease in probands carrying 2 mutated *FBN1* alleles. Finally, although homozygosity and compound heterozygosity are rarely found in molecular diagnosis of MFS, they should not be overlooked and no predictive evaluation of severity should be provided.

210

Clinician perspectives on inconclusive genetic test results for osteogenesis imperfecta in children with unexplained fractures: are families at risk if they engage in parental testing for VUS? *E. Youngblom[1], M.L. Murray[2], P.H. Byers[2].* 1) Institute of Public Health Genetics, University of Washington, Seattle, WA; 2) Departments of Pathology and Medicine (Medical Genetics), University of Washington, Seattle, WA.

Genetic testing can identify children with osteogenesis imperfecta (OI) among those evaluated for non-accidental injury (NAI). In this setting, testing of parents to identify the origin of variants of uncertain significance (VUS) can alter the interpretation of the initial result in ways that influence further treatment of the child and increase the likelihood that parents are subject to legal risk. To understand how inconclusive results for OI are interpreted and used by medical practitioners in the context of perceived NAI, a 15-question survey was sent to physicians who had requested testing for OI (analysis of the genes *COL1A1* and *COL1A2*). The survey gathered information about clinical experiences, policies, and follow-up procedures for VUS test results. The participants were gathered from individuals who requested the tests at the Collagen Diagnostic Laboratory, University of Washington from 2005-2013. 89 out of 292 (27%) eligible participants responded, and all but one were geneticists. Participants saw an average of 6-7 patients per year in which the differential diagnosis included OI and NAI, 36% of which were estimated to be legal cases. The factor that most influenced physicians' decisions to refer a patient for OI testing was the presence of clinical features of OI (85% of the time). The most common reasons that follow-up studies on VUS would not be carried out were: financial (63%), lack of access to child (32%), and family request (20%). When parents received a VUS result, the most frequent reaction was confusion (72%). Only 20% of respondents indicated that their clinic recontacts patients' families if their VUS is later reclassified as benign or pathogenic. The context in which a VUS is observed may lead to different steps to determine if it could have clinical meaning. When identified in a family member with a known genetic condition, testing of other family members can lead to reclassification of the variant. In an abuse investigation, the study of other family members may decrease the likelihood that the VUS contributes to the fracture and change the evaluation of the child and the parents. On the basis of the findings in this study, there are no clear guidelines for how to interpret a VUS result and what procedures should be followed to determine if they could have clinical relevance. In the legal setting, this class of results has ramifications for family members and careful guidance may be necessary to avoid unintended consequences.

211

Biallelic loss of human *CTNNA2*, encoding α-N-catenin, links ARP2/3-mediated actin regulation to neuronal migration. *A.E. Schaffer[1,2], A.O. Caglayan[3], N. Al-Sanaa[4], H.Y. Al-Abdulwahed[4], R.O. Rosti[1], B. Copeland[1], S.T. Baek[1], E. Scott[1], M.S. Zaki[5], G.M.H. Abdel-Salam[5], T. Ben-Omran[6], A. Karimenejad[7], H. Kayserili[8], F. Mojahedi[9], M. Kara[10], N. Cai[1], J. Silhavy[1], E. Yosunkaya[11], B.A. Barshop[12], B. Kara[13], R. Nachnani[1], H. Megahed[5], F. Incecik[14], S. Danda[15], I. Miller[16], W.B. Dobyns[17], S. Gabriel[18], K. Bilguvar[3], M. Gunel[3], J.G. Gleeson[1,2].* 1) Laboratory for Pediatric Brain Disease, Howard Hughes Medical Institute, Rockefeller University, NY, USA; 2) Department of Neuroscience, Howard Hughes Medical Institute, University of California, San Diego, La Jolla, CA, USA; 3) Department of Neurosurgery, Neurobiology, and Genetics, Yale University School of Medicine, New Haven, CT, USA; 4) Department of Pediatrics, Dhahran Health Center, Saudi Aramco Corporation, Dhahran, Kingdom of Saudi Arabia; 5) Clinical Genetics Department, Human Genetics and Genome Research Division, National Research Centre, Cairo, Egypt; 6) Clinical and Metabolic Genetics Division, Department of Pediatrics, Hamad Medical Corporation, Doha, Qatar; 7) Karimenejad-Najmabadi Pathology and Genetic Center, Tehran, Iran; 8) Medical Genetics Department, Istanbul Medical Faculty, Istanbul University, Millet Caddesi, Fatih/Istanbul, Turkey; 9) Mashhad Medical Genetic Counseling Center, Mashhad, Iran; 10) Department of Pediatrics, Tripoli Children's Hospital, Tripoli, Libya; 11) Department of Medical Genetics, Cerrahpasa School of Medicine, Istanbul University, Istanbul, Turkey; 12) University of California, San Diego Department of Biochemical Genetics, Rady Children's Hospital, San Diego, CA, USA; 13) Kocaeli University, Medical Faculty, Department of Pediatric Neurology, Umuttepe, Kocaeli, Turkey; 14) Department of Pediatric Neurology, «ukurova University Medical Faculty, Balcali, Adana, Turkey; 15) Department of Clinical Genetics, Christian Medical College and Hospital, Vellore, Tamil Nadu, India; 16) Neurology Department, Miami Children's Hospital, Miami, FL, USA; 17) Seattle Children's Research Institute, Centre for Integrative Brain Research, Seattle, WA, USA; 18) Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA.

Neuronal migration defects (NMDs), including lissencephaly, represent the most severe developmental brain defects in humans. We studied a cohort of 107 families with NMDs by exome sequencing, identifying two homozygous C-terminal truncating mutations in *CTNNA2*. *CTNNA2* encodes α-N-catenin, one of three paralogues of the α-catenin family, involved in epithelial integrity and cell polarity. One of these paralogs, α-E-catenin, has been well studied and shown to function by regulating Wnt signaling, binding directly to actin, and by repressing ARP2/3-mediated actin branching in epithelial tissues; however, controversy exists as to whether α-N-catenin has similar roles. α-N-catenin was strongly expressed in developing mouse and human cerebral cortex and loss-of-function mutations in CRISPR/Cas9-edited and patient-derived neural cells led to failed neurite stability and severe migration defects. Wnt target gene expression was not altered in *CTNNA2*-mutant cells; even in the presence of exogenous Wnt ligand. Using a gene replacement stratagy, we determined the neural migration defect was dependent on the presence of the putative F-actin binding domain of α-N-catenin. Moreover, we found recombinant α-N-catenin was sufficient to bind and bundle purified F-actin as well as repress ARP2/3-mediated actin polymerization. ARP2/3 association with F-actin was increased in patient-derived neurons and small molecule inhibition of ARP2/3 activity in *CTNNA2*-mutant cells was sufficient to restore neurite stability and migration. We thus identify *CTNNA2* as the first catenin family member with bi-allelic mutations in human, critical for brain development.

212

Missense mutations in the middle domain of DNM1L cause infantile encephalopathy in humans and peroxisomal and mitochondrial defects in *Drosophila* and humans. *L. Robak[1], Y. Chao[1], F. Xia[1], M. Koenig[2], C. Bacino[1], F. Scaglia[1], M. Wangler[1].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) University of Texas Medical School at Houston, Department of Pediatric Neurology, Houston TX.

DNM1L is a gene previously linked to a lethal defect of mitochondrial and peroxisomal fission in one patient with a *de novo* p.A395D pathogenic variant with biochemical evidence of mitochondrial and peroxisomal defects. No further cases have been reported. We identified Patient 1, a 5 year-old male with global developmental delay (GDD), status epilepticus, and progressive volume loss on brain MRI. WES revealed a VUS in *DNM1L*, c.1048G>A, p.G350R. Very long chain fatty acids were within normal limits. Because his phenotype differed from that previously reported, but the *DNM1L* variant changed a highly conserved amino acid, we considered this a phenotypic expansion. In order to test this hypothesis, we studied the function of this variant in *Drosophila* homologue *drp1*. We cloned the human wild-type *DNM1L* cDNA and constructs with the p.G350R and p.A395D variants for expression in *Drosophila* neurons both in *Drosophila drp1*-/- and *drp1*-/+ animals. Our analysis revealed rescue of *Drosophila drp1* mutants by expression of the human *DNM1L* gene. In contrast, the brains contained aggregated mitochondria with trafficking defects due to expression of the p.A395D and the p.G350R pathogenic variants. Subsequently we identified another case. Patient 2 is a 10 month-old female with lactic acidosis, diffuse hypotonia, GDD, poor growth and agenesis of the corpus callosum. Global metabolomic testing revealed mild elevations in peroxisomal lipids. WES revealed two *de novo* changes in mitochondrial-related genes: a *de novo* VUS in the *PDHA*1 gene (c.448G>A, p.G150R), known to be associated with pyruvate dehydrogenase E1 deficiency and a *de novo* VUS in the *DNM1L* gene (c.1135G>A, p.E379K). This raised the possibility that one or both of these variants were contributing to the patient's phenotype. A comparison of Patient 1 and Patient 2 suggests the *DNM1L* variant is pathogenic in Patient 2. Both patient 1 and patient 2 have pathogenic variants affecting the middle domain of DNM1L, an important domain for complex assembly associated with organelle fission. Our work combining detailed clinical evaluation, metabolomics and WES in patients with rare mitochondrial phenotypes alongside *Drosophila* mitochondrial functional studies allows for the elucidation of the effect of specific human variants on organelle dynamics and aid in clinical WES interpretation.

213

Mosaic and constitutional mutations of *MTOR* cause a spectrum of developmental brain disorders from focal cortical dysplasia to diffuse megalencephaly. *G. Mirzaa[1], C. Campbell[2], N. Solovieff[2], L. Jansen[3], A. Timms[4], V. Conti[5], C. Adasm[1], E. Boyle[6], S. Collins[1], G. Ishak[7], S. Poliachik[7], S. Gunter[3], R. Leary[2], S. Mahan[2], M. Doerschner[8], S. Jhangiani[9, 10], D. Muzny[9], E. Boerwinkle[10, 11], R. Gibbs[9, 10], J. Lupski[9, 10, 12, 13], J. Shendure[14], R. Saneto[15], E. Novotny[15], W. Sellers[2], L. Murphy[2], M. Morrissey[2], J. Ojemann[16], R. Guerrini[5], W. Winckler[2], W. Dobyns[1].* 1) Human Genetics, Seattle Children's Research Institute, Seattle, WA; 2) Novartis Institutes for Biomedical Research Inc., Cambridge, MA; 3) University of Virginia, Neurology, Charlottesville, VA, USA; 4) Center for Developmental Biology and Regenerative, Medicine, Seattle Children's Research Institute, Seattle, Washington, USA; 5) Pediatric Neurology, Neurogenetics and Neurobiology Unit and Laboratories, A. Meyer Children's Hospital-University of Florence, Florence, Italy; 6) Department of Genetics, Stanford University School of Medicine, Stanford, California, USA; 7) Department of Radiology, Seattle Children's Hospital, Seattle, Washington, USA; 8) Department of Pathology, University of Washington, Seattle, Washington, USA; 9) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA; 10) Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA; 11) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, USA; 12) Department of Pediatrics, Baylor College of Medicine, Houston, Texas, USA; 13) Texas Children's Hospital, Houston, Texas, USA; 14) Department of Genome Sciences, University of Washington, Seattle, Washington, USA; 15) Division of Pediatric Neurology, University of Washington, School of Medicine, Seattle, Washington, USA; 16) Department of Neurosurgery, University of Washington, Seattle, Washington, USA.

Focal cortical dysplasia (FCD), hemimegalencephaly (HMEG) and megalencephaly constitute a spectrum of malformations of cortical development with shared neuropathologic features. Collectively, these disorders are associated with significant childhood morbidity and mortality. FCD, in particular, represents the most frequent cause of intractable focal epilepsy in children. Multiple lines of evidence have demonstrated functional defects in the phosphatidylinositol 3-kinase (PI3K)-AKT-MTOR pathway by western blot, immunohistochemistry, AKT kinase and RNA expression analyses, suggesting that detailed molecular studies are needed in this group of disorders.We performed whole exome sequencing (WES) on eight children with FCD or HMEG using standard depth (~50-60X) sequencing in peripheral samples (blood, saliva or skin) from the affected child and their parents, and deep (~150-180X) sequencing in affected brain tissues. We also used both targeted sequencing or WES to screen a cohort of 105 children with molecularly unexplained diffuse or focal brain overgrowth (62 with FCD-HMEG, and 43 with diffuse megalencephaly). Histopathological and functional assays of PI3K-AKT-MTOR pathway activity in resected brain tissue with FCD were performed to validate mutations. We identified low-level mosaic mutations of *MTOR* in brain tissues in four children with FCD type 2a (with alternative allele fractions ranging from 0.012–0.086). Molecular and functional analysis in two children with FCD type 2a from whom multiple affected brain tissue samples were available revealed a gradient of alternate allele fractions with an epicenter in the most epileptogenic area. We also identified an intermediate level mosaic mutation of *MTOR* (p.Thr1977Ile) in three unrelated children with diffuse megalencephaly and cutaneous pigmentary mosaicism (alternative allele fractions 0.07–0.23). Finally, we identified a constitutional *de novo* mutation of *MTOR* (p.Glu1799Lys) in three unrelated children with diffuse megalencephaly and intellectual disability. Our data show that mutations of *MTOR* are associated with a spectrum of brain overgrowth phenotypes extending from FCD type 2 to diffuse megalencephaly, distinguished by different mutations and levels of mosaicism. Our data also show the first compelling demonstration of the pattern of mosaicism in brain in FCD, and substantiate the link between mosaic mutations of *MTOR* and pigmentary mosaicism in skin (sometimes designated "hypomelanosis of Ito").

214

***MTIF2* mutations cause a novel disorder of mitochondrial translation.** *S.B. Pierce[1], R. Ganetzky[2,3], J.A. Foster[4], D. Xu[4], S. Wakefield[4], N. Sondheimer[2], S.P. Yang[4].* 1) Department of Medicine (Medical Genetics), University of Washington, Seattle, WA; 2) Section of Biochemical Genetics, Children's Hospital of Philadelphia, Philadelphia PA; 3) Division of Medical Genetics, Children's Hospital of Philadelphia, Philadelphia PA; 4) Department of Medical Genetics, Shodair Children's Hospital, Helena MT.

Mutations in nuclear genes required for translation of mitochondrial proteins can lead to oxidative phosphorylation deficiencies. Clinical features may include encephalopathy, developmental delay, congenital anomalies, and endocrine dysfunction, with a wide range of severity. Primary ovarian failure is an uncommon feature of oxidative phosphorylation deficiency, but is part of Perrault syndrome, which can be caused by mutations in genes involved in mitochondrial translation (*HARS2* [MIM 600783] and *LARS2* [MIM 604544]) and mitochondrial DNA replication (*C10orf2* [MIM 606075]), as well as in the mitochondrial protease gene *CLPP* (MIM 601119). Mitochondrial translation initiation factor 2 (MTIF2 [MIM 603766]) is essential for mitochondrial translation. It serves the role of its prokaryotic homologue, prokaryotic initiation factor 2 (pIF2), and has a 37 amino acid domain that provides the function of prokaryotic initiation factor 1 (pIF1), for which there is no mitochondrial homologue. We evaluated two siblings with clinical features consistent with mitochondrial dysfunction. The 18-year-old female proband presented with microcephaly, moderate developmental delay, partially absent septum pellucidum, inverted nipples, unusual fat distribution, lactic acidosis, and primary amenorrhea. Her 3-year-old brother presented with microcephaly, severe developmental delay, partially absent septum pellucidum, inverted nipples, 2-3 toe syndactyly, short stature, and lactic acidosis. Analysis was performed by whole exome sequencing. The siblings were compound heterozygous for MTIF2 p.H191Q (c.573C>G, maternal) and MTIF2 p.V300Afs*9 (c.899_909del, paternal). *MTIF2* c.573C>G was shown experimentally to increase exon 8 skipping, leading to a frameshift and premature stop; p.H191Q was also predicted to damage protein function. *MTIF2* c.899_909del was shown experimentally to lead to nonsense-mediated decay. In patient fibroblasts, *MTIF2* transcript expression was reduced more than 75% and MTIF2 protein was undetectable. Pulse labeling of mitochondrial proteins in patient fibroblasts with [^{35}S]-methionine/cysteine indicated that translation of all mitochondrial proteins was decreased. These results suggest that mutations in *MTIF2* impair mitochondrial translation, leading to phenotypes that include microcephaly and developmental delay, and ovarian failure in females.

215

Overexpression of the chromosome 21 gene *ATP5O* results in enteric hypoganglionosis: the missing link between Down Syndrome and Hirschsprung disease? *R.K. Chauhan[1], R. Lasabuda[1], Z. Azmani[1], H.C. van der Linde[1], A.S. Brooks[1], S. Edie[2], R.H. Reeves[2], A.J. Burns[1,4], I.T. Shepherd[3], R.M.W. Hofstra[1,4].* 1) Department of Clinical Genetics, Erasmus MC, Rotterdam, Netherlands; 2) Johns Hopkins University School. of Medicine, Department of Physiology and McKusick-Nathans Institute for Genetic Medicine, Baltimore, USA; 3) Department of Biology, Emory University, Atlanta, USA; 4) Birth Defects Research Centre, UCL Institute of Child Health, London, United Kingdom.

Hirschsprung disease (HSCR) is characterized by the absence of enteric ganglia in a variable length of the gastrointestinal tract, leading to severe intestinal obstruction. Around 12% of individuals with HSCR have a chromosomal abnormality and the most common live born one is trisomy 21, leading to Down Syndrome (DS). As individuals with DS have a 40-fold higher risk of developing HSCR than people in the general population, human chromosome 21 (Hsa21) genes may be involved in the etiology of HSCR.To identify genes contributing to HSCR phenotype in DS , we used zebrafish with the reporter transgene, *Tg (-8.3bphox2b:kaede)* to assay the potential candidate genes and look for ENS phenotypes. This reporter line expresses the fluorescent kaede protein in the enteric neuron precursor cells, making it easy to visualize and quantify the number of enteric neurons of the zebrafish gut *in vivo*. we prioritized 28 genes of Hsa21 for overexpression in zebrafish, based on the expression data generated from mouse enteric neural crest stem cells (ENCSCs) and literature survey. To overexpress candidate genes, we micro-injected capped mRNAs of candidate Hsa21 genes at the single-cell stage. Embryos were maintained and scored for enteric nervous system (ENS) defects and abnormal phenotypes at 5 days post fertilization. Expression of mRNAs from the 28 Hsa21 genes induced various phenotypic defects in the zebrafish model. Of note, we showed that overexpression of *ATP50 (*ATP synthase, H^+ transporting, mitochondrial F1 complex, O subunit*)* led to enteric hypoganglionosis. The protein encoded by this gene is a component of the F-type ATPase found in the mitochondrial matrix and participates in ATP synthesis coupled proton transport. *ATP50* is highly expressed in mouse ENCSCs that give rise to mature neurons and glial cells within the gut. Hypoganlionosis observed in the zebrafish is a strong indication that *ATP50* overexpression in DS affected individuals could contribute to their HSCR phenotype.These studies therefore identified a Hsa21 gene that can (partly) explain the association between DS and HSCR. The underlying mechanisms by which enteric neural crest stem cells are affected by *ATP5O* upregulation at an early developmental stage, leading to neuronal loss is currently being further investigated.

216

Mutations in *PPP2R5D* are a novel cause of intellectual disability, macrocephaly, hypotonia, and autism. *L.B. Henderson[1], L. Shang[2], M.T. Cho[1], C.T. Fong[3], K.M. Haude[3], N. Shur[4], J. Lundburg[4], N. Hauser[5], J. Carmichael[5], J. Innis[6,7], J. Schuette[6,7], Y.W. Wu[8], S. Asaikar[9], M. Pearson[10], L. Folk[1], K. Retterer[1], K.G. Monaghan[1], W.K. Chung[2,11].* 1) GeneDx, Gaithersburg, MD; 2) Department of Pediatrics, Columbia University Medical Center, New York, NY; 3) University of Rochester Medical Center, Rochester, NY; 4) Albany Medical Center, Albany, NY; 5) Valley Children's Hospital, Madera, CA; 6) Division of Pediatric Genetics, University of Michigan Health System, Ann Arbor, MI; 7) Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI; 8) Departments of Neurology and Pediatrics, University of California San Francisco, San Francisco, CA; 9) Child and Adolescent Neurology Consultants, Sacramento, CA; 10) District Medical Group, Scottsdale, AZ; 11) Department of Medicine, Columbia University Medical Center, New York, NY.

Identifying the etiology of intellectual disability (ID) and autism spectrum disorder (ASD) poses a challenge due to the heterogeneity of these disorders. Whole exome sequencing (WES) provides an effective strategy to identify the molecular cause of disease, which in many cases has not been identifiable by multiple other genetic tests. Furthermore, trio-based WES affords the opportunity to detect *de novo* mutations, which account for a significant portion of ID and ASD due to reduced genetic fitness of affected individuals. Using clinical WES, we identified *de novo* variants in the *PPP2R5D* gene in seven individuals with shared clinical characteristics of global developmental delay and ID, macrocephaly, and hypotonia. Most of these patients also had ASD, and additional features present in some individuals included broad-based gait, congenital heart defects, dysmorphic features, seizures, scoliosis, and short stature. *PPP2R5D* encodes a regulatory B-type subunit of protein phosphatase 2A (PP2A) involved in regulating tau phosphorylation and other key neuronal processes. Four distinct missense variants were identified, two of which were recurrent. Each of these *de novo* variants alters a highly conserved glutamic acid residue to lysine, likely resulting in a pathogenic gain of PP2A function. Our findings implicate *PPP2R5D* as a novel cause of ID, macrocephaly, hypotonia, and ASD.

217

To elucidate the genetic of recessive cognitive disorders: 104 novel genes identified using deep sequencing. *H. Najmabadi[1], H. Hu[2], Z. Fattahi[1], L. Musante[2], S.S. Abedini[1], M. Hosseini[1], F. Larti[1], M. Mohseni[1], P. Jamali[3], M. Beheshtian[1], F. Mojahedi[4], T.F. Wienker[2], K. Kahrizi[1], H.H. Ropers[2].* 1) Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran; 2) Department Human Molecular Genetics, Max Planck Institute for Molecular Genetics, Berlin, Germany; 3) Shahroud Welfare Organization, Semnan, Iran; 4) Mashhad Medical Genetic Counseling Center, Mashhad, Iran.

Intellectual disability (ID), the widespread impairment with an enormous socio-economic burden, affects 2-3% of the overall population. Genetic causes of ID comprise a high spectrum of molecular mechanisms which may vary from point mutations to large cytogenetic abnormalities. It is believed that about 13-24% of the patients in Western populations are due to recessive forms of ID which is highly heterogeneous. It is has been proposed that more than 2500 genes are implicated in autosomal ID and so far, more than 650 genes have been identified for autosomal recessive ID. We applied combination of exome and whole genome sequencing on 420 consanguine families mostly Iranian from different ethnicities with two and more affected. We were able to identify 104 novel ARID genes in 111 families (27%) of our subjects, and eighty nine known ARID genes in 114 (28%) families. Majority of our novel genes contribute to ARID with additional features. This project revealed a novel ciliopathy gene in two families. Twenty out of 104 novel genes were involved in developmental function. Of 104 novel ID genes, 34 genes play roles in cell process. Thirteen out of 104 novel genes contribute to metabolic function. Of 104 novel ID genes, we found seven novel genes involved in response to DNA damage and DNA repair function. 5 out of 104 our novel ID genes play roles in synaptic function. Of 104 novel ID genes, 5 genes were implicated in RNA processing that encodes a mitochondrial tryptophanyl-tRNA synthetase. In this cohort, 9 out of 104 novel genes implicated in regulation and transcription functions. The exact function of 10 out of 104 novel genes has not been identified yet.

218

Variants in *TAF1* are associated with a new syndrome with severe intellectual disability and characteristic dysmorphic features. *G.J. Lyon[1,2], J.A. O'Rawe[1,2], Y. Wu[1,2], A. Rope[3], P.Y. Au[4], K. Kosma[5], C. Smith[4], S. Kitsiou-Tzeli[5], J. Schuette[6,7], F. Martinez[8], C. Orellana[8], M. Rosello[8], S. Oltra[8], A. Caro-Llopis[8], L. Jimenez Barrón[1,9], J. Swensen[10], H. Fang[1], D. Mittelman[11], C. Keegan[6,7], R. Robison[12], E. Yang[13], J. Parboosingh[4], K. Wang[14], J. Parboosingh[6], V. Kalscheuer[15], M. Hammer[16], M. Kousi[17], E. Davis[17], N. Katsanis[17], E. Wang[18].* 1) Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, USA; 2) Graduate Program in Genetics, Stony Brook University, Stony Brook, NY, USA;; 3) Department of Medical Genetics, Northwest Kaiser Permanente, Portland, OR, USA;; 4) Department of Medical Genetics and Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, Canada; 5) Department of Medical Genetics, Medical School, University of Athens, and Research Institute for the Study of Genetic and Malignant Disorders in Childhood, Aghia Sophia, Children's Hospital, Athens, Greece; 6) Dept. of Pediatrics, Division of Genetics, University of Michigan, Ann Arbor, MI, USA; 7) Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA; 8) Unidad de Genetica. Hospital Universitario y Politecnico La Fe, Valencia, Spain; 9) Centro de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Morelos, MX; 10) Caris Life Sciences, Phoenix, Arizona, USA; 11) Gene by Gene, Ltd., Houston, TX, USA; 12) Utah Foundation for Biomedical Research, Salt Lake City, UT, USA; 13) Department of Radiology, Boston Children's Hospital, Boston, MA, USA; 14) Zilkha Neurogenetic Institute, Department of Psychiatry and Preventive Medicine, University of Southern California, Los Angeles, CA, USA; 15) Department of Human Molecular Genetics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany; 16) Division of Biotechnology, University of Arizona, Tucson, AZ, USA; 17) Depts of Cell Biology and Pediatrics, Duke University, North Carolina, USA; 18) Department of Pharmacology, University of Washington, Seattle, WA, USA.

We describe the discovery of a new X-linked genetic syndrome, driven initially by a whole genome sequencing study for one Caucasian family from Utah with two affected male brothers, presenting with severe intellectual disability (ID), a characteristic caudal prominence, and very distinctive facial features, including a broad, upturned nose, sagging cheeks, downslanted palpebral fissures, prominent periorbital ridges, deeply set eyes, relative hypertelorism, thin upper lip, a high palate, prominent ears with thickened helices, and a pointed chin. Illumina-based whole genome sequencing (WGS) was performed on 10 members of this family, with additional Complete Genomics-based WGS performed on the mother, father and two affected sons. The boys carry a maternally inherited missense variant in the X-chromosomal gene *TAF1*, which encodes the largest subunit of the general transcription factor IID (TFIID) multi-protein complex. Simultaneous studies using diverse strategies led to >5 other families with de novo or maternally inherited variants in *TAF1* and with a remarkably similar clinical presentation. All the variants are novel and the majority of the variants are missense however, one introduces a leaky cryptic splice site leading to a prematurely truncated protein and one is a large 0.42 Mb duplication that includes *TAF1*. A recent population-scale study also reported *TAF1* as being ranked 53rd among the top 1,003 constrained human genes, and the identified variants fall in regions of TAF1 that are under-represented in population-wide sequencing of "normal controls". We are currently undertaking functional studies in zebrafish, and we are also conducting cellular complementation assays with a ts13 mutant cell line that carries a mutation in *TAF1*, causing them to arrest in the late G1 phase of the cell cycle at the non-permissive temperature of 39.5°C. The cell cycle defect of ts13 cells can be fully complemented by expression of full-length WT TAF1, and we are currently testing the mutated versions of *TAF1*. To investigate how the *TAF1* variants identified in the above families may influence protein structure and packaging, we built a structure model for the region of residues 1080 to 1569 using I-TASSER. Our results implicate mutations in TAF1 as playing a critical role in the development of this new intellectual disability syndrome.

219

A Pathway-centric Approach to Rare Variant Association Analysis. *T.G. Richardson[1], N.J. Timpson[1], C. Campbell[2], T.R. Gaunt[1].* 1) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, Univ, Bristol, United Kingdom; 2) Intelligent Systems Laboratory, University of Bristol, Bristol, United Kingdom.

Current endeavours in rare variant analysis are typically underpowered when investigating signals from individual genes. We undertook a novel approach to rare variant analysis by utilising biological pathway information to analyse functionally relevant genes together.Using whole genome sequence data from the UK10K project, we collapsed all rare variants together that were located in genes that resided along the same pathway, according to definitions from three curated databases. Variants were filtered according to predicted consequence and predicted deleterious impact. The sequence kernel association test (SKAT) was used to test association between collapsed variants and cardiovascular traits after applying thresholds of 1% or 0.5% minor allele frequency (MAF). Two pathways provided strong evidence of association with cardiovascular traits after filtering rare variants according to a strict filter based on predicted deleterious impact (MAF $\leq$ 0.5%). One of these pathways (the Reactome pathway for transcriptional regulation of white adipocyte differentiation) also showed evidence of replication of its observed association with diastolic blood pressure ($P=3.22x10^{-5}$) using imputed data from the ALSPAC cohort ($P=9.85x10^{-3}$). Our follow-up analyses found that the strength of evidence diminished when analysing genes in this pathway individually, suggesting that they would have been overlooked in a conventional gene-based analysis.In conclusion, we have undertaken a novel approach to rare variant analysis and identified signals from two biological pathways which collectively provide much stronger evidence of association in contrast to analysing their variants using single gene-based approaches. Future studies which adopt similar approaches to investigate polygenic effects should yield value in better understanding the genetic architecture of complex disease.

220

PolyTest – a novel method for joint analysis of genome-wide association studies and functional annotations. *D. Golan[1,2], A. Raj[1], K. Gaulton[3], S. Jain[4], D. Calderon[5], Y. Field[1], T. Raj[1,6,7], J. Pritchard[1,8,9].* 1) Department of Genetics, Stanford University, Stanford, CA 94305, USA; 2) Department of Statistics, Stanford University, Stanford, CA 94305, USA; 3) Wellcome Trust Centre for Human Genetics, Oxford UK OX3 7BN; 4) Undergraduate School, Stanford University, Stanford, CA 94305, USA; 5) Department of Biomedical Informatics , Stanford University, Stanford, CA 94305, USA; 6) Departments of Neurology and Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA; 7) The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; 8) Department of Biology, Stanford University, Stanford, CA 94305, USA; 9) Howard Hughes Medical Institute.

A key challenge of human geneticists is to leverage genetic variation for better understanding of the biological mechanisms driving human disease. One highly successful approach is to test for enrichment of GWAS hits within functional annotations such as tissue- or cell-specific active enhancers, thus implicating those tissues and cells that are involved in the etiology of the disease. For example, SNPs associated with type-2 diabetes are enriched in genomic regions that are annotated as active enhancers in pancreatic islets. We present polyTest, a novel method for joint analysis of GWAS test statistics and functional annotations. PolyTest is inspired by the success of random-effects modeling in the context of heritability estimation. We model the effect of each SNP as a random variable whose variance is governed by a given list of annotations, where relevant annotations imply higher variance and lead to larger effects. This approach has several benefits: (a) it can accumulate information across millions of SNPs (b) the model can be represented as a non-standard generalized linear model, yielding very efficient methods for estimation and inference, thus allowing the joint analysis of millions of SNPs and hundreds of annotations of any type, (c) it provides a natural way to control for and interpret the effects of confounders (such as allele frequency) (d) it can easily account for the effects of linkage disequilibrium. These benefits set polyTest apart from existing methods that are limited in one or more of these aspects. We demonstrate the utility of polyTest in three ways. First, we use polyTest to jointly analyze seven well-studied GWAS and well-studied annotations of active promoters in 34 tissues and cell-types. Remarkably, we show that even after removing all SNPs with p-value<0.001, as well as any SNP falling within a 100Kb window around them, we still see significant associations between, e.g., type-2 diabetes and adipose tissue, Schizophrenia and mid-frontal lobe, and Alzheimer's disease and Treg cells. This result demonstrates the highly polygenic nature of these diseases. Second, we use k-mer counts around SNPs to identify short motifs that are associated with increased or decreased association. Our results coincide with a recently discovered enrichment of specific 2-mers in general enhancers. Lastly, we explore the idea of using trans-QTL associations as annotations for the purpose of identifying genes that are involved in disease pathways.

221

Sex-specific gene co-expression networks. *B.E. Engelhardt[1], C. Gao[2], C.D. Brown[3].* 1) Department of Computer Science, Princeton University, Princeton, NJ; 2) Department of Statistical Science, Duke University, Durham, NC; 3) Department of Genetics, University of Pennsylvania, Philadelphia, PA.

Gene expression profiles that differ across various covariates, such as age or sex, have proven useful to study condition-specific and condition-differential processes. Differential gene co-expression networks, which capture pairs of genes that are co-expressed differentially across conditions, or condition-specific gene co-expression networks, which capture genes that are co-expressed uniquely in one condition, enable the study of condition-specific processes at a higher fidelity by recovering gene interactions. Up to now, differential and condition specific co-expression networks have been identified on an edge-by-edge basis. Here, we propose a method for identifying complex, connected differential and condition-specific gene co-expression networks globally from gene expression data in a supervised setting. Our method involves sparse supervised estimation of precision matrices from gene expression data, controlling for all other gene co-expression signals in the data. We apply this method to more than 8,500 gene expression profiles from the Genotype Tissue Expression (GTEx) project v6 data, considering covariates such as sex, age, tissue type, and BMI. We found a number of known and novel co-expressed genes specific to and differential across these covariates. For example, in human sex-differential co-expression networks, three genes we found central to the network structure were *UTX*, *ZFX*, and *USP9X*, all of which are known to play a role in sex determination or sex-differential regulation. Similarly, the sex-specific networks—the female-specific network in particular—recovered as the second most central gene *FIGNL1*, whose disregulation in mice testes is associated with reduced testis size. We validated our networks by finding trans-eQTLs that were identified using the condition-specific co-expression networks. In particular, for genes with cis-eQTLs, we tested for association of the cis-eQTL SNP with each of the gene's neighbors in the condition-specific network, limiting the association tests to samples with that condition (e.g., male). We show that trans-eQTLs in the network neighbors of the cis-eQTL target gene are enriched in samples with the specific condition but not across conditions, validating our condition-specific networks. We have initiated other analyses based on these careful models of condition-specific sources of co-variation in gene expression data, including functional and evolutionary studies and heritability studies of network genes.

222

Detection of Master Regulatory SNPs in expression and methylation quantitative trait loci studies. *J. Shi[1], W. Wheeler[2], A. Battle[3], S. Mostavi[4], X. Zhu[5], M.M. Weissman[6], J.B. Potash[7], S.B. Montgomery[5], N.E. Caporaso[1], M.T. Landi[1], D.F. Levinson[5].* 1) National Cancer Institute, Bethesda, MD; 2) Information Management Services, Bethesda, MD; 3) John Hopkins University, Baltimore, MD; 4) University of British Columbia, Vancouver, British Columbia, Canada; 5) Stanford University, Stanford, CA; 6) Columbia University and New York State Psychiatric Institute, New York, NY; 7) University of Iowa Hospitals & Clinics, Iowa City, IA.

One central but challenging problem in expression or methylation QTL studies is to identify master regulatory SNPs (MRS) that are associated with many traits in *trans.* Identifying MRS helps in understanding gene regulation and the biological mechanisms of associations between genetic loci and diseases. Our theoretical analyses show that the poor performance of methods to detect MRS is mainly due to the extensive correlations among traits, caused by either shared biological regulation or uncorrected hidden factors. We developed a statistical test for detecting MRS which effectively eliminates the impact of extensive correlations by adjusting for the empirical null distribution. The significance of the test is evaluated by permutations while retaining the correlations of traits. Simulation studies with correlations based on real data confirmed that the new method has superior performance. We applied our method to DNA methylation QTL data with 210 normal lung tissues and ~340K CpG probes. The top SNP rs1214759 (P=3.8×10^{-6}) was associated with the methylation of 80 CpG probes in *trans*, of which 53 were replicated (P<0.05) in 65 normal lung tissues from TCGA. We then applied the method to eQTL data based on RNA sequencing of 922 blood samples. We identified 22 MRS with P<5×10^{-8} and 33 MRS with FDR<5%, each associated with expression of between 5 and 242 genes. For a few MRS, *trans*-associations were mediated, mostly partially, by a *cis*-regulated gene. For four MRS located on 6p21.33, 7p21.3, 8q21.13 and 9p24.1, their *trans*-regulated genes were strongly enriched in immune system process (GO analysis). Rs1354034 was previously reported as the SNP most strongly associated with platelet counts (PLT) and mean platelet volume (MPV). In our study, rs1354034 was identified as the strongest MRS (P=1.5×10^{-26}) and it *trans*-regulated 242 genes that were strongly enriched in GO categories for blood coagulation and platelet activation/degranulation, providing a mechanism for the association with PLT and MPV. Finally, rs4895441 was associated with multiple traits including beta thalassemia, glycated hemoglobin, MPV, PLT, HbA2 levels, hemoglobin E disease, corpuscular hemoglobin and red/white blood cell traits. These associations may be explained by the fact that rs4895441 *trans*-regulated multiple genes including *HBE1*, *HBBP1*, *HBG2* and *HBG1.* In summary, our method improves power to detect MRS, providing biological explanations for associations with multiple traits.

223

Proper Use of Allele-Specic Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data. *Y.J. Hu[1], W. Sun[2], J.Y. Tzeng[3,4], C.M. Perou[5].* 1) Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA; 2) Department of Biostatistics, University of North Carolina, Chapel Hill, NC; 3) Department of Statistics, North Carolina State University, Raleigh, NC; 4) Department of Statistics, National Cheng-Kung University, Tainan Taiwan; 5) Department of Genetics, University of North Car- olina, Chapel Hill, NC.

Studies of expression quantitative trait loci (eQTLs) offer insight into the molecular mechanisms of loci that were found to be associated with complex diseases and the mechanisms can be classified into cis- and trans-acting regulation. At present, high-throughput RNA sequencing (RNA-seq) is rapidly replacing expression microarrays to assess gene expression abundance. Unlike microarrays, RNA-seq also provides information on allele-specific expression (ASE), which can be used to distinguish cis-eQTLs from trans-eQTLs and, more importantly, enhance cis-eQTL mapping. However, assessing the cis-effect of a candidate eQTL on a gene requires knowledge of the haplotypes connecting the candidate eQTL and the gene, which cannot be inferred with certainty. The existing two-stage approach that first phases the candidate eQTL against the gene and then treats the inferred phase as observed in the association analysis tends to attenuate the estimated cis-effect and reduce the power for detecting a cis-eQTL. In this article, we provide a maximum-likelihood framework for cis-eQTL mapping with RNA-seq data. Our approach integrates the inference of haplotypes and the association analysis into a single stage, and is thus unbiased and statistically powerful. We also develop a pipeline for performing a comprehensive scan of all local eQTLs for all genes in the genome by controlling for false discovery rate, and implement the methods in a computationally efficient software program. The advantages of the proposed methods over the existing ones are demonstrated through realistic simulation studies. An application to the breast cancer data from The Cancer Genome Atlas project identified 2,486 eQTLs, determined their cis- or trans- mechanisms, and suggested functional roles of 6 eQTLs that overlap with brease cancer-associated SNPs.

224

Tensor decomposition uncovers trans eQTL networks in the multi-tissue EuroBATS study. *J. Marchini[1,2], V. Hore[1], A. Vinuela[3], A. Buil[4], M. McCarthy[2], K. Small[3].* 1) Dept Statistics, Oxford Univ, Oxford, UK; 2) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK;; 3) Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom; 4) Dep. Genetic Medicine and Development, University of Geneva, Geneva, Switzerland;.

Uncovering trans eQTL networks in gene expression studies of multiple tissues is a challenging statistical problem. To tackle this problem, we have developed a general framework for decomposing matrices and tensors of multi-tissue gene expression datasets into sparse latent factors, where latent factors consist of networks of co-varying genes. The model also determines the subset of tissues in which each latent factor is active. We fit our model using variational Bayes, which allows for relatively fast inference on large data sets, and handles missing data. We then use the individual scores vector of each factor (or component) as a phenotype in a GWAS to identify genetic variants that drive the gene network associated with that component. We have applied our method to data from the EuroBATS project which consists of gene expression measured via RNA sequencing on 845 related individuals from the TwinsUK cohort in LCLs, adipose tissue and skin tissue. Our method uncovers many sparse components, many of which exhibit strong statistical and biological significance. Some notable findings include the following (a) we clearly identify the role of *KLF14* as a master regulator of gene expression in adipose tissue, (b) we find a component that links the two transactivators *CIITA* and *RFX5* with genes in the MHC class II, as well as other known targets of the *CIITA* transcription factor, and we uncover a cis-eQTL SNP for the *CIITA* gene ($p<1e^{-10}$), (c) similarly we are able to link the transactivator *NLRC5*/*CITA* with genes in the MHC class I, (d) we find a component with a cis eQTL in the region of the *CEBPD* gene ($p<1e^{-10}$), which is a bZIP transcription factor important in the regulation of genes involved in immune and inflammatory responses. This component shows strong enrichment ($p<1e^{-32}$) for genes involved in inflammatory response, (e) We uncover a link between genetic variation in the *SENP7* gene on chr3 ($p<1e^{-10}$) and a cluster of zinc finger genes on chromosome 19. In addition, we are able to show that several of the dense components found by our method correlate strongly ($p<1e^{-20}$) with confounding variables such as insert size and GC content of the sequencing. Overall, these results illustrate the utility of this method to uncover real biological signals in multi-tissue gene expression studies, whilst detecting and correcting for confounding effects.

225

Integrative Genome-wide Gene Expression and Metabolomics Networks in Pregnant Women Identify Vitamin D Variants Associated with Incident Preeclampsia Cases. *J. Lasky-Su[1], A. Sharma[1], C. Clish[2], A. Litonjua[1], S. Weiss[1], IMPACT.* 1) Channing Department of Network Medicine, Brigham & Women's Hosp, Boston, MA; 2) Broad Institute, Cambridge, MA.

Preeclampsia (PE) is a leading cause of maternal and fetal mortality and morbidity worldwide, affecting up to 8% of pregnancies. Despite this, the pathophysiology of PE remains elusive. To date, gene expression and metabolomic studies for PE have been limited in scope and size. We used 74 PE cases and 165 matched controls that were enrolled in the Vitamin D Antenatal Asthma Reduction Trial, a clinical trial that randomized pregnant women to high and low doses of vitamin D during pregnancy (400vs.4400IU). We generated genome-wide gene expression and metabolomic profiles from whole blood and plasma that was taken in the second (10-18wks) and third (32-38wks) trimesters of pregnancy using PE cases and matched controls. We first generated network to predict PE with genome-wide gene expression and metabolomic data using weighted gene co-expression network analysis. The networks were then interconnected using established physical protein interactions that were obtained from the use of several curated databases, containing a total of 13,460 proteins that are interconnected by 141,296 physical interactions. The gene expression network identified a gene module that fully encompasses the vitamin D pathway. We observed that 1322 genes in this module were a part of the largest connected component (LCC) of the module (Zscore=7.5). Further, in the LCC, we found that 45 genes contained partially methylated domains seen in the placenta of PE subjects. We then performed articulation point analysis, to identify the crucial genes whose removal would disconnect the module. In addition to immune pathways, the vitamin D pathway was fully included in the module (p=0.0065), with a significant number of vitamin D-specific probes contained in the LCC that were also identified as articulation points. This module was then validated using three external gene expression datasets from GEO. While the vitamin D pathway was also identified in the metabolomic network, the differences in network hubs suggested that each network within the multidimensional PE network provided unique information relevant to the disease outcome that may not be captured in the other networks, suggesting that the nodes central in one network type might play a specific functional role and lose this property in the transition to other network types. This highlights the importance of the generation of a multidimensional PE networks in elucidating our understanding of the biological transitions that result in PE.

226

Are Genetic Interactions Influencing Gene Expression in Humans Evidence for Biological Epistasis or Statistical Artifacts? *A. Fish[1], J.A. Capra[1,2], W.S. Bush[3].* 1) Vanderbilt Institute for Genetics, Vanderbilt University, Nashville, TN; 2) Department of Biological Sciences, Vanderbilt University, Nashville, TN; 3) Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University.

Interactions between genetic variants, also called epistasis, are observed in a variety of model organisms and are hypothesized to account for heritability in complex human traits. The extent and importance of genetic interactions in humans remains unknown because statistical interactions between variants can be produced through processes other than biological epistasis. In this study, we accounted for technical artifacts, statistical artifacts, and biological phenomena other than epistasis that are capable of producing signatures of interactions in common statistical tests to evaluate the evidence for true epistasis impacting a human trait. We first identified 1,093 significant statistical interactions between pairs of *cis*-regulatory variants that impact gene expression in human lymphoblastoid cell lines. We then determined whether these interactions could be explained by an underappreciated type of population stratification, ceiling/floor effects, haplotype effects, or the tagging of single variants through linkage disequilibrium. Overall, we identified 15 interacting loci that are robust to established alternate explanations and are consistent with true biological epistasis. Since the majority of interactions were consistent with multiple explanations, we used functional genomics data to determine if biological epistasis was plausible mechanistically. Interacting variants were strongly enriched within enhancers, promoters, and the binding sites for many transcription factors, including CTCF and cohesin, which mediate chromatin looping and suggest a physical mechanism underlying interactions. We therefore conclude that genetic interactions impacting gene expression likely exist in humans. We additionally demonstrate that while many statistical interactions were consistent with other biological explanations, some of these associations would have gone undetected in a standard single-marker analysis. Ultimately, we identified new, complex genetic architectures underlying the regulation of 23 genes in lymphoblastoid cells; this suggests that single-SNP analyses of *cis*-regulatory regions may miss important modifiers relevant for a host of traits.

227

Building a platinum assembly from single haplotype human genomes generated from long molecule sequencing. *K. Meltz Steinberg[1], T.A. Graves-Lindsey[1], V.A. Schneider[2], R.S. Fulton[1], J. Chin[3], M. Kremitzki[1], W.C. Warren[1], D.M. Church[4], E.E. Eichler[5,6], R.K. Wilson[1].* 1) McDonnell Genome Institute, Washington University, St. Louis, MO; 2) National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; 3) Pacific Biosciences of California, Inc., Menlo Park, CA 94025; 4) Personalis, Inc., Menlo Park, CA 94025; 5) Department of Genome Sciences, University of Washington, Seattle, WA 98195; 6) Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195.

The human reference sequence has provided a foundation for studies of genome structure, human variation, evolutionary biology, and disease. At the time the reference was originally completed there were some loci recalcitrant to closure; however, the degree to which structural variation and diversity affected our ability to produce a representative genome sequence at these loci was still unknown. Many of these regions in the genome are associated with large, repetitive sequences and exhibit complex allelic diversity such producing a single, haploid representation is not possible. To overcome this challenge, we have sequenced DNA from two hydatidiform moles (CHM1 and CHM13), which are essentially haploid. CHM13 was sequenced with the latest PacBio technology (P6-C5) to 52X genome coverage and assembled using Daligner and Falcon v0.2 (GCA_000983455.1, CHM13_1.1). Compared to the first mole (CHM1) PacBio assembly (GCA_001007805.1, 54X) contig N50 of 4.5Mb, the contig N50 of CHM13_1.1 is almost 13Mb, and there is a 13-fold reduction in the number of contigs. This demonstrates the improved contiguity of sequence generated with the new chemistry. We annotated 50,188 RefSeq transcripts of which only 0.63% were split transcripts, and the repetitive and segmental duplication content was within the expected range. These data all indicate an extremely high quality assembly. Additionally, we sequenced CHM13 DNA using Illumina SBS technology to 60X coverage, aligned these reads to the GRCh37, GRCh38, and CHM13_1.1 assemblies and performed variant calling using the SpeedSeq pipeline. The number of single nucleotide variants (SNV) and indels was comparable between GRCh37 and GRCh38. Regions that showed increased SNV density in GRCh38 compared to GRCh37 could be attributed to the addition of centromeric alpha satellite sequence to the reference assembly. Alternatively, regions of decreased SNV density in GRCh38 were concentrated in regions that were improved from BAC based sequencing of CHM1 such as 1p12 and 1q21 containing the SRGAP2 gene family. The alignment of PacBio reads to GRCh37 and GRCh38 assemblies allowed us to resolve complex loci such as the MHC region where the best alignment was to the DBB (A2-B57-DR7) haplotype. Finally, we will discuss how combining the two high quality mole assemblies can be used for benchmarking and novel bioinformatics tool development.

228

Building a Better Human Genome Reference and Targeting Structure using Single Molecule Technologies. *R. Sebra[1], M. Pendleton[1], A. Pang[2], A. Ummat[1], O. Franzen[1], T. Rausch[3], A. Stütz[3], W. Stedman[2], T. Anantharaman[2], A. Hastie[2], H. Dai[2], M. Fritz[3], H. Cao[2], A. Cohain[1], G. Deikus[1], L. Newman[1], S. Scott[1], A. Uzilov[1], R. Durrett[4], S. Blanchard[5], R. Altman[4], C. Chin[6], E. Paxinos[6], J. Korbel[3,7], R. Darnell[8,9], W. McCombie[10,11], P. Kwok[12], C. Mason[4,13], E. Schadt[1], A. Bashir[1].* 1) Icahn School of Medicine, Mount Sinai, NY School of Natural Sciences, NYC, NY, USA; 2) BioNano Genomics, San Diego, CA, USA; 3) European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany; 4) The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, 1305 York Ave., Weill Cornell Medical College, New York, NY 10065, USA; 5) Department of Physiology and Biophysics, Weill Cornell Medical College, 1300 York Avenue, New York, New York 10064, USA; 6) European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute (EMBL EBI), Wellcome Trust Genome Campus, Hinxton, Saffron Walden CB10 1SD, UK; 7) Pacific Biosciences, Menlo Park, California, USA; 8) Laboratory of Neuro-Oncology, The Rockefeller University, New York, NY, USA; 9) Howard Hughes Medical Institute, New York, NY, USA; 10) The Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA; 11) The Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA; 12) Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA; 13) Department of Medicine, Division of Hematology/Oncology, 1300 York Ave., Weill Cornell Medical College, New York, NY 10065, USA.

Advances in high-throughput technologies have increased our ability to survey human genomes, yet assessing large-scale variation in genome architecture mostly unresolved. Even with complex cloning and sequencing strategies, the ubiquitous advances in next-generation sequencing (NGS) technologies and data processing algorithms, a large portion of the genome cannot be disambiguated. Here, we present the first comprehensive analysis of a human genome that combines single-molecule sequencing (PacBio RSII) with single molecule genome maps (BioNano Genomics' Irys). The resulting hybrid assembly dramatically improves upon assembly contiguity observed in shotgun sequencing approaches, with N50s >30 Mb and even resolves gaps in the recently released hg38 assembly. By comparing to the human reference genome, we are able to identify complex structural variations otherwise missed by NGS approaches. Furthermore, by combining Illumina short read data with long reads, we are able to resolve low-accuracy intervals and phase both SNVs and structural events - achieving haplotypes that are over 99% consistent with previous trio-based studies. Using novel algorithms to integrate single molecule and next generation sequencing technologies, we believe that one can generate clone-free genomes that rival, or surpass, the best human reference assemblies available today. As these single molecule methods mature, fully *de novo* WGS approaches and targeted gene characterization will increasingly become a standard practice, and inference of variation will be replaced by a more direct, comprehensive characterization of genome variation to accelerate our understanding of complex phenotypes that are induced. Beyond resolving structural information using WGS, we present methods for targeting whole genes of interest using long-range, high fidelity PCR and/or other capture methods in tandem with long read sequencing. These targeted long read methods motivate the potential for deriving niche diagnostics in specific genes (examples include BRCA1, BRCA2, CYP2D6, HLA) for the comprehensive assessment of genetic variation in pathologically relevant regions to explore the relationship of structurally complex domains using long reads. The combination of these results reinforce the pressing need to derive higher resolution human genome references and targeted gene sequencing methods to expand utility beyond that achieved with traditional, high throughput sequencing methods.

229

Genome in a Bottle: You may have sequenced, but how well did you do? *J.M. Zook[1], H. Parikh[1], M. Salit[2,3], Genome in a Bottle Consortium and Global Alliance for Genomics and Health.* 1) Genome-scale Measurements Group, National Institute of Standards and Technology, Gaithersburg, MD; 2) Genome-scale Measurements Group, National Institute of Standards and Technology, Stanford, CA; 3) Bioengineering Dept, Stanford University, Stanford, CA.

Purpose: Clinical laboratories, research laboratories and technology developers all need DNA samples with reliably known genotypes in order to help validate and improve their methods. The Genome in a Bottle Consortium (genomeinabottle.org) has been developing Reference Materials with high-accuracy whole genome sequences to support these efforts. **Methodology:** Our pilot reference material is based on Coriell sample NA12878 and was released in May 2015 as NIST RM 8398 (tinyurl.com/giabpilot). To minimize bias and improve accuracy, 11 whole-genome and 3 exome data sets produced using 5 different technologies were integrated using a systematic arbitration method [1]. The Genome in a Bottle Analysis Group is adapting these methods and developing new methods to characterize 2 families, one Asian and one Ashkenazi Jewish from the Personal Genome Project, which are consented for public release of sequencing and phenotype data. We have generated a larger and even more diverse data set on these samples, including high-depth Illumina paired-end and mate-pair, Complete Genomics, and Ion Torrent short-read data, as well as Moleculo, 10X, Oxford Nanopore, PacBio, and BioNano Genomics long-read data. We are analyzing these data to provide an accurate assessment of not just small variants but also large structural variants (SVs) in both "easy" regions of the genome and in some "hard" repetitive regions. We have also made all of the input data sources publicly available for download, analysis, and publication.**Results:** Our arbitration method produced a reference data set of 2,787,291 single nucleotide variants (SNVs), 365,135 indels, 2744 SVs, and 2.2 billion homozygous reference calls for our pilot genome. We found that our call set is highly sensitive and specific in comparison to independent reference data sets. We have also generated preliminary assemblies and structural variant calls for the next 2 trios from long read data and are currently integrating and validating these.**Discussion:** We combined the strengths of each of our input datasets to develop a comprehensive and accurate benchmark call set. In the short time it has been available, over 20 published or submitted papers have used our data. Many challenges exist in comparing to our benchmark calls, and thus we have worked with the Global Alliance for Genomics and Health to develop standardized methods, performance metrics, and software to assist in its use.[1] Zook et al, Nat Biotech. 2014.

230

An Accurate Read Mapper for Graph Genomes. *W. Lee[1], K. Ghose[1], V. Semenyuk[1], D. Kural[1], R. Brown[2], A. Jain[2], B. Murray[2], B. Pollex[2], J. Browning[1], A. Stachyra[1], F. Sung[1].* 1) R&D, Seven Bridges Genomics, Cambridge, MA; 2) R&D, Seven Bridges Genomics, London, UK.

Classical short-read mapping algorithms utilize a linear genome reference sequence to align reads from a newly sequenced individual. Many reads fail to map or are incorrectly mapped because each new genome typically contains genomic variations not present in the reference sequence. As a result, while it is possible to detect SNPs and very short INDEL variants using such mappings, longer INDELs and structural variations are often missed. Furthermore, undetected structural variants in a new sample often cause mismappings that lead to false positive variant predictions.

Projects, such as the 1000 Genomes Project, have analyzed the genomes of thousands of individuals from different populations, allowing us to understand how genomes vary between humans. A key insight from these projects is that most of the variants in an individual are shared by the population. This has led to the hypothesis that, by incorporating known variants into the current linear reference we can improve short read alignments and variant calling.

We have developed a novel whole-genome read mapper that takes known variations into account when mapping reads. A graph is constructed by combining the linear reference genome with a list of variants. The graph mapper takes FASTA/Q and VCF as inputs, and generates standard BAM files ensuring compatibility with the majority of other available bioinformatics tools. Constructing a graph with 1000 Genomes Project variant list (81 million variations) takes 698 seconds and requires 16 Gb. The mapper processes ~500 reads per second per thread.

We simulated 18,827 insertions with lengths ranging from 2 to 96 bp. We aligned simulated reads against a graph constructed with human genome and these variants. We passed the graph aligned BAM files to GATK, Samtools and Freebayes to call variants. The overall insertion detection accuracy improved by 15% (from 78% to 93%) when compared to alignments using a linear mapper (BWA).Our experiments suggest that large gains in variant calling sensitivity can be achieved by incorporating a graph genome based mapper into bioinformatics pipelines. We believe that such a graph based pipeline will be of great use to scientists and clinicians who need to perform fast and accurate comparisons of new samples to existing populations, such as that required for precision medicine.

231

Anchored Pseudo-*De Novo* Assembly of Human Genomes Identifies Extensive Sequence Variation in Unmapped Sequence Reads. *K.H. Brown, J.J. Faber-Hammond.* Department of Biology, Portland State University, Portland, OR.

The Human Reference Genome (HGR) completion marked the beginning of the genomics era and has been instrumental in broad ranging discoveries. Despite this, limitations arising from using limited numbers of individuals from a single ethnic population exist, highlighted by the number of high quality sequence reads that failing to map using common genome resequencing workflows. While the portion of high quality raw sequences failing to map generally represents 2-5% of total reads, these sequences may harbor regions that would enhance our understanding of population variation, human evolution, and genetic disease. To examine genomic content in these regions, we developed a bioinformatics pipeline treating sequenced mate-pairs as separate reads for mapping while exporting unmappable reads. This method isolated 4.5X more unmappable reads than traditional paired-end mapping methods. Isolated reads are then assembled de novo for individuals before being combination into a secondary reference assembly. Using 45 diverse individuals from the 1000 Genomes Project, we identified 353,412 unmapped contigs covering 197.2 Mb of non-redundant sequence. 31,250 contigs are represented in mutiple individuals with ~40% showing high sequence complexity. Genome map coodinates were generated for 99.8%, with 23.7% exhibiting high quality mapping scores. Comparative genomic analyses with archaic human and primate species revealed significant sequence alignment and comparisons with model organism RefSeq gene datasets identified novel human genes. This study expands the HRG and highlights the continuing need for human genome analysis of unmapped sequences to explore biological functions contributing to human phenotypic variation, disease and functionality for personal genomic medicine.

232

A Diploid Personal Human Genome Reference from Diverse Sequence Data – A Model for Better Genomes. *K.C. Worley[1, 2], Y. Liu[1], D.S.T. Hughes[1], S.C. Murali[1], R.A. Harris[1, 3], A.C. English[1], O.A. Hampton[1], C.R. Beck[2], Y. Han[1, 2], M. Wang[1, 2], H. Doddapaneni[1, 1], C.L. Kovar[1], W.J. Salerno[1, 2], S. Richards[1, 2], J. Rogers[1, 2], J.R. Lupski[2], D.M. Muzny[1, 2], R.A. Gibbs[1, 2].* 1) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 2) Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX; 3) Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, TX.

High quality reference genomes are the foundation of genomic research. The human reference genome, although finished to highest quality, is a haploid mosaic sampled from several individuals and each parental haplotype within those individuals– moreover, it represents a haploid reference genome. The few published assembled genomes from a single individual all have limitations such as poor contiguity, and none of these represent the diploid nature of a personal genome. We report here the assembly of data from a single individual (HS1011). The data include a variety of Illumina sequence reads and libraries (180 bp, 300 bp and 500 bp paired end data; 3kb, 6.5 kb and 8 kb mate-pair data), as well as Illumina Hi-C data, supplemented with 20x PacBio RS long read data and BioNano optical mapping data. The assembly is highly contiguous with a 394 kb Contig N50 and 148 Mb Scaffold N50. All contigs are placed on chromosomes. Data from the parents of HS1011 allow us to phase variants within this genome. With this assembly and the underlying data, we are exploring genomic features beyond the coding regions of the exome. Structural variants, particularly insertions and inversions are difficult to characterize with the common human sequencing strategies that utilize exome sequencing or short sequence reads from small fragments. Using the diverse HS1011 data, we have identified 78 putative novel insertions and 70 putative tandem duplications. The insertions were defined using assembly of reads that did not map to the GRCh38 reference. These calls were supported by 6.5 kb mate-pair data, and were confirmed in the *de novo* WGS assembly contigs and the 3 kb mate-pair data. Tandem duplication identification utilizes the hierarchy of fragment sizes sampled in the data with the 300 bp and 500 bp read pairs identifying the boundaries and the larger read pairs covering the duplication and confirming the size of the duplicated region. These data allow us to explore the utility of the different data types in the definition of structural variation and to define which heterozygous variants are located on the same haplotype to enable downstream analyses that benefit from haplotype-aware methods. With these data, we examine the relative merits of the different data types to understanding the comprehensive clinical genome.

233

Genome-wide copy number detection using a hybrid clinical NGS assay. *J. Harris, G. Bartha, S. Luo, A. Patwardhan, S. Garcia, S. Chervitz, M. Morra, D. Church, J. West, R. Chen.* Personalis, Inc., Menlo Park, CA.

Whole exome sequencing has enabled cost-effective sequencing in the 100-500X fold coverage range: a depth critical for detecting low-frequency inherited mosaic or cancer variants in medically important regions, and which is typically outside the practical range for whole genome sequencing, in terms of cost and data burden. However, robust detection of copy-number variations (CNVs) using whole-exome data can be problematic, as clinically relevant CNVs may lie beyond the bounds of the exome. Even if a CNV does intersect the exome, its breakpoints may not be captured, making it difficult to reliably detect. We present a hybrid whole exome/whole genome approach that supplements exome sequencing with thin whole-genome sequencing (thinWGS) in a single clinical NGS assay. We show that such a hybrid whole exome/whole genome approach provides a cost-effective clinical assay that can simultaneously achieve sensitivity for low frequency inherited mosaic and somatic cancer variants while also delivering genome-wide sensitivity to large CNVs. We have developed a CNV detection algorithm for the thinWGS assay, based on measuring the mean read depth in large (10—20 kbp) bins of the alignment reference. CNVs in the sample manifest as regions that deviate sharply from the baseline coverage level. We employ a Hidden Markov Model (HMM) to define event breakpoints. We further refine the raw CNV detections by demanding that the morphology of the read-depth profile matches expectations for a true CNV signal. For example, in a deletion, the read depth should drop to zero (for a homozygous deletion), or to half the depth of the flanking bins (for a heterozygous deletion), and the transition between read-depth states should be very sharp. We find that including filters based on these and other morphological features decreases the false-positive detection rate of the method significantly. Using Coriell samples with known deleterious CNVs and internally-developed gold-standard CNV call sets, we evaluate the performance of the thinWGS assay in combination with our CNV detection and refinement algorithm, as a function of CNV type and size. In particular, we employ a set of 28 samples for which we have both 60x WGS data and thinWGS data. We use the WGS data to construct a set of ground-truth CNVs in each of these samples, and then test our ability to detect these CNVs in the thinWGS data. We find that the thinWGS assay is >=90% sensitive to CNVs down to a size of ~60 kbp.

234

A reference panel of common CNVs and SVs identified from high depth sequencing data on over 2000 samples of European ancestry. *M.A. Bekritsky[1], J. O'Connell[1], S.S. Ajay[2], M.A. Eberle[2].* 1) Illumina Cambridge Ltd., Chesterford Research Park, Saffron Walden, Essex, CB10 1XL, UK; 2) Illumina Inc., 5200 Illumina Way, San Diego, CA 92122, USA.

Large population-level studies such as the 1000 Genomes Project have greatly improved our understanding of the variation of SNPs and small indels within the human genome. In addition to cataloging genetic variation, SNP calling has made remarkable progress due to the work of the 1000 Genomes Project and variant calling truthsets such as the Platinum Genomes. To build upon the success of these initial efforts, we are compiling a comprehensive and accurate catalogue of CNVs and SVs that can provide reliable truthsets to assess and improve detection of larger and more complex mutations. Additionally, as sequencing technology matures, we must have a thorough understanding of genomic regions that remain difficult to interpret. The latest generation of high throughput sequencing platforms enables sequencing studies on increasingly large populations, allowing us to leverage aggregate statistical analysis of highly accurate small variant calls to improve our understanding of larger, more complex genomic variation. To do this, we are developing SNP-based methods to identify regions of the reference genome that consistently violate Hardy Weinberg equilibrium and other genotyping assumptions, such as the percent of samples with a genotype call or allelic balance. We can further analyze these regions to identify why these assumptions are violated. For example, some might be due to difficulties in error-prone regions of the genome, such as long simple repeats or low complexity regions. Others may deviate from diploid expectation because they overlap common CNVs and SVs.We are applying these methods to high-depth (~40x) sequencing data from over 2,200 samples of European ancestry containing ~9.5 million SNPs (MAF>1%). We observe ~85,000 regions of the genome where at least 2 consecutive SNPs violate Hardy Weinberg equilibrium. Based on our preliminary analyses, we predict that these represent ~49,000 common deletions covering over 36 Mb and ~27,000 common amplifications covering over 7 Mb, as well as ~9,000 complex events covering over 9 Mb that may either be parts of the genome prone to sequencing errors or complicated SVs. This analysis can also be extended to compare CNV and SV frequencies among ancestral populations. This study provides a novel means of leveraging population data as we enter the era of large-scale sequencing studies, taking advantage of basic population genetics concepts to create a powerful genotyping, annotation, and benchmarking tool and dataset.

235

Integrative analysis of five cancer GWAS meta-analyses with eQTLs and splice QTLs from relevant normal tissues in GTEx proposes causal regulatory processes and genes for cancer risk. *A.V. Segrè[1], D.S. DeLuca[1], T. Sullivan[1], E. Gelfand[2], S.B. Gruber[3,4], G. Casey[4], D.J. Hunter[5], B. Henderson[6], T. Sellers[7], C.I. Amos[8], D.G. MacArthur[9], S. Lindstroem[5], P. Kraft[5], A. Kristin[2], G. Getz[1,10,11], The GTEx Consortium, GAME-ON Network, CORECT, DRIVE, ELLIPSE, FOCI and TRICL Consortia.* 1) Cancer Program, Broad Institute, Cambridge, MA; 2) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA; 3) Department of Medicine, Keck School of Medicine, University of Southern California, Los Angeles CA; 4) Norris Comprehensive Cancer Center, University of Southern California, Los Angeles CA; 5) Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston MA; 6) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA; 7) Moffitt Cancer Center and Research Institute, Tampa FL; 8) Department of Community and Family Medicine, Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover NH; 9) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 10) Department of Pathology, Harvard Medical School, Boston, MA; 11) Cancer Center, Massachusetts General Hospital, Boston, MA.

The majority of common variant associations with different cancers lie in noncoding regions, suggesting that causal mechanisms of cancer risk alleles may largely be due to alterations in gene regulation. Emerging evidence suggests that DNA variants associated with changes in gene expression levels (eQTLs) or splicing (sQTLs) may help explain underlying mechanisms of GWAS signals and prioritize causal genes, though the extent to which regulatory QTLs contribute to cancer risk is unknown. Here we systematically analyze 5 cancers: breast, prostate, lung, colorectal and ovarian, with GWAS meta-analyses conducted in GAME-ON Initiative (4-16K cases; 9-18K controls), testing whether their genome-wide associations are enriched for *cis*-eQTLs, or sQTLs, from relevant normal human tissues computed in the Genotype-Tissue Expression (GTEx) project (85-280 samples), and if so, whether top eQTL target genes point to specific pathways. Of the genome-wide significant associations per cancer, 3-43% were found to be in linkage disequilibrium (LD) (r^2>0.5) with significant eQTLs (FDR<5%), in a relevant tissue (1,400-7,200 eGenes in lung, mammary breast, transverse or sigmoid colon, prostate or ovary; release v6). New insights are suggested, e.g., while GTEx detects a previously reported eQTL for *CHRNA5* (P=10^{-5}) in lung cancer locus 15q25 (r^2=0.8), it proposes an additional candidate gene, *RP11-650L12*, an antisense RNA with a stronger eQTL (P=10^{-7}, r^2=1 with GWAS SNP). We developed a computational method that tests for enrichment of modest GWAS associations amongst significant QTLs, accounting for confounding factors (e.g. MAF, distance to TSS, LD). If enrichment is found, target genes of eQTLs with top ranked GWAS *p*-values are tested for enrichment in biological pathways. Results for the 5 cancers will be presented. Preliminary analysis of the TRICL lung cancer GWAS meta-analysis, using GTEx pilot phase lung eQTLs (best-eQTL per gene) or sQTLs (Altrans; best sQTL per exon-exon junction), suggests significant enrichment of lung cancer associations in lung eQTLs (P=5x10^{-5}, enrichment fold=1.4; 35 new modest associations) and sQTLs (P<10^{-5}). The top eQTL target genes are nominally enriched for several REACTOME and KEGG pathways, e.g. DNA repair, apoptosis, and complement and coagulation cascades. Alleles decreasing expression of apoptotic genes, *DSP* and *PLEC* are proposed to increase lung cancer risk. This work suggests new candidate cancer risk genes and pathways for follow-up.

236

Mutations in a promoter of *APC* cause a syndrome of gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS) without colorectal involvement. *G. Chenevix-Trench[1], J. Li[1], S. Healey[1], H. Sivakumaran[1], J. French[1], S. Edwards[1], K. Nones[1], N. Waddell[1], P. Pichurin[2], P. Hulick[3], K.J. Hamman[4], U. Rudloff[5], K. Calzone[5], J.J. Waterfall[5], D. Huntsman[6], P. Meltzer[5], D. Neklason[7], D. Goldgar[8], F. Carneiro[9], C. Kiraly-Borri[10], L. Schofield[10], D. Worthley[11], N. Lindor[12], G. Suthers[13], I. Schrader[6].* 1) QIMR Berghofer, Brisbane, QLD, Select a Country; 2) Mayo Clinic College of Medicine, Rochester, MN, USA; 3) NorthShore University Health System, Evanston, IL, USA; 4) Oregon Health & Science University, Portland, OR, USA; 5) National Cancer Institute, Bethesda, MD, USA; 6) University of British Columbia, Vancouver, BC, Canada; 7) Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA; 8) University of Utah, Salt Lake City, Utah, USA; 9) Institute of Molecular Pathology and Immunology of the University of Porto, Portugal; 10) Genetic Services of Western Australia, King Edward Memorial Hospital, Perth, WA, Australia; 11) University of Adelaide, Adelaide, SA, Australia; 12) Mayo Clinic, Scottsdale, Arizona, USA; 13) University of Adelaide, South Australia.

Gastric adenocarcinoma and proximal polyposis of the stomach is a rare cancer syndrome with a significant risk of gastric adenocarcinoma [MIM 613659]. It is characterised by autosomal dominant transmission of fundic gland polyposis, including areas of dysplasia or intestinal-type gastric adenocarcinoma, restricted to the proximal stomach, with no evidence of colorectal or duodenal polyposis, or other heritable cancer syndromes. Using a large Australian pedigree with over 30 affected individuals, we mapped the gene to an interval of 18Mb at 5q22 containing 51 genes, including *APC* [MIM 611731]. In total, we have identified five families with GAPPS, in all of which coding and large rearrangements of *APC* have been excluded by extensive sequencing and multiplex ligation-dependent probe amplification analysis. Extensive targetted (mean coverage = 27X), whole exome (59X) and whole genome sequencing (WGS; 24X using HiSeq) in the largest family failed to identify any novel or rare coding mutations, or obvious regulatory mutations. Whole genome copy number analysis showed loss of heterozygosity (LOH) only on 5q in 7/14 fundic gland polyps from four affected individuals, with loss of the wildtype allele. The minimal common region of loss overlaped the linkage region by 12 Mb, centered around *APC.* We used Sanger sequencing of the promoters to find informative polymorphisms for allelic imbalance analysis at *APC*. This revealed a novel mutation in promoter 1B in the large pedigree (missed by previous WGS because of low coverage, but subsequently found by WGS using XTen). X Ten WGS also identified somatic mutations in *APC* in 4/7 polyps without LOH, providing the 'second hit'. The other four small GAPPS families from North America all had the same novel mutation four base pairs away, which cosegregated with disease. Sanger sequencing of a 3' UTR polymorphism in cDNA showed reduced expression of the mutant allele in blood from affected individuals. Electrophoretic mobility shift assays showed that both mutations reduced transcription factor YY1 binding, as predicted. Furthermore, luciferase assays showed that both mutations abrogated activity of the *APC* 1B promoter in gastric and colorectal cancer cell lines, suggesting that YY1 normally acts as an activator of this promoter. In summary, we have shown that GAPPS is caused by point mutations in the 1B promoter of *APC*, with LOH or somatic mutations involving *APC* common in the fundic gland polyps from affected individuals.

237

***TNS1* mutations and gastrointestinal stromal tumors and defective mitochondria in the *blistery* (*Tns1* knockout) *D. melanogaster*.** *F. Faucz[1], T. Silva[1], S. Reincke[1], A. Sen[2], R. Cox[2], M. Miettienen[3], S. Lo[4], J. Carney[5], C. Stratakis[1].* 1) Section on Endocrinology and Genetics, Program on Developmental Endocrinology and Genetics, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD; 2) Department of Biochemistry and Molecular Biology, F. Edward Hébert School of Medicine, Bethesda MD; 3) Laboratory of Pathology, National Cancer Institute, Center for Cancer Research, Bethesda, MD; 4) Department of Biochemistry and Molecular Medicine, University of California - Davis, Sacramento, CA; 5) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN.

TNS1 gene encodes the actin-binding protein Tensin 1 (TNS1). Tensin1 is an adaptor phosphoprotein linking integrin to the actin cytoskeleton; it is also involved in fibrillar adhesion formation. *Tns1* knockout (KO) mice develop kidney abnormalities; female mice have also reduced fertility. A *Drosophila melanogaster Tns1* KO model is known as *blistery* due to defective wing unfolding and consequent blister formation. *PTEN* (phosphatase and tensin homolog), the known tumor suppressor gene, is a paralog of *TNS1*. Genome-wide association studies have linked *TNS1* to asthma, lung disease and colorectal cancer, and TNS1 expression is reduced or absent in several prostate and breast cancer cell lines. Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal neoplasms of gastrointestinal tract. Most of GISTs carry activating mutations in the *KIT* or *PDGFRA* genes; a smaller number are due to germline mutations in succinate dehydrogenase subunit genes (SDHx). However, up to 10% of GISTs do not have mutations in those genes and are known as wild type (WT). Recently, mitochondrial structural defects were described in WT GISTs (including those without SDHx mutations), in conjunction with their epigenetic down-regulation of *SDHC* gene. After whole-exome sequencing identified a germline *TNS1* defect in a patient with WT GIST, we screened *TNS1* gene in 24 patients with WT GISTs. We identified a total of 12 distinct *TNS1* gene variants, which were predicted to be damaging through *in silico* analysis. The frequency of patients with at least one variation was 50% (12/24). Six of the variants were new, including 4 missense (p.Y327C, p.G1306R, p.G1227D, and p.R894W), an insertion of two amino acids (c.1978insCAGCAG) and a deletion of one base that was absent at the germline level, being present only in the tumor DNA (c.2669_2669delC). The remaining consisted of a deletion of one amino acid, 4 missense substitutions, and a splice site mutation. All 6 previously described variants had low frequency in the general population. We also studied *blistery* fly mitochondria by electron microscopy, identifying several defects: some of the mitochondria were devoided of cristae, exhibited structural abnormalities, and have bigger size similar to the structural defects seen in WT GISTs. Collectively, these data suggest that *TNS1* may play a role in the development of GIST and imply that inactivating *TNS1* variants could lead to defective mitochondria.

238

Exome sequencing provides evidence of pathogenicity for genes implicated in the development of colorectal cancer. *E.A. Rosenthal[1], B.H. Shirts[2], L. Amendola[1], C. Gallego[1], M. Horike-Pyne[1], P.D. Robertson[3], P.H. Byers[3,4], C. Nefcy[5,6], D. Veenstra[7], F. Hisama[1,8], R. Bennett[1], M.O. Dorschner[2,3,4], D. Nickerson[3,9], D. Crosslin[3], R. Nassir[10], N. Zubair[11], T. Harrison[11], U. Peters[11,12], G.P. Jarvik[1,3,12].* 1) Dept Med Gen, Univ Washington School of Medicine, Seattle, WA; 2) Dept Lab Med, Univ Washington, Seattle, WA; 3) Dept Genome Sciences, Univ Washington School of Medicine, Seattle, WA; 4) Dept Pathology, Univ Washington School of Medicine, Seattle, WA; 5) Dept Biostat, Univ Washington School of Public Health, Seattle, WA; 6) Dept Radiology, Univ Washington School of Medicine, Seattle, WA; 7) School of Pharm, Univ Washington, Seattle, WA; 8) Dept Neurology, Univ Washington School of Medicine, Seattle, WA; 9) Dept Bioengineering, Univ Washington, Seattle, WA; 10) Dept of Biochem and Molecular Medicine, Univ California, Davis, CA; 11) Cancer Prevention Division, Fred Hutchinson Cancer Research Cencer, Seattle, WA; 12) Dept Epidemiology, Univ Washington School of Public Health, Seattle, WA.

The lifetime risk to develop colorectal cancer (CRC [MIM 114500]) is 4.5% in the U.S. Approximately 5% of people with CRC carry an identified pathogenic variant in known causal genes. In another 20%, CRC appears to be inherited but no known pathogenic variant has been detected. This is partly due to lack of evidence to classify variants as pathogenic when they occur in genes that are implicated (GWAS, linkage, or biological pathway information), but not proven, to be associated with CRC. To provide evidence of association for 1143 suspected CRC associated genes, we compared the number of rare (minor allele frequency < 0.005), potentially disruptive variants (PDV) (stop gain (SG), splice acceptor/donor change (SA, SD), and frameshift (FS)) found in 169 CRC cases and 3524 controls. Cases included individuals with unexplained CRC from the Clinical Sequencing Exploratory Research NEXT Medicine study (CSER, N=78), Women's Health Initiative (WHI, N=76), and Northwest Institute of Genetic Medicine Family Polyposis Study (NWIGM, N=15). Controls were selected randomly with respect to CRC from the Exome Sequencing Project (ESP), and were not known to have any Lynch [MIM 120435] associated cancers or to carry a known pathogenic CRC variant. Exome enrichment was performed using Agilent SureSelect All Exon v5 (WHI) and Roche NimbleGen SeqCap EZ v3 (CSER, NWIGM) targets. The analysis pipeline covered similar regions of captured genes and the quality control measures were the same for cases and controls. Average sequencing depth was >50X for all samples. We found a significant association between case status and rare PDV carrier status: 27% of CRC cases carried a rare PDV compared to 5% of controls (p<2e-16, OR 7.5, 95% CI 5.0-11.0). 99 genes had 161 rare PDVs in this data set: 92 SGs, 22 SDs, 15 SAs and 32 FSs. Rare PDVs occurred in cases only for 38 genes, in controls only for 55 genes, and in both for 6 genes. Among cases, there were 11 SGs, 2 SDs, 6 SAs and 28 FSs, 85% of which occurred in case specific genes. Among controls, there were 81 SGs, 20 SDs, 9 SAs and 4 FSs, 92% of which occurred in control specific genes. Case specific genes had ≤2 rare PDVs: those with 2 were *MKL2* and *PMS1*. This study shows the power of aggregate information to find evidence for disease association in a subset of implicated genes.

239

Recurrent acquisition of super enhancer function drives druggable oncogenic expression programs in colorectal cancer. *A.J. Cohen[1], O. Corradin[1], A. Saiakhova[1], C.F. Bartels[1], J.M. Luppino[1], G. Dhillon[1], I.M. Bayles[1], L. Beard[2], L. Myeroff[3], S.D. Markowitz[1,2,3], P.C. Scacheri[1,2].* 1) Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH; 2) Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH; 3) Department of Medicine, Case Medical Center, Case Western Reserve University, Cleveland, OH.

Cancer is a disease of disordered gene expression driven by somatic mutations as well as epigenetic alterations on chromatin. While the relationship between cancer and epigenetic alterations at promoters has been well studied, distal regulatory elements have received less attention. Through ChIP-seq of the characteristic enhancer histone marks, H3K4me1 and H3K27ac, we mapped putative gene enhancer elements across the epigenomes of a cohort of more than forty colorectal cancer (CRC) cell lines, primary CRC tumors, and normal colon epithelium samples, representing the most comprehensive characterization of enhancer elements in a single type of cancer to date. We identified thousands of loci that gained or lost the H3K4me1/H3K27ac-enhancer histone marks in CRC compared to normal colon, which we call variant enhancer loci, or VELs. Of particular interest, we identified a set of highly recurrent VELs gained in nearly all CRC samples. The vast majority of these recurrent VELs are constituents of super enhancers, most of which are present in CRC and absent from normal colon. Gene targets of these recurrent VELs are consistently misexpressed in primary CRC tumors, and include novel and known genes implicated in CRC pathogenesis. Additionally, 40% of all CRC risk loci identified through GWAS contained variants that co-localized with recurrent VELs, strongly supporting their role in CRC pathogenesis. To further test the significance of recurrent VELs, we treated CRC samples with JQ1, a small molecule inhibitor of the BET family of enhancer binding proteins. JQ1 potently inhibited growth of CRC cells both *in vitro* and in xenograft mouse models, and these anti-proliferative effects coincided with downregulation of genes associated with the recurrent VELs. Collectively, our results suggest that in addition to DNA mutations, colorectal cancers are dependent on a signature set of epigenetic alterations at enhancer elements, and that these recurrent VELs and their associated genes may represent promising targets for anti-cancer therapies.

240

Genomic landscape of colorectal tumors shapes the microbiome of the tumor microenvironment. *R. Blekhman[1], M. Burns[1], E. Montassier[2], D. Knights[2].* 1) Genetics, Cell Biology, and Development, University of Minnesota, St. Paul, MN; 2) BioTechnology Institute, Computer Science and Engineering, University of Minnesota, Minneapolis, MN.

Recent studies have found an association between the composition of the gut microbiome and colorectal cancer, the second most diagnosed cancer in the USA. Understanding interactions between colorectal tumors and the microbiome at the tumor microenvironment is critical for elucidating the potentially causal role of bacteria in colorectal cancer, and for development of therapeutics that target the microbiota. To that end, it is especially important to investigate how different tumor genetic profiles affect the tumor's microbial communities. Here, we jointly analyze microbiome and somatic tumor genetic variation to learn about the interactions between tumors and their associated microbiota. We used a combined approach of whole-exome sequencing and microbiome profiling in 44 tumor biopsies and 44 normal colonic tissue samples from the same individual. Using these data, we find strong concordance between tumor mutational patterns and shifts in the microbiome. We show that the number of loss-of-function mutations in tumors is correlated with the diversity of the tumor's microbiota, whereby hypermutated tumors have a significantly more diverse microbiome. Moreover, we show that somatic mutations in certain genes are correlated with changes in abundance of specific microbial taxa, including a link between loss-of-function mutations in *APC* and the abundance of *Fusobacterium*, a previously characterized cancer-associated taxon. We also show similar patterns at the pathway level; for example, we find that loss-of-function mutations in glucose transport pathways in the tumor are strongly correlated in abundance of pathways related to energy utilization in the microbiome, indicating a competition for resources (in this case, glucose) between the tumor and its associated bacterial communities. Lastly, we built a risk index from a panel of microbes correlated with each of several prevalent tumor-driving mutations. We show that we can use this index to accurately predict the existence of loss-of-function mutations in *ZFN717* using solely the tumor's microbiome profile. To summarize, our results highlight a significant and clear interaction between the genetic profile of colorectal tumors and the composition of the tumor-attached microbiome. These results serve as a starting point for colon cancer prognostics and treatments that target and use the microbiome.

241

***PRKACA* defects and adrenal tumors: Human and animal studies and gene dosage effects.** *P. Salpea[1], A. Angelousi[1], B. Yuan[2], F.R. Faucz[1], I. Levy[1], B. Delemer[3], S. Hieronimus[4], B. Feve[5], F. Kelestimur[6], G. Raverot[7], J. Bertherat[8], J.R. Lupski[2], M. Serpe[9], C.A. Stratakis[1].* 1) Program on Developmental Endocrinology and Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, USA; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA; 3) Department of Anatomy and Pathology, University Hospital of Reims, 51092 Reims, France; 4) Department of Endocrinology, Diabetology and Reproductive Medicine, University Hospital of Nice, Nice, France; 5) Centre de Recherche Saint-Antoine, Hôpital Saint-Antoine, Assistance Publique Hôpitaux de Paris, Paris, France; 6) Department of Endocrinology, Erciyes University Medical School, 38039 Kayseri, Turkey; 7) Lyon Neuroscience Research Center, Neuro-Oncology & Neuro-Inflammation Team, Lyon Department of Endocrinology, Groupement Hospitalier Est, Hospices Civils de Lyon, Lyon, France; 8) Centre National de la Recherche Scientifique Unité Mixte de Recherche, Institut Cochin, Université Paris Descartes, Paris, France; 9) Unit on Cellular Communication, Program in Cellular Regulation and Metabolism, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, USA.

Genetic defects that cause abnormal c-AMP dependent Protein Kinase A (PKA) signaling constitute the leading cause for adrenal tumors and ACTH-independent Cushing syndrome (CS). Recently two types of activating defects of the PKA catalytic subunit gene (*PRKACA)* were described leading to abnormal PKA signaling and adrenal tumors. Interestingly, adrenal disease due to *PRKACA* activating defects appears to correlate with *PRKACA* gene dosage and function of the catalytic subunit. We studied DNAs from 25 unilateral sporadic cortisol-producing adenomas (CPAs) that were Sanger-sequenced: 4 (21%) carried the somatic constitutively-activating defect L206R in *PRKACA.* We also screened 35 patients with adrenal hyperplasias (AH) with TaqMan copy number assay and array comparative genomic hybridization (aCGH): 8 of them (22%) had in the germline additional copies of *PRKACA*. We chose the *D. melanogaster* model to test the dosage-dependent effect of *PRKACA*. Previous studies described hypomorphic alleles of the PKA system in flies. We created transgenic flies carrying the various types of PKA main catalytic subunit (*PKA-C1*) activating defects we detected in our patient cohort. The first type of fly contained one additional copy of the *PKA-C1*, these flies had elevated PKA activity but no apparent phenotypic defects. The introduction of two extra *PKA-C1* copies led to phenotypic defects. The second type of transgenic fly had a constitutively active *PKA-C1* (PKA-act) under the control of *phantom-Gal4*, a prothoracic gland (PG) specific promoter. This highly increased PKA activity in the PG (*phm>PKA-act*) effectively blocked the onset of pupariation due to decreased production and/or release of ecdysone, a hormone that controls developmental timing in insects. Our human genetics and animal model studies showed that the dose of PKA main catalytic subunit correlates with the type of adrenal disease and developmental defects in a dose-dependent manner: germline duplication or triplication of the gene leads to AH of correspondingly increasing severity and somatic constitutive activation of the gene leads to CPAs that cause severe CS in humans. In the transgenic *D. melanogaster* models one extra copy of the gene did not cause any detectable phenotypic change but two extra copies caused morphological defects in the adult fly. Finally the highest gene dosage defect was in animals where *PKA-C1* was constitutively expressed in the PG, resulting in complete arrest of developmental progression.

242

CLIC5: a new transcriptional target of ETV6 in childhood acute lymphoblastic leukemia. *B. Neveu[1], C. Richer[1], K. Lagacé[1], P. Cassart[1], M. Lajoie[1], J.F. Spinella[1], D. Sinnett[1,2].* 1) Centre Hospitalier Universitaire Sainte-Justine Research Center, University of Montreal, Montreal, Canada; 2) Department of Pediatrics, Faculty of Medicine, University of Montreal, Montreal, Canada.

The t(12;21) translocation is observed in about 25% of pediatric acute lymphoblastic leukemia (ALL) cases, making it the most common genetic aberration found in this type of leukemia. However expression of the resulting chimeric ETV6-AML1 fusion protein is not sufficient to initiate leukemogenesis, suggesting that additional mutations are required to induce full-blown disease. Loss of the residual ETV6 allele is observed in approximately 75% of t(12;21) positive patients in which case complete inactivation of ETV6 function would promote leukemia onset; the mechanisms through which ETV6 is involved in leukemic transformation remain largely undescribed. Given its role in transcriptional repression, we postulated that loss of ETV6 could result in deregulated expression of downstream target genes and perturb key cellular processes/pathways leading to oncogenesis, but the target genes it regulates remain unknown. In order to investigate the function of ETV6 and identify its transcriptional targets, we performed RNA-seq on the ETV6-AML1 positive ALL cell line REH that overexpress, or not, ETV6. By comparing differential gene expression profiles, we identified 88 potential gene targets whose expression may be directly or indirectly modulated by ETV6. Among the genes that showed significantly decreased expression in the presence of ETV6 was CLIC5, a member of the chloride intracellular channel gene family. Transcriptome analysis performed on 20 leukemia patients also demonstrated that CLIC5 expression is specifically increased in t(12;21) positive cases suggesting that loss of ETV6 and concomitant increased expression of CLIC5 may contribute to leukemia onset within these patients. ChIP experiments revealed an interaction between ETV6 and the proximal promoter of CLIC5, suggesting that CLIC5 is a direct target of ETV6. To further evaluate the functional effects of CLIC5, we generated cell lines overexpressing CLIC5 and showed that overexpression of CLIC5 leads to resistance to oxidative-stress induced apoptosis which may allow for increased mutational burden thereby contributing to leukemogenesis. Functional studies are ongoing to dissect the molecular mechanisms through which ETV6-mediated expression of CLIC5 is involved in childhood ALL. Interestingly, inhibition of ionic channels in cancer has recently emerged as a promising new therapeutic avenue. Accordingly, CLIC5 may prove to be a novel actionable target in t(12;21) positive childhood ALL patients.

243

Towards personalized cellular adoptive immunotherapy targeting tumor specific neo-antigens in microsatellite unstable colorectal cancers. *P. Maby[1], M. Hamieh[1], H. Kora[1], D. Tougeron[2], B. Mlecnik[3], G. Bindea[3], H.K. Angell[3,4], T. Fredriksen[3], N. Elie[5], E. Fauquembergue[1], A. Drouet[1], J. Leprince[6], J. Benichou[7], J. Mauillon[8], F. Le Pessot[9], R. Sesboüé[1], T. Frebourg[1,8], J. Galon[3], J-B. Latouche[1].* 1) Inserm U1079, Rouen University, France; 2) Gastroenterology Department, Poitiers University Hospital and EA 4331, Poitiers University, France; 3) Inserm U1138, Paris Descartes University, Pierre et Marie Curie University, Paris, France; 4) AstraZeneca Pharmaceuticals, Cheshire, UK; 5) Imaging Core Facility, Caen University Hospital, France; 6) Inserm U982, Rouen University, France; 7) Inserm U657, Rouen University, France; 8) Department of Genetics, Rouen University Hospital,France; 9) Department of Pathology, Rouen University Hospital, France.

Colorectal cancers with microsatellite instability (MSI-CRCs) represent 15% of all CRCs and are observed in Lynch syndrome, the most frequent hereditary form of CRC. MSI-CRCs have a higher density of tumor-infiltrating lymphocytes (TILs) than other CRCs. This feature is thought to result from frameshift mutations within coding repeat sequences, leading to the synthesis of neo-antigens expressed as immunogenic neo-peptides recognized by $CD8^+$ T lymphocytes (CD8+ TLs). However, a clear link between $CD8^+$ TIL density and frameshift mutations in MSI-CRCs has yet to be established. With this aim, we screened 103 MSI-CRCs from two independent cohorts for frameshift mutations in 19 genes, using 2 multiplex PCRs, and $CD3^+$, $CD8^+$ and $FOXP3^+$ TIL densities were quantified by immunohistochemistry, using tissue microarrays. We found that $CD3^+$ and $CD8^+$ TIL densities were positively correlated with the total number of frameshift mutations, and that $CD8^+$ TIL density was especially higher when a frameshift mutation was present in *ASTE1*, *HNF1A* or *TCF7L2* gene. We then developed a personalized cellular adoptive immunotherapy strategy based on (1) the characterization of frameshift mutations in the patient tumor and (2) the stimulation of the patient's TLs against neo-peptides derived from these mutations. We constructed Artificial Antigen Presenting Cells (AAPCs), expressing the main costimulatory molecules, B7.1, ICAM-1 and LFA-3, and efficiently presenting a transgene-encoded peptide on the most frequent HLA class I molecule, HLA-A2.1. In the tumor of the first $HLA\text{-}A2^+$ MSI-CRC Lynch patient, we detected a single nucleotide deletion within *TGFBR2*, *TAF1B* and *ASTE1* genes, leading to the putative synthesis of 3 neo-peptides predicted to have a high affinity for the HLA-A2.1 molecule. We cultured this patient's TLs with AAPCs expressing each one of these frameshift mutation-derived peptides. After expansion, activated TLs were able to specifically kill cells, including MSI-CRC tumor cells, presenting the relevant peptides. Then, we performed similar experiments on 2 other MSI-CRC $HLA\text{-}A2^+$ Lynch patients and on 3 $HLA\text{-}A2^+$ control donors. After specific activation with the same AAPCs, only MSI-CRC $HLA\text{-}A2^+$ Lynch patients' peripheral TLs could recognize neo-peptides derived from frameshift mutations present in their tumor. Our results establish a preclinical rationale for developing personalized cellular adoptive immunotherapy strategies to treat MSI-CRCs in Lynch syndrome patients.

244

Functional correction of dwarfism in a mouse model of achondroplasia using the tyrosine kinase inhibitor NVP-BGJ398. *D. Komla Ebri[1], E. Dambroise[1], I. Kramer[2], C. Benoist-Lasselin[1], N. Kaci[1], P. Busca[3], G. Prestat[3], F. Barbault[4], D. Graus-Porta[2], A. Munnich[1], M. Kneissel[2], F. Di Rocco[1], M. Biosse-Duplan[1], L. Legeai-Mallet[1].* 1) INSERM U1163, University Paris Descartes, Sorbonne Paris Cité, Institut Imagine, Paris, France; 2) Novartis Institutes for BioMedical Research, Basel, Switzerland; 3) University Paris Descartes, UMR 8601 CNRS, Paris, France; 4) University Paris Diderot, Sorbonne Paris Cité, ITODYS, UMR CNRS 7086, Paris, France.

Missense mutations localized in the *FGFR3* (*Fibroblast Growth Factor Receptor 3*) gene lead to the most frequent form of dwarfism: achondroplasia (ACH). *FGFR3* mutations induce an increased phosphorylation of this tyrosine kinase receptor causing enhanced activation of its downstream signaling pathways leading to strong endochondral and membranous ossification defects. Nowadays several preclinical studies have been carried out (CNP analog BMN111, PTH injections, soluble FGFR3, statin and meclozine). Since ACH is due to FGFR3 over-activation, the use of selective small molecule Tyrosine Kinase Inhibitors (TKIs) to arrest FGFR3 phosphorylation seems to be a relevant therapeutic approach. For this reason we performed experiments with the recently discovered pan-FGFR TKI "NVP-BGJ398". *In vitro*, NVP-BGJ398 reduced FGFR3 phosphorylation and inhibited FGFR3 downstream signaling pathways, MAPKs and PLCγ, in transfected chondrocytes and ACH chondrocyte lines. Studying in culture femur and calvaria (16.5 and 18.5 dpc) of fetal mutant mice mimicking ACH ($Fgfr3^{Y367C/+}$) we observed that NVP-BGJ398 at 100 nM corrected the abnormal femoral growth plate cartilage and the calvarial defect. Moreover, we demonstrated that a low dose of NVP-BGJ398 (2 mg/kg), was able to penetrate into the growth plate of living $Fgfr3^{Y367C/+}$mice reducing the constitutive phosphorylation of FGFR3 and activation of its downstream signaling pathways. Importantly, improvements of the appendicular and axial skeleton were noticeable after only 15 days of treatment. The size increase was dramatic for femur (21%), tibia (32%), humerus (12%), ulna (22%), radius (24%), tail (27%) and vertebra L4-L6 (12%). Immunohistological analyses showed a significant rescue of the growth plate disorganization and an improvement of the defective proliferation and differentiation in long bones. µCT analyses demonstrated that treatment with NVP-BGJ398 improved skull anomalies, prevented synchondroses loss and changed foramen magnum shape. In addition, we demonstrated for the first time that the cartilage endplate and intervertebral disc of the vertebrae were disturbed by FGFR3 gain-of-function mutation and improved with NVP-BGJ398. In conclusion, most of the hallmarks of ACH, e.g. shortening of the long bone, macrocephaly, kyphosis, reduced foramen magnum were largely improved with NVP-BGJ398. Our findings support the idea that TKIs could be a novel therapeutic tool for improvement of growth in ACH.

245

Potential AAV5 gene therapy for MPS IIIB mice brain. *S.H. Kan, S.Q. Le, Q.D. Bui, P.I. Dickson.* Medical Genetics, LA Biomedical Research Institute, at Harbor-UCLA, Torrance, CA.

Mucopolysacccharidosis (MPS) IIIB is an inherited lysosomal storage disorder characterized by mild somatic features but severe neurologic manifestations with high mortality. MPS IIIB is caused by the deficiency of an enzyme, α-*N*-acetylglucosaminidase (NAGLU) associated with storage of heparan sulfate. Current studies have shown that intracerebroventricular (ICV) enzyme replacement therapy (ERT) with IGF-II fusion protein is a feasible treatment in MPS IIIB mice to correct central nervous system (CNS) phenotypes. It overcomes two impdiments to ERT: the absence of mannose 6-phosphate (M6P) on recombinant human NAGLU (rhNAGLU) and the blood brain barrier.In this study, hNAGLU-IGF-II and hNAGLU cDNA were cloned into an adeno-associated virus (AAV) vector plasmid and the recombinant AAV serotype 5 (rAAV5) was generated by triple transfection. Recombinant AAV5 expressing either hNAGLU-IGF-II or hNAGLU targets the choroid plexus epithelium to express and secrete the missing enzyme (NAGLU) into the cerebrospinal fluid (CSF) of MPS IIIB mice.2.5x10^9 vector genomes (v.g.) of rAAV5 for either constructs or vehicle was administered into both lateral ventricles in MPS IIIB mice at postnatal day 2. Biochemical and histological analyses were performed 10 weeks after rAAV5/vehicle injection and evaluated with control (heterozygous) mice.NAGLU enzyme activity reached 3.68 and 0.13 times the control level in the brain section around the injection site of mice treated by rAAV5-hNAGLU and rAAV5-hNAGLU-IGF2, respectively. β-Hexosaminidase activity, which is elevated in MPS IIIB, was reduced in the rAAV5-rhNAGLU treated mice to the carrier level throughout the brain except cerebellum/brain stem area.Tissue evaluations by immunohistochemistry showed hNAGLU-IGF-II or hNAGLU expression in the choroid plexus epithelium in lateral ventricles; Lamp1 expression was significantly reduced around the injection sites (hippocampus, frontal cortex) in the rAAV5-hNAGLU treated mice. β-Hexosaminidase activity was reduced half way to the normal level, though no detectable NAGLU enzyme activity was measured. Current results suggest that the choroid plexus-targeted viral gene therapy with rAAV5 NAGLU may overcome the major obstacles for ERT with proper M6P modification as a permanent, efficient distribution of NAGLU throughout the brain.

246

Quantitative cell image-based high content screening identifies brain permeable small molecules that rescue peroxisome assembly defects in cells from patients with Zellweger spectrum disorder. *N. Huang[1], P.K. Dranchak[2], A. Flores[1], C. Argyriou[3], R. Luo[3], X. Wang[4], E.N. Oliphant[1], A.B. Moser[5], R. MacAruther[2], C. Hsu[6], J. Inglese[2], N.E. Braverman[3], J.G. Hacia[1].* 1) Department of Biochemistry and Molecular Biology, University of Southern California, Los Angeles, CA, USA; 2) National Center for Advancing Translational Sciences, NIH, MD, USA; 3) Department of Human Genetics, McGill University, Quebec, Canada; 4) Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA; 5) Peroxisome Disease Lab, Hugo W Moser Research Institute, Baltimore, MD, USA; 6) Department of Medical Bioinformatics, University of California, San Diego, CA, USA.

Zellweger spectrum disorder (ZSD) is a disease continuum caused by biallelic defects in a subset of *PEX* genes required for peroxisome assembly and functions. Although ZSD results in high morbidity and mortality, the majority of patients have a milder, but progressive, disease whose clinical manifestations include vision and hearing loss, intellectual disabilities, and liver dysfunction. Recently implemented newborn screening for peroxisomal disorders on a national level provides new opportunities for the early adoption of targeted therapies that address the molecular basis for disease.Nevertheless, current medical interventions are palliative in nature.Here, we developed a quantitative cell image-based high content screening (qHCS) assay for targeted ZSD small molecule therapies that can be conducted in 1536-well format. The HCS assay is based on patient fibroblasts, harboring common hypomorphic missense and null *PEX1* mutations, which express a GFP reporter protein modified with a peroxisome targeting signal (GFP-PTS1). These patient cells show a cytosolic localization of the GFP-PTS1 reporter in contrast to its peroxisomal localization in cells from healthy donors. We previously used this assay in 96-well format to screen >2,000 small molecules at a single concentration and uncovered molecular chaperones, including flavonoids, that rescue peroxisome assembly and functions in patient cells.We used the miniaturized assay to obtain a pharmacological profile for >4,000 compounds, including the entire FDA drug collection, with qHCS, employing seven different concentrations. Using a suite of cell imaging software, including CellProfiler and IN Cell Analyzer, Z-factors above 0.4 were consistently achieved. We validated prior results by identifying the flavonoid apigenin as a bioactive molecule. We also uncovered a novel class of compounds that are active the micromolar range and rescue peroxisome functions in patient cells based on follow-up cell imaging, biochemical, and protein-based assays. They share a structural motif that suggests they could stabilize the structure of the common PEX1-G843D mutant protein by interacting with its ATP-binding domain. Two compounds in this class, naltriben and naltrindole, cross the blood-brain-barrier and have been extensively tested in rodent models of substance abuse and are well-tolerated over extended periods of time. Studies in yeast, plant, fly, and mouse models with patient-specific mutations are ongoing and planned.

247

The Fibrodysplasia Ossificans Progressiva mutation *ACVR1^{R206H}* causes disease by allowing the receptor ACVR1 to respond to Activin A. *A.N. Economides[1,2], S.J. Hatsell[1], V. Idone[1], D.M. Alessi Wolken[1], L. Huang[1], H.J. Kim[1], L. Wang[1], X. Wen[1], K.C. Nannuru[1], J. Jimenez[1], L. Xie[1], G. Makhoul[1], R. Chernomorsky[1], D. D'Ambrosio[1], R.A. Corpina[1], C. Schoenherr[1], K. Feeley[1], H. Nistala[2], P.B. Yu[3], G.D. Yancopoulos[1], A.J. Murphy[1].* 1) Regeneron Pharmaceuticals, Tarrytown, NY; 2) Regeneron Genetics Center, Tarrytown, NY 10591; 3) Brigham & Women's Hospital, Boston, MA 02115.

Fibrodysplasia Ossificans Progressiva (FOP) is a rare genetic disorder characterized by episodically exuberant heterotopic ossification (HO), whereby skeletal muscle is abnormally converted into histologically "normal" bone. This HO leads to progressive immobility with catastrophic consequences, including death by asphyxiation. FOP results from mutations in the intracellular domain of the type I BMP receptor ACVR1; the most common alters Arginine 206 to Histidine (*ACVR1^{R206H}*), and was previously thought to drive inappropriate bone formation due to hyperactivity. We unexpectedly found that – *in vitro* – this single mutation renders ACVR1 responsive to a set of 'non-canonical' ligands, including Activin A, which cannot normally activate this receptor to induce bone formation. To test the implications of this finding *in vivo*, we engineered mice with the *Acvr1^{R206H}* mutation. As mice constitutively expressing Acvr1[R206H] die perinatally, we generated a genetically humanized 'conditional-on' knock-in model for this mutation. When Acvr1[R206H] expression is induced, mice develop HO resembling that in FOP. This process can be inhibited by broad-acting BMP ligand blockers, confirming ligand-dependence. We furthermore show that Activin A induces HO only in mice expressing Acvr1[R206H], and that inhibition of Activin A with a fully human monoclonal antibody completely blocks formation of HO in this model of FOP. Our results indicate that the FOP mutation is due to gain of response to a non-canonical ligand, and that in a physiologic knock-in model of FOP, Activin A is necessary and sufficient for driving HO; thus, our fully human antibody to Activin A presents a potential near-term therapeutic option for FOP.

248

Necroptosis in Niemann-Pick Disease, type C1: A Potential Therapeutic Target. *A.C. Cougnoux[1], C.V. Cluzeau[1], J.M. Picache[1], C.A. Wassif[1], S.M. Cologna[1,2], F.D. Porter[1].* 1) NICHD, bethesda, MD; 2) departement of chemistry, university of illinois at chicago.

Background. Niemann-Pick disease, type C (NPC1) is a lysosomal storage disorder characterized by progressive cerebellar ataxia and dementia. Neurological dysfunction, neuroinflammation, and neuronal loss contribute to the neurological signs and symptoms found in NPC1. Necroptosis is a distinct mechanism of cell death mediated by receptor-interacting proteins kinase 1 and 3 (RIPK1 and RIPK3) and associated with an inflammatory response. **Results**. Based on an observation of increased cell death in cultured NPC1 fibroblasts, we hypothesized that increased necroptosis could contribute to cellular death in NPC1. We subsequently demonstrated significant activation of the necroptotic cell death pathway in both NPC1 fibroblasts and cerebellar tissue from NPC1 patients and mice. Based on these observations, we investigated whether inhibition of RIPK1 would decrease phenotypic signs and increase survival in *Npc1* mutant mice. This hypothesis was tested in BALB/c cNctr-*Npc1$^{mln/mln}$* (*Npc1$^{-/-}$* mice). Control (*Npc1$^{+/+}$*) and *Npc1* mutant mice were treated with Necrostatin 1 (Nec1, an allosteric inhibitor of RIPK1), inactive Necrostatin 1 (Nec1i) or vehicle only (PBS) by intraperitoneal injections every other day starting at 3 weeks of age. We found that treatment with Nec1, but not with Nec1i, resulted in a significant increase in lifespan compared to the vehicle treated control animals: 85±4 ($p<0.01$), 76±6 ($p=0.3$) and 72±4 respectively. In addition to increased lifespan, we also observed delayed progression of neurological signs evaluated by rearing activity scoring. Histopathological analysis demonstrated a two-week delayed cerebellar Purkinje cell loss in mutant animal treated with Nec1.**Conclusion**. These experiments clearly demonstrate that inhibition of necroptosis is effective in ameliorating disease related pathology in the NPC1 mouse model. The increased survival observed with Nec1 treatment, although significant, is rather modest. The limited efficacy is likely due to the pharmacological properties of Nec1, specifically a half-life of approximately 1 hour. This work provides a proof of concept of the involvement of this cell death mechanism in NPC1 pathology, and thus investigation of novel necroptosis inhibitors with improved pharmacological properties in NPC1 is necessary.

249

Clinical, molecular, and metabolomic studies reveal targets for therapy and new mechanisms of pathology in Barth Syndrome. *H. Vernon[1,2], R. Thompson[3], B. DeCroes[4], R. McClellan[2], K. Mercier[5], W. Pathmasiri[5], S. Dhungana[5], J. Carlson[5], S. McRitchie[5], S. Sumner[5], Y. Sandlers[6].* 1) McKusick Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD; 2) Department of Neurogenetics, Kennedy Krieger Institute, Baltimore, MD; 3) Department of Pediatric Cardiology, Johns Hopkins University, Baltimore, MD; 4) Department of Physical Therapy, Kennedy Krieger Institute, Baltimore, MD; 5) RTI International, Research Triangle Park, NC; 6) Department of Chemistry, Cleveland State University.

Barth Syndrome is a rare X-linked recessive disorder of cardiolipin remodeling caused by mutations in the *TAZ* gene on chromosome Xq28, which encodes for the transacylase tafazzin. The cardinal biochemical abnormality in these patients is an elevation in the monolysocardiolipin to tetralinoleyl cardiolipin ratio (MLCL/CL4). Clinical features of Barth syndrome include cardiomyopathy, neutropenia, and skeletal muscle weakness. While the primary biochemical defect in these patients is known, downstream metabolic abnormalities have not been elucidated. Moreover, metabotype (MLCL/CL4)/phenotype/genotype correlations have not been systematically defined. We conducted a detailed, multi-disciplinary study in a cohort of 42 patients with Barth Syndrome in which we investigated cardiac and skeletal muscle characteristics, and then examined their relationship to the genotype and metabotype. We then pursued untargeted metabolomic studies in plasma from a subset of patients to investigate for novel mechanisms of disease pathology and potential new targets for study. Our studies reveal that MLCL/CL4 correlates to both functional exercise capacity and cardiac mass. Additionally, we uncovered a potential association with increased MLCL/CL ratio and older age, implying a progressive or cumulative biochemical defect. We further established a range of genotype/phenotype/metabotype relationships representing the most and least severely affected patients. We then employed untargeted metabolomics analysis via ^{1}H NMR spectroscopy and semi-targeted metabolomic analysis via LC-MS in order investigate for novel discriminating biochemical features between Barth syndrome patients and controls. Multivariate analysis using orthogonal projection to latent structures (OPLS-DA) of the spectroscopy data revealed a clear differentiation between cases and controls, not dependent on the age or body mass index of the subjects. Discriminating metabolites were used for pathway analysis, and involved insulin/glucose regulation, regulation of lipid metabolism, and choline metabolism. LC-MS analysis of a subset of metabolites further clarified specific targets. This work represents the most comprehensive study in a cohort of Barth syndrome patients to date, and establishes metabotype, phenotype, and genotype relationships, targets for clinical and therapeutic monitoring, and novel targets for metabolic pathology.

250

Modulating Ryanodine Receptors by Dantrolene Attenuated Neuropathic Phenotype in Gaucher Disease Mice. *B. Liou[1], V. Inskeep[1], Y. Peng[1], R. Li[1,2], G.A. Grabowski[3], Y. Sun[1,2].* 1) Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 2) Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH; 3) Synageva BioPharma Corp, Lexington, MA.

Gaucher disease is caused by mutations on *GBA1* gene encoding lysosomal acid β-glucosidase (GCase). Defective GCase results in accumulation of substrates, glucosylceramide and glucosylsphingosine that cause pathological insults in the visceral and central nervous systems (CNS). Neuronopathic Gaucher disease (nGD, types 2 and 3) manifests severe CNS symptoms in patients and currently has no effective treatment available. The goal of this study is to determine the role of ryanodine receptors in Gaucher disease and as a potential target for the treatment of nGD. In an nGD cell model CBE (GCase inhibitor) treatment led to accumulation of glucosylceramide and glucosylsphingosine and decreased mitochondrial ATP production and oxygen consumption rate in N2a cells demonstrating nGD pathology. The CBE treated N2a cells showed enhanced calcium release induced by caffeine compared to N2a cells. The calcium release was antagonized by antagonists, either ryanodine or dantrolene, suggesting substrate accumulation mediated ER-calcium release through ryanodine receptors. In the brain of nGD mouse model (4L;C* mice) decreased ryanodine receptors expression were detected in end-stage 4L;C* brain by RNAseq and immunoblot analyses. Block of ryanodine receptors by dantrolene in 4L;C* mice delayed neurological deterioration and prolonged survival. By Gait analysis, the treated 4L;C* mice showed significantly improved gait at 40 days of age compared to the untreated mice. Reduced CD68 positivity in the dantrolene treated 4L;C* brain indicates an attenuation of CNS inflammation. Increased residual GCase activity in dantrolene treated 4L;C* brain also suggested the effect of the treatment on protein folding of mutant GCase. These data demonstrated that maintaining calcium homeostasis through modulating ryanodine receptor function can enhance the activity of mutant GCase and have neuroprotective effect in nGD mouse model. Our results support the hypothesis that deregulated ryanodine receptors play the role in nGD. This study demonstrates that calcium signaling stabilizer, such as dantrolene, could be potential disease modifying therapy for nGD. .

251

A Large-Scale Survey Conducted by the eMERGE Network of Patient Perspectives on Broad Consent and Data Sharing in Biospecimen Research. *M.E. Smith[1], S. Sanderson[2,3], N. Mercaldo[4], A. Antommaria[5], S.A. Aufox[1], M. Brilliant[6], K. Brothers[7], M.B. Claar[8], E.W. Clayton[9, 13], J.J. Connolly[10], P. Conway[11], M. Fullerton[12], N.A. Garrison[13], H. Hakonarson[10,14], C.R. Horowitz[2], G.P. Jarvik[15], D. Kaufman[16], T. Kitchner[6], R. Li[16], E. Ludman[17], C. McCarty[11], J.B. McCormick[18, 19], M. Myers[5], K.E. Nowakowski[19], J. Schildcrout[4], M.J. Shrubsole[20], S. Stallings[8], J.L. Williams[21], S. Ziniel[22], I.A. Holm[23].* 1) Center for Genetic Medicine, Northwestern University, Chicago, IL; 2) Icahn School of Medicine at Mount Sinai, New York, NY; 3) Health Behaviour Research Centre, University College London, London UK; 4) Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN; 5) Cincinnati Children's Hospital Medical Center, Cincinnati OH; 6) Marshfield Clinic Research Foundation, Marshfield, WI; 7) Department of Pediatrics, University of Louisville, Louisville KY; 8) Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN; 9) Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN; 10) Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA; 11) Essentia Institute of Rural Health, Duluth, MN; 12) Department of Bioethics and Humanities, University of Washington, Seattle, WA; 13) Center for Biomedical Ethics & Society, Vanderbilt University Medical Center, Nashville, TN; 14) Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 15) Division of Medical Genetics, University of Washington Medical Center, Seattle, WA; 16) National Human Genome Research Institute, NIH, Bethesda, MD; 17) Group Health Cooperative, Seattle, WA; 18) Division of Health Care Policy and Research, Division of General Internal Medicine, Mayo Clinic, Rochester MN; 19) Biomedical Ethics Program, Mayo Clinic, Rochester, MN; 20) Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN; 21) Genomic Medicine Institute, Geisinger Health System, Danville, PA; 22) Center for Patient Safety and Quality Research, Program for Patient Safety and Quality, Boston Children's Hospital; Division of Adolescent and Young Adult Medicine, Boston Children's Hospital, Department of Pediatrics, Harvard Medical School, Boston, MA; 23) Division of Genetics and Genomics and Manton Center for Orphan Disease Research, Boston Children's Hospital; Department of Pediatrics, Harvard Medical School, Boston, MA.

The Advanced Notice of Proposed Rulemaking (ANPRM) would require consent for research on de-identified human data and specimens, significantly impacting biobank and other research using human biospecimens. Understanding patients' perspectives would be valuable to policy-makers. The eMERGE (Electronic Medical Records and Genomics) Network is an NHGRI-supported national consortium of 11 institutions combining DNA biorepositories with electronic medical record systems for large-scale, high-throughput genetic research. The collaborative nature of eMERGE, plus the diverse patient populations represented, makes the eMERGE Network ideally suited to conduct a large survey. Our primary aims were to: (1) assess respondents' willingness to provide broad consent for sharing of biosamples and data; (2) gain insights into what biospecimen and biobanking-related practices impact willingness to provide broad consent; and (3) compare attitudes towards broad vs. specific consent and open vs. controlled data-sharing models. Respondents were randomly assigned to 1 of 3 hypothetical scenarios: (A) specific consent + controlled data-sharing; (B) broad consent + controlled data-sharing; (C) broad consent + open data-sharing. Survey questions were informed by a systematic literature review, developed by a multi-disciplinary eMERGE workgroup and validated by an iterative cognitive interview process. Measures included willingness to participate in a biobank; willingness to enroll one's child in a biobank; reasons for being un/willing to participate; trust; and attitudes regarding privacy. A stratified sampling approach used census information to oversample individuals with low education, from rural areas, and minorities. The survey was piloted to 1500 individuals to assess the wording of the questionnaire, review response patterns, and provide preliminary summaries on biobank views. The sampling strata classified the 163 survey respondents as 26% Hispanic, 13% African American, 46% from rural areas and 26% with less than a high school education. The survey-adjusted average willingness to participate score (1=not at all, 5=yes, definitely) was 3.5 (95% CI: 2.9-4.1). Modification of the survey instructions were made based on the pilot and the full survey was sent to 90,000 individuals; data is currently being collected. These data may provide recommendations to inform future policy for the ethical conduct of human subjects research, and potential revisions to the Common Rule.

252

Adolescents' Opinions on Disclosure of Non-actionable Secondary Findings in Whole Exome Sequencing. *S.B. Hufnagel[1,2], L.J. Martin[1], A. Cassedy[1], R.J. Hopkin[1], A.H. Antommaria[1].* 1) Cincinnati Children's Hospital Medical Center, Cincinnati OH 45229; 2) Children's National Medical Center, Washington DC 20011.

When predictive genetic testing for Huntington Disease was developed, a consensus emerged that it should not be offered to asymptomatic minors in order to allow them to decide for themselves as adults. Evidence of significant harm of testing was not produced and some began to argue that testing provided psychological benefit. This consensus has been further challenged by whole exome sequencing, which has the potential to identify secondarily adult onset conditions that are not medically actionable in childhood. While the literature suggests that most parents want access to this type of results, little information is available about adolescents' views. This study's goal is to examine adolescents' views about this topic including which arguments about disclosure they consider most and least compelling. We conducted a cross-sectional study of adolescents in grades 7-12 who attend one of three Cincinnati public schools. One investigator presented background information on genetic testing and then administered a survey. No parents opted their child out of participation (n=282). Sixty-three percent of the participants identified themselves as Black or African American. Most participants (83%) preferred to know non-actionable results. The most common reason for wanting to know (39%) was future planning. Reasons for not wanting to know were endorsed at comparable rates. Seventy-two percent of participants believed they should be able to make this decision themselves (19%) or jointly with their parents (53%). Further, 73% of participants believed that parents of children less than 12 years old should have access to non-actionable results. With the exception of age, options were not associated with demographics. Children under 12 were less likely to want non-actionable findings (55% vs. 87%, $p<0.0001$) and to believe that parents of children under 12 should have access to them (likert response, strongly agree of 5, median 4 versus 5, with a $p<0.0001$). There was a modest correlation (Spearman rho = 0.28, $p<0.0001$) between increasing age and the belief that participants were capable of making the decision at their current age. These results demonstrated that the majority of adolescents surveyed desire to have access to and participation in decision-making about secondary genetic findings for adult onset conditions that are not medically actionable in childhood. This study articulates previously ignored stakeholders' preferences, information essential to guide policy.

253

Responses of Primary Care Physicians to Unsolicited Secondary Findings about Lynch Syndrome. *K.D. Christensen[1], M.T. Scheuner[2], J.E. Garber[3], H.L. Rehm[1,4], R.C. Green[1,5].* 1) Department of Medicine, Division of Genetics, Brigham & Women's Hospital and Harvard Medical School, Boston, MA; 2) VA Greater Los Angeles Healthcare System, Los Angeles, CA & David Geffen School of Medicine, University of California, Los Angeles, CA; 3) Dana Farber Cancer Institute, Boston, MA; 4) Laboratory for Molecular Medicine, Partners HealthCare Personalized Medicine and the Broad Institute, Cambridge, MA; Department of Pathology, Brigham & Women's Hospital and Harvard Medical School, Boston, MA; 5) Partners HealthCare Personalized Medicine, Cambridge, MA.

Background: Guidelines for genomic sequencing recommend the disclosure of certain secondary findings (SFs) that are unrelated to the original purposes of testing, yet may have important implications for patient health. Primary care physicians (PCPs) may have substantial responsibilities to respond to genomic SFs because of their roles in managing the overall clinical care of their patients.Methods: We invited members of the American Academy of Family Physicians to complete a web-based survey. Physicians were randomized to review a genomic sequencing report about 3 Lynch Syndrome SFs that were reported as benign, variant of uncertain significance (VUS), or pathogenic. Physicians provided open-ended responses about the clinical follow-up they would request. They also ranked how they would prioritize their response to the SF in comparison to five other tasks common to PCP appointments.Results: Of 2,000 invited PCPs, 149 (7.5%) completed the survey (78% non-Hispanic white, 55% female, mean age 50). The majority of PCPs reported that they would explain the results to their patients, although they were less likely to do so in response to a benign variant than a VUS or a pathogenic variant (75% vs 91% vs 100%, respectively, overall $p<0.001$). Responding to SF reports was prioritized higher if reports showed a pathogenic variant than a VUS or a benign variant (2.7 vs 3.6 vs 4.3, respectively, on 1-6 scale, $p<0.001$). PCPs anticipated discussing referrals more often after reviewing a pathogenic variant than a VUS or benign variant (77%, 66%, and 16%, respectively, all pairwise comparisons $p<0.03$). PCPs who reviewed either a pathogenic variant or a VUS were more likely than PCPs who reviewed a benign variant to anticipate discussing patients' personal or family histories of cancer (39% vs 7%, $p<0.001$), cancer screening (36% vs 9%, $p<0.001$), further workup (28% vs 2%, $p<0.001$), and testing of patients' relatives (23% vs 2%, $p<0.001$), with no differences observed between anticipated responses to pathogenic variants and VUSs (all $p>0.46$).Conclusions: How PCPs manage genomic SFs will depend on the variants' reported pathogenicity. Disclosure of benign SFs is unlikely to lead to further clinical workup. However, PCPs may respond to VUSs and pathogenic variants in similar ways. Findings suggest that unless laboratories set stringent criteria for SF disclosure, genomic sequencing may substantially increase demands for healthcare services as its use becomes more common.

254

The return of whole exome sequencing results in a pediatric cancer setting: What is being said? *S. Scollon[1,2], L.B. McCullough[3], K. Bergstrom[1,2], R.A. Kerstein[1,2], D.W. Parsons[1,2], S.E. Plon[1,2], R.L. Street Jr.[4].* 1) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 2) Texas Children's Cancer Center, Texas Children's Hospital, Houston, TX; 3) Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX; 4) Department of Communication, Texas A&M University, College Station, TX.

Background: Patient-centered communication, defined as providing clear, understandable information with active patient participation, can be especially challenging in the return of genome-scale sequencing results such as whole exome sequencing (WES) due to the volume, difficulty and complexity of information, and what actions, if any, may be taken based on the findings. There has been little research on physician, genetic counselor (GC) and parent communication patterns regarding disclosure of WES results. Methods: The Baylor College of Medicine Advancing Sequencing in Childhood Cancer Care (BASIC3) study examines the clinical utility of tumor and germline WES in the pediatric oncology setting. Results are disclosed to families by the participant's oncologist (n=10) and one of two study GCs. These visits (n=90 encounters) were audio-recorded, transcribed and analyzed using a previously validated coding system to code for clinician and parent communication behaviors. The unit of analysis for the coding of this portion of data is the utterance, the oral analogue of a sentence. In addition, 20 encounters identified to have a diagnostic or incidental finding on the germline report and another 20 randomly selected encounters without such findings were further analyzed for themes related to content of parental questions. Results: The primary findings were: (a) oncologists varied greatly in how they chose to utilized GC services during the encounters (b) GCs made more recommendations based on the WES findings than did oncologists; (c) clinicians gave more information when there were significant findings; (d) clinicians engaged in more partnership-building and supportive talk after they gained more experience presenting WES reports; and (e) checking for parent understanding was infrequent. On average less than 20% of the conversation was contributed by parents. However, this participation varied with parents and family member utterances ranging from 3 to 167 per disclosure when discussing the germline reports. Conclusions: Oncologists vary significantly in use of GCs to disclose WES results and may benefit from interventions to improve checking for understanding of WES results and partnership-building. Further analysis of the specific content of parental questions may provide insight into topics of importance to families, which can guide future research towards improving communication of WES results. Supported by NHGRI/NCI 1U01HG006485.

255

Communication and Management of Genomic Sequencing Results by Non-Geneticist Physicians. *J. Krier[1], C. Blout[1], D. Lautenbach[2], J. Vassy[1,3], J. Oliver Robinson[4], M. Helm[1], K. Lee[4], M. Murray[5], R. Green[1,6].* 1) Division of Genetics, Brigham and Women's Hospital, Boston, MA; 2) Illumina, Inc., San Diego, CA; 3) VA Boston Healthcare System, Boston, MA; 4) Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX; 5) Geisinger Health System, Danville, PA; 6) Partners HealthCare Personalized Medicine, Boston, MA.

Background: Experts have raised concerns about the ability of non-geneticist clinicians to understand and communicate genome sequencing (GS) results, but little data exist to support or contradict these fears. The MedSeq Project explores the incorporation of GS into clinical medicine by enrolling primary care physicians (PCPs) and cardiologists to return GS results on interpreted genome reports (GR) to their patients, without the mediation of genetic counselors (GCs). **Methods**: In the MedSeq Project, 11 PCPs and 9 cardiologists and 204 of their patients were enrolled, with patients randomized to receive either GR and family history reports (FHR) or FHR alone. The GR include potentially actionable variants in genes associated with monogenic diseases, recessive carrier states, and polygenic risk estimates for 8 cardiovascular phenotypes. At study start, PCPs completed a 6-hour genomics education module. Before patient enrollment, physicians were asked about their anticipated use of the Genome Resource Center (GRC), staffed by GCs and clinical geneticists, to support their understanding of the GR and subsequent decision-making. Interviews were consensus-coded and analyzed and GRC utilization tracked. The GRC reviewed GR disclosure session transcripts for the accuracy of genetic information conveyed and understanding of genetic concepts underlying clinical decisions. If communication or management concerns were identified, physicians were contacted immediately or at the end of the study depending on the urgency and potential safety impact. **Results:** Of 18 physicians interviewed, 13 articulated an intention to utilize the GRC. Six physicians have contacted the GRC in the first 28 months of the study. Of the first 45/100 GS disclosure transcripts reviewed, 20 minor miscommunications were noted for end-of-study educational follow-up, and 3 for real-time intervention (2 related to carrier risk, 1 to inheritance pattern misinterpretation). No immediate patient safety issues were identified. Analysis of the remaining 55 GS disclosure transcripts will be completed in the coming months. **Conclusion**: While the counseling provided by non-geneticist physicians in the MedSeq Project contained some mistakes, serious errors have not occurred. Physicians acknowledged the value of the GRC, but only a minority utilized it. When provided with genetic education and resources to easily query experts, non-genetics providers may be able to correctly and safely convey GS results.

256

Impact of genome sequencing on the medical care of healthy adults. *J.L. Vassy[1,2], K.D. Christensen[1], D. Dukhovny[3], C.L. Blout[1], J. Oliver Robinson[4], J.B. Krier[1], M.F. Murray[5], A.L. McGuire[4], R.C. Green[1,6] for the MedSeq Project.* 1) Brigham and Women's and Harvard Medical School; 2) VA Boston Healthcare System; 3) Oregon Health and Science University; 4) Baylor College of Medicine; 5) Geisinger Health System; 6) Partners HealthCare Personalized Medicine.

Background: Genome sequencing (GS) in healthy asymptomatic adults might enable early disease prevention and detection. Alternatively, it might precipitate a cascade of costly medical interventions that do not improve health outcomes and may even cause harm. To quantify the risks and benefits of GS in healthy adult patients, we measured its impact on their primary care physicians' (PCP) clinical decision-making, concordance with appropriate prevention guidelines, and downstream healthcare utilization and costs. **Methods:** In the MedSeq Project, 9 PCPs from one health system were enrolled to participate with their patients (n=100) in a clinical trial of GS in primary care. Eligible patients were 40-65 years old and deemed generally healthy by their PCPs. Patients were randomized to receive a family history report only (FHx arm) or a FHx report plus an interpreted GS report (FHx+GS arm). The GS reports included any potentially clinically relevant variants in ~4600 genes associated with monogenic disease, recessive carrier states, and polygenic risk estimates for 8 cardiometabolic traits. The PCP and patient met to discuss the reports and any next steps in clinical management, and the PCP completed a survey identifying any clinical action taken as a result of the patient's reports. Six months after each disclosure visit, medical chart review was used to abstract data on utilization and concordance with U.S. Preventive Services Task Force (USPSTF) Grade A and D guidelines. Healthcare costs were determined from billing data and Medicare price weights. We compared outcomes in the 2 arms with non-parametric tests.**Results:** After 95 results discussions to date, PCPs report ordering significantly more cardiac tests for FHx+GS patients vs. FHx patients (7 vs. 0, p=0.01), but not more lab tests (7 vs. 3, p=0.49), imaging studies (2 vs. 0, p=0.49), or referrals (6 vs. 6, p>0.99). Six-month chart review of the first 41 patients suggests the arms do not differ in the proportions concordant with USPSTF guidelines for preventive care, including colorectal cancer screening and aspirin use, or in total imaging tests (15 vs. 13, p=0.46) or specialty visits (54 vs. 33, p=0.43). Median 6-month costs are $226 in the FHx arm and $616 in the FHx+GS arm (p=0.25). **Conclusions:** Introducing GS to the care of healthy adults impacts PCPs' immediate clinical decision-making but might not change concordance with prevention guidelines or increase downstream healthcare utilization and costs.

257

Short-term costs of integrating genome sequencing into clinical care: preliminary results from the MedSeq Project. *D. Dukhovny[1], K.C. Christensen[2], J.L. Vassy[2,3], D.R. Azzariti[4], C. Lu[5], H.L. Rehm[2,4,6], A.L. McGuire[7], R.C. Green[2,4,6] for the MedSeq Project.* 1) Oregon Health & Science University, Portland , OR; 2) Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 3) VA Boston Healthcare System, Boston, MA; 4) Partners HealthCare Personalized Medicine, Boston, MA; 5) Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; 6) The Broad Institute of MIT and Harvard, Cambridge, MA; 7) Baylor College of Medicine, Houston, TX.

Genome sequencing (GS) has the potential to offer significant health benefits to patients and their families, but there is apprehension about the healthcare costs of GS implementation into routine clinical practice. The MedSeq Project is a clinical trial examining the impact of GS in the care of healthy patients and patients with cardiomyopathy in primary care and cardiology settings. Patient participants are randomized to receive either review their family history alone or family history plus GS analysis. Financial costs being examined prospectively include: cost of the disclosure appointments (based on the length of sessions); DNA sequencing (market price); variant confirmation, interpretation, and reporting (based on time demands for lab personnel); and pre-disclosure counseling and consent for participants randomized to GS (based on time demands for a registered nurse). Additional follow-up healthcare costs are being assessed over a 6-month time horizon from a third party payer perspective using medical record data and price weights from Medicare reimbursement schedules.To date, 73 patient participants (49% randomized to GS, 59% from primary care, mean age 55, 51% female, 86% non-Hispanic white) of an expected 200 total participants have completed the study through the six-month follow-up. Disclosure sessions that reviewed GS results were 15 minutes longer, on average, than sessions that reviewed only family history information (34.5 min vs 19.6 minutes, respectively, p<0.001). The mean total cost per patient within Partners HealthCare System was greater for patients randomized to receive GS than for patients randomized to receive family history analysis alone ($8,163 vs $510, respectively, p<0.001). This difference was less, and did not reach statistical significance, when considering only the costs of disclosure and follow-up care ($980 vs $510, p=0.11). Preliminary results suggest that integrating GS into clinical care compared to family history alone will increase short-term healthcare costs. The potential health benefits that accrue from GS analyses over a longer time horizon may justify those costs, although they must also be weighed against potential for over-utilization of health care resources due to pursuit of information of unknown clinical significance. Future analyses on MedSeq Project data will examine the impact of GS on health-related quality of life, patient out-of-pocket expenditure and overall cost-effectiveness of GS.

258

Incorporation of whole genome sequencing results into the electronic medical record: Attitudes of MedSeq Project participants. *C.L. Blout[1], J.O Robinson[2], A.L. McGuire[2], P.M. Diamond[3], K.D. Christensen[1,4], L. Jamal[2], R.C. Green[1,4,5,6] for the MedSeq Project.* 1) Division of Genetics, Brigham and Women's Hospital, Boston, MA; 2) Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston TX; 3) Division of Health Promotion and Behavioral Sciences, University of Texas School of Public Health, Houston TX; 4) Harvard Medical School, Boston, MA; 5) The Broad Institute of MIT and Harvard, Cambridge, MA; 6) Partners Personalized Medicine, Boston, MA.

Background: Electronic medical record (EMR) usage has increased steadily over the last decade. Optimizing the clinical utility of whole-genome sequencing (WGS) will require incorporation of genomic data into the EMR, but patients may be concerned about insurance discrimination or privacy issues. Little is known about patient attitudes around the inclusion of WGS data into EMRs. Our objective was to assess the attitudes of MedSeq Project participants about the incorporation of WGS results into their EMRs. Methods: The MedSeq Project is a randomized, controlled trial exploring the use of WGS in cardiology and primary care. 202 adult participants were randomized to receive family history alone or family history plus WGS. WGS results were uploaded into participants' EMRs. Participants were surveyed about attitudes and comfort level regarding the storage of WGS data in their medical record prior to learning their randomization status and 6 weeks after disclosure (6wk). Non-parametric Wilcoxon signed-rank tests were used to assess changes to participants' attitudes over time. Results: Participants were on average 55 years old, mostly non-Hispanic white (87%), and educated (81% ≥ college degree); 64% had annual household income ≥ $100,000. At baseline 57% of participants agreed that genetic information should be part of a standard medical record, while 17% disagreed. Analyses from the first 123 participants who completed a 6wk follow-up survey showed an increase in the percentage of participants who agreed that genetic information should be part of the EMR compared to baseline (66% vs 57%, p=0.05). Participants who received WGS were more comfortable with their genetic information going into their EMR at 6wk compared to baseline (77% very comfortable at 6wk vs 52% at baseline, p=0.03). Similarly, participants who received WGS were more likely to prefer to have all of their genetic information in their EMR at 6wk compared to baseline (63% vs 41%, respectively, p<0.001). Conclusion: Most MedSeq Project participants supported genetic information inclusion into EMRs, and attitudes were more favorable at 6wk post disclosure compared to baseline. After receiving WGS, more participants were comfortable with genetic information being incorporated into their EMR compared to baseline, suggesting patient willingness to embrace incorporation of genetic information into EMRs despite the previously mentioned concerns.

259

Contributions of "healthy genomes" to expand our understanding of Mendelian conditions. *D.L. Perry, J. Kakishita, A. Khouzam, E. Thorpe, E. Ramos, V.M. Raymond, A. Chawla, A. Livengood, S. Chowdhury, T.M. Hambuch.* Illumina, Inc., San Diego, CA.

Background: The Illumina Clinical Services Laboratory (ICSL) is one of the first laboratories to offer clinical whole genome sequencing (cWGS) to healthy individuals. The cWGS testing includes predisposition and carrier screening for Mendelian conditions.**Methods:** Individuals were evaluated as part of a cWGS test that included 1,600 genes associated with 1,221 monogenic conditions (n=443) or as part of an expanded test that encompasses 1,691 genes associated with 1,232 monogenic conditions (n=95). Evidence was evaluated by a team of geneticists and genetic counselors and variants were classified according to the American College of Medical Genetics and Genomics guidelines. Clinical reports were issued to the ordering physician in accordance with CLIA/CAP regulations.**Results:** Of the 538 adults who obtained cWGS since May 2012, 189 (35%) had variants classified as pathogenic or likely pathogenic (P/LP) which are expected to be clinically significant (heterozygous for a dominant condition, homozygous or compound heterozygous for a recessive condition). Of these 189 individuals, 27 (14%) had P/LP variants associated with highly penetrant disorders (e.g. Neurofibromatosis, Type 1). Furthermore, 176 individuals (93%) had P/LP variants associated with conditions of low to intermediate penetrance. Of these, more than half (59%) had P/LP variants in one of the following five conditions: Mannose-Binding Protein Deficiency (15%), Factor V Leiden Thrombophilia (15%), Hereditary Hemochromatosis (11%), Prothrombin-Related Thrombophilia (11%) and Familial Periodic Fever (7%).**Implications:** Over one third of presumably healthy adults who underwent cWGS for the purpose of predisposition and carrier screening were found to have P/LP variants in Mendelian genes with potential personal health implications. While this proportion may seem high, the vast majority of variants returned were associated with relatively common conditions with known low to intermediate levels of penetrance. Importantly, 27 of 538 (5%) individuals screened had variants associated with highly penetrant disorders. These results contribute to broadening our understanding of genetic variation in Mendelian disorders within presumably healthy adults in the general population. This information has the potential to redefine our understanding of the phenotypic spectrum of inherited syndromes that were previously thought to be well-characterized and provide insight regarding the use of cWGS in precision medicine.

260

Large scale analysis of interracial differences in genetic polymorphisms of *CYP2C9, CYP2C19, CYP2D6, CYP3A4*, and *CYP3A5* in the U.S. Population. *H. Yao, J.M. Harrington, T. Hsieh, T.Y. Jacobson, M.S. Shumaker, J.A. Byers, M. Nakano, C.J. Sailey, W. Mo.* Molecular Testing Labs, Vancouver, WA.

Approximately 70% of all prescribed drugs are metabolized by five common cytochrome P450 enzymes: CYP2C9, CYP2C19, CYP2D6, CYP3A4, and CYP3A5. A variety of DNA polymorphisms have been identified in genes encoding these enzymes, causing significant inter-individual variations in enzymatic activities and hence responses to medications. Frequencies of polymorphisms in the U.S. population have been challenging to characterize due to the country's large racial diversity and the cost associated with analysis. At Molecular Testing Labs, a total of 65,338 patients have undergone Phamacogenetics(PGx) testing, of which 35,578 patients have provided their ethnic background. By utilizing this large data set, we determined the interracial allele frequencies of the *CYP2D6, CYP2C9, CYP2C19, CYP3A4,* and *CYP3A5* genes in the U.S. population. For CYP2D6, as an example, we determined the frequencies of *2, *3, *4, *5, *6, *7, *8, *9, *10, *11, *15, *17, *29, *41, and *Xn. We observed that the *3 and *4 alleles are more common in Caucasians (1.54% and 19.30% respectively); *17 and *XN are more common in African Americans (16.81% and 11.10%); while *10 is more common in Asians (47.18%). Consequently, phenotypes of CYP2D6 metabolism are also different across races. Poor metabolizers (PM) in the Caucasian population are 6.5%, while the prevalence of PMs in the Asian population is only 0.56%. In CYP2C19, the highest PM phenotype observed is in Asians (13.3%), while PMs in Caucasians are relatively low (2.22%). Our results also determined rare variant frequencies in the U.S. population for variants with no previously reported frequency. These rare alleles are unevenly distributed across different racial backgrounds. For instance, while the frequency of CYP2C19*3 alleles is only 0.23% in the overall population, its frequency is 5.93% in Asians and 4.52% in Hawaiian/ Pacific Islander populations. Similarly, the frequency of CYP2D6*6 is 0.86% and 1.12% in the general population and in Caucasians respectively. Interestingly, many rare alleles are more prevalent in African Americans than other racial groups, such as CYP2D6 *15, 2C9 *11, *5, *6 and CYP2C19*9. In conclusion, our large scale analysis has provided novel insight into the frequencies of rare polymorphisms for these five important CYP genes. .

261

Protein Truncating Mutations in the *ARID2* Gene Are Associated with a Novel Neurodevelopmental Disorder. *B.E. Friedman[1], L. Shang[2], M.T. Cho[2], K. Retterer[1], L. Folk[1], J. Humberson[3], L. Rohena[4], A. Sidhu[5], S. Saliganan[5], A. Iglesias[2], P. Vitazka[1], J. Juusola[1], W.K. Chung[2,6].* 1) GeneDx, Gaithersburg, MD, USA; 2) Department of Pediatrics, Columbia University Medical Center, New York, NY, USA; 3) Department of Pediatrics, Division of Genetics and Metabolism, University of Virginia, Charlottesville, VA, USA; 4) Department of Pediatrics, Division of Genetics, San Antonio Military Medical Center, San Antonio, TX, USA; 5) 5Department of Pediatrics and Human Development, Michigan State University, East Lansing, MI, USA; 6) Department of Medicine, Columbia University Medical Center, New York, NY, USA.

Whole exome sequencing (WES) is a powerful genomic tool that can be used to identify novel molecular causes of disorders with multiple etiologies. In this study of 970 probands with neurodevelopmental disorders, a clinical WES approach identified *de novo*, loss-of-function variants in the *ARID2* gene in four individuals. WES data were evaluated using a custom analysis pipeline for alignment, variant calling and annotation, and interactive filtering of WES variants. In this cohort, we identified four unrelated probands with overlapping clinical features who were each heterozygous for a loss-of-function variant in the *ARID2* gene. The four probands ranged in age from six to 15 years. All exhibited global developmental delays, cognitive delay, hypotonia, and behavioral issues, most notably attention deficit hyperactivity disorder. All patients also had short stature and dysmorphic facial features including micrognathia or retrognathia, low set or posteriorly rotated ears, epicanthal folds, down-slanting palpebral fissures, high arched palate, and frontal bossing. Wormian bones were observed in two of the patients. The variants included two novel frameshift and two novel nonsense variants, not present in any of the available population databases, and predicted to lead to protein truncation. Three of these variants were confirmed to be *de novo*, and parents of the fourth proband were unavailable for testing. ARID2 is one of three ARID proteins in SWI/SNF subunits and is an intrinsic part of the PBAF complex, which is evolutionarily well conserved and is involved in the regulation of gene expression, embryogenesis, organ development, and tumorigenesis. The ATP-dependent SWI/SNF chromatin modifier has been previously implicated in neurodevelopmental disorders including Nicolaides-Baraitser syndrome, Coffin-Siris syndrome, autism, and schizophrenia. Our data delineate the clinical and molecular features of a novel neurodevelopmental disorder characterized by intellectual disability, behavior abnormalities, short stature, and dysmorphic features due to *de novo*, loss-of-function variants in the *ARID2* gene. This study highlights the value of WES for gene discovery in clinical diagnostic testing, especially for clinically and genetically heterogeneous neurodevelopmental disorders.

262

Molecular Diagnoses of Acute Hepatic Porphyrias: Comparisons of Mutation Positive Results for Various Physician Specialties. *H. Naik[1], D. Doheny[1], J. Overbey[2], R. Srinivasan[1], R.J. Desnick[1], Porphyrias Consortium of the NIH Rare Diseases Clinical Research Network.* 1) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY.

The three autosomal dominant Acute Hepatic Porphyrias (AHPs) and their causative genes are Acute Intermittent Porphyria (AIP; *HMBS*), Hereditary Coproporphyria (HCP; *CPOX*), and Variegate Porphyria (VP; *PPOX*). Diagnosing the AHPs can be difficult due to the rarity of the diseases, nonspecific presentation of symptoms, non-specialist physicians' selection of inappropriate tests, and misinterpretation of disease-specific biochemical testing. A retrospective analysis of mutation-positive diagnoses for suspected AHP patients whose samples were sent to the Mount Sinai Genetic Testing Laboratory between 2005 and 2015 was performed. The ordering physician's specialty was identified and positive diagnostic rates were determined for physicians by specialty and porphyria experts. Blood samples from probands were tested for pathogenic mutations including small and large deletions by dosage analysis for either a specific AHP gene or the three AHP gene panel, which was requested for ~25% of samples received. Of the 2644 individuals tested, 324 (12%) were positive for a pathogenic mutation, including: 23% of 1100 tested for *HMBS;* 5% of 755 tested for *PPOX*, and 3% of 790 tested for *CPOX*. For AIP the most frequent referring specialties were: Geneticists (G=21%), Porphyrias Consortium experts (PCE=19%), Internists (I=14%), Hematologists (H=11%), Family Practitioners (FP=10%), Pediatricians (P=5%), Neurologists (N=3%), and Gastroenterologists (GI=3%). Positive diagnoses were modeled using logistic regression with the specialty as the dependent variable. Odds ratios were computed using PCEs as the reference group. Positive diagnostic rates by specialty were PCE=58%, G=24%, GI=17%, I=16%, P=14%, H=12%, N=11%, and FP=8%. The odds of making a positive diagnosis compared to PCE were G=0.22, GI=0.14, I=0.13, P=0.12, H=0.10, N=0.09, and FP=0.06. Porphyrias Consortium experts, as expected, and second Geneticists had significantly higher odds of making a positive diagnosis than the other specialties. The PCE diagnostic rate may be even higher as they often use genetic testing to exclude an AHP diagnosis. That few samples were from GI and <1% from Emergency Medicine physicians was surprising, as AHPs present with excruciating abdominal pain. These results identify specialties to target for educational programs on the diagnosis/management of AHPs.

263

Utility of whole genome sequencing for detection of newborn screening disorders in a population cohort of 1696 neonates. *D.L. Bodian[1], E. Klein[1], R.K. Iyer[1], W.S.W. Wong[1], P. Kothiyal[1], D. Stauffer[1], K.C. Huddleston[1], A.D. Gaither[2], I. Remsburg[2], A. Khromykh[1], R.L. Baker[3], G.L. Maxwell[4], J.G. Vockley[1], J.E. Niederhuber[1], B.D. Solomon[1,5].* 1) Inova Translational Medicine Institute, Inova Health System, Falls Church, VA; 2) Inova Children's Hospital, Inova Health System, Falls Church, VA; 3) Fairfax Neonatal Associates, Inova Health System, Falls Church, VA; 4) Dept of Obstetrics and Gynecology, Inova Health System, Falls Church, VA; 5) Dept of Pediatrics, Inova Health System, Falls Church, VA.

Purpose: To assess the potential of whole genome sequencing (WGS) to replicate and augment results from conventional blood-based newborn screening (NBS).Methods: Research-generated WGS data from an ancestrally diverse cohort of 1,696 infants and both parents were analyzed for variants in 163 genes involved in disorders included or under discussion for inclusion in United States NBS programs. WGS results were compared to results from state NBS and related follow-up testing.Results: NBS genes are generally well-covered by WGS. There is a median of 1 (range 0-6) database-annotated pathogenic variant in the NBS genes per infant. Results of WGS and NBS in detecting 28 state-screened disorders and 4 hemoglobin traits were concordant for 88.6% of true positives (n=35) and 98.9% of true negatives (n=45,757). Of the 5 infants affected with a state-screened disorder, WGS identified 2 whereas NBS detected 4. WGS yielded fewer false positives than NBS (0.037% vs. 0.17%) but more results of uncertain significance (0.90% vs. 0.013%).Conclusion: WGS may help rule-in and rule-out NBS disorders, pinpoint molecular diagnoses, and detect conditions not amenable to current NBS assays.

264

Diagnostic Utility of Whole Genome Sequencing as an Alternative to Chromosomal Microarray Analysis in Pediatric Medicine. *D.J. Stavropoulos[1], R. Jobling[2], D. Merico[3], S. Bowdin[2], N. Monfared[4], M.S. Meyn[2,5,11], M. Girdea[5], M. Szego[3,6], R. Zlotnik-Shaul[7], B. Thiruvahindrapuram[3], G. Pellecchia[3], T. Nalpathamkalam[3], M. Brudno[5,8], C. Shuman[2,5,11], R. Hayeems[9], C. Carew[4], R. Erickson[10], R.A. Leach[10], P.N. Ray[1,3,4,5,11], R.D. Cohn[2,4,5,11], S.W. Scherer[3,4,5,11], C.R. Marshall[1].* 1) Genome Diagnostics, Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, ON, Canada; 2) Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, ON, Canada; 3) The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada; 4) Centre for Genetic Medicine, The Hospital for Sick Children, Toronto, ON, Canada; 5) Program in Genetics and Genome Biology, Hospital for Sick Children, Toronto, Ontario, Canada; 6) Joint Centre for Bioethics, University of Toronto, Toronto, ON, Canada; 7) Department of Bioethics, The Hospital for Sick Children, Toronto, ON, Canada; 8) Department of Computer Science, University of Toronto, Toronto, ON, Canada; 9) Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, ON, Canada; 10) Complete Genomics Inc, Mountain View, CA, USA; 11) Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada.

Chromosome microarray analysis (CMA) is the current standard as a first tier genetic test for those individuals presenting with developmental delay and/or congenital abnormalities. However, up to 90% of patients who undergo CMA do not obtain a genetic diagnosis leading physicians to seek out other forms of molecular genetic testing. Whole genome sequencing (WGS) promises to capture all classes of genetic variation in a single test, but the diagnostic yield of WGS compared to CMA in the clinical setting has not been established. We established the SickKids Genome Clinic to address this and other questions related to pediatric genomic medicine. As part of this multidisciplinary project, we prospectively performed WGS (Complete Genomics) on 100 consecutive patients referred to a pediatric genetics service with clinical indication(s) for CMA. In 32% of cases, WGS identified causative variants for the primary reason of referral; a 4-fold increase in diagnostic yield over CMA (8%) alone and >2-fold increase compared to CMA plus targeted molecular testing (15%). WGS identified all rare reportable CNVs that were detected by CMA including *de novo* pathogenic CNVs affecting 4p16.3 and 22q11.2 associated with Wolf-Hirschhorn and 22q11.2 microdeletion syndromes, respectively. In an additional 24 patients, WGS revealed clinically significant sequence level variants presenting in a dominant (68%; including variants in EP300, GDF5, PIK3R2, PACS1, CCM2, SPTAN1) or a recessive (32%; including variants PANK2, LARP7, TSEN54 and NGLY1) manner. Similar to previous whole exome sequencing studies we found that 4% of cases had variants in at least two genes involved in distinct genetic disorders, contributing to a more complex clinical phenotype. Clinical implementation of WGS as a single genetic test will provide a higher diagnostic yield than conventional testing while decreasing the number of genetic tests, and ultimately the time before reaching a genetic diagnosis in a pediatric population.

265

Panel testing for familial breast cancer: tension at the boundary of research and clinical care. *I. Campbell[1], E. Thompson[1], S. Rowley[1], N. Li[1], S. McInerny[2], L. Devereux[1], M. Wong-Brown[3], A. Trainer[1,2], M. Mitchell[1,2], R. Scott[3,4], P. James[1,2], Lifepool.* 1) Research Div, Peter MacCallum Cancer Ctr, East Melbourne, Australia; 2) Familial Cancer Centre, Peter MacCallum Cancer Centre, East Melbourne, Australia; 3) The University of Newcastle and Hunter Medical Research Institute, Newcastle, Australia; 4) Hunter Area Pathology Service, Newcastle, Australia.

Gene panel sequencing is revolutionizing germline risk assessment for hereditary breast cancer. Despite scant evidence supporting the role of many of these genes in breast cancer predisposition, results are often reported to families as the definitive explanation for their family history. We assessed the frequency of mutations in 18 genes commonly included in hereditary breast cancer panels among 2,000 index cases from breast cancer families and 1,997 population controls. Cases were predominantly breast cancer-affected women referred to specialized familial cancer centers (BRCA1 and BRCA2 wild-type). Controls were cancer-free women from the LifePool study (www.lifepool.org). Sequencing data were filtered for known pathogenic or novel loss of function mutations.The frequency of pathogenic mutations in BRCA1 and BRCA2 in the control group was 0.2% (4 mutations) and 0.4% (8 mutations), respectively, which is consistent with previous indirect estimates for Caucasian populations but to our knowledge this is the largest direct assessment of their prevalence.Excluding 18 mutations identified in BRCA1 and BRCA2 among the cases and controls, a total of 69 cases (3.5%) and 26 controls (1.3%) were found to carry an “actionable mutation”. PALB2 was most frequently mutated (22 cases, 3 controls), while no mutations were identified in PTEN or STK11. Among the remaining genes, loss of function mutations were rare with similar frequency between cases and controls. The frequency of mutations in most breast cancer panel genes among individuals selected for possible hereditary breast cancer is low and in many cases similar or even lower than that observed among cancer-free population controls. While multi-gene panels can significantly aid in cancer risk management, they equally have the potential to provide clinical misinformation and harm at the individual level if the data is not interpreted cautiously.

266

Yield of Pathogenic/Likely Pathogenic Variants in Women with Breast Cancer Undergoing Hereditary Cancer Panel Testing. *L.M. Andolina, R. Nusbaum, L.R. Susswein, S.R. Solomon, K.S. Hruska, R.T. Klein.* GeneDx, Gaithersburg, MD.

BACKGROUND: Most studies investigating hereditary predisposition to breast cancer have focused on *BRCA1/2*. The availability of hereditary cancer panel testing via next-generation sequencing has allowed for broader testing of genes predisposing to cancer risk. However, data regarding the diagnostic yield of panels in women with breast cancer have been sparse. We sought to determine the frequency of pathogenic variants/likely pathogenic variants (PV/LPVs) in hereditary cancer genes in women with breast cancer. Current cancer panel offerings at GeneDx include combinations of high-risk, moderate-risk and newer genes. High-risk genes are clinically well-characterized, confer a significantly increased risk for cancer and have published management guidelines. Moderate-risk genes are generally associated with a 2- to 3-fold increased risk of cancer and have published management guidelines. Newer genes have been identified in familial cancer cases, but lifetime cancer risks have not been robustly determined. METHODS: We reviewed the results of cancer panel testing for 9773 women with breast cancer tested at GeneDx. Panel tests included analysis of up to 29 genes associated with hereditary cancer risk. Single heterozygous *MUTYH* PV/LPVs were excluded because *MUTYH*-associated polyposis is a recessive disorder. RESULTS: In total, 909 PV/LPVs were identified in 883 women (9.1%); 23 women had more than one PV/LPV. PV/LPVs were identified in *CHEK2* (222), *BRCA2* (153), *BRCA1* (130), *ATM* (104), *PALB2* (79), *BRIP1* (28), *PMS2* (20), *FANCC* (19), *RAD51C* (19), *TP53* (19), *NBN* (18), *PTEN* (18), *MSH6* (17), *BARD1* (15), *RAD51D* (13), *CDH1* (8), *MLH1* (7), *XRCC2* (6), *MSH2* (4), *MUTYH* (4 PV/LPVs in 2 women), *APC* (3), *AXIN2* (1), *CDKN2A* (1), *VHL* (1). Of all PV/LPVs, 385 (42.4%) were detected in high-risk genes, 405 (44.6%) in moderate-risk and 119 (13.1%) in newer genes. CONCLUSION: Overall, 283 (31.1%) of PV/LPVs were identified in *BRCA1/2*, while the majority of PV/LPVs were identified in non-*BRCA1/2* genes and would not have been detected by testing only *BRCA1/2*. Most of the non-*BRCA1/2* PV/LPVs (507/626; 81.0%) were detected in high-or moderate-risk genes with published management guidelines. Inherited cancer panels beyond *BRCA1/2* should be considered as an initial test for women with breast cancer who are being evaluated for hereditary cancer risk, as our data demonstrate that panel tests identified PV/LPVs with published clinical guidelines in 768 patients in our clinical series (7.9%).

267

Extremely high resolution 3D maps of human and mouse genomes across lineages and during differentiation reveal principles of chromatin looping. *S. Rao[1,2,3,4,10], M. Huntley[1,2,3,4,5,10], N. Durand[1,2,3,4], E. Stamenova[1,2,3,4], I. Bochkov[1,2,3], J. Robinson[1,4], A. Sanborn[1,2,3,6], I. Machol[1,2,3], A. Omer[1,2,3], E. Lander[4,7,8], E. Lieberman Aiden[1,2,3,4,9].* 1) The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; 3) Department of Computer Science, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA; 4) Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA; 5) School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA; 6) Department of Computer Science, Stanford University, Stanford, CA 94305, USA; 7) Department of Biology, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA; 8) Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA; 9) Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA; 10) Co-first author.

We use *in situ* Hi-C to probe the three-dimensional architecture of genomes, constructing haploid and diploid maps of ten cell types. The densest, in human lymphoblastoid cells, contains 4.9 billion contacts, achieving 1-kilobase resolution. We find that genomes are partitioned into contact domains (median length, 185kb), which are associated with distinct patterns of histone marks and segregate into six subcompartments. We identify ~10,000 loops. These loops frequently link promoters and enhancers, correlate with gene activation, and show conservation across cell types and species. Loop anchors typically occur at domain boundaries and bind CTCF. CTCF sites at loop anchors occur predominantly (>90%) in a convergent orientation, with the asymmetric motifs "facing" one another. We examine murine differentiation from the embryonic stem cell state and find that loops and domains vary in a developmentally regulated manner and regulate key developmental genes, such as *SOX2*. The inactive X-chromosome splits into two massive domains and contains large loops anchored at CTCF-binding repeats. This work was funded by NSF grants DGE0946799 and DGE1144152, an NIH New Innovator award (1DP2OD008540-01), an NIH CEGS (P50HG006193), an NVIDIA Research Center award, an IBM University Challenge Award, a Google Research Award, a Cancer Prevention Research Institute of Texas Scholar Award (R1304), a McNair Medical Institute Scholar Award, the President's Early Career Award in Science and Engineering, and an NHGRI grant (HG003067).

268

Identifying the transcription factors mediating enhancer–target gene regulation in the human genome. *Y.-C. Hwang[1,2], P.P. Kuksa[2,3], B.D. Gregory[1,4], L.-S. Wang[1,2,3].* 1) Genomics and Computational Biology Graduate Group, University of Pennsylvania Perelman School of Medicine; 2) Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine; 3) Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine; 4) Department of Biology, University of Pennsylvania, Philadelphia, PA.

The majority of genetic variations associated with disease or trait phenotype reported by GWAS are located in non-coding regions. One class of the non-coding elements is enhancer elements, which can regulate gene expression through bindings of transcription factor complexes and form long-range interactions with the protein-coding promoters. To identify all possible enhancers and the genes they regulate genome-wide, we developed a novel method and reanalyzed the latest Hi-C datasets of human cells (GM12878) with ultra-high read depth (~3.0B reads) for physical DNA–DNA interactions. We discovered the GC-content of the Hi-C reads are slightly higher than the genomic background, suggesting that there could be regulation roles of the long-range interactions. Unlike the standard binning approach for mapping chromatin interactions, our analysis aims at identifying the precise loci of the DNA-interacting sites. The proposed model utilizes multiple sources of information including (1) read strandness; (2) read distances to the closest restriction sites; and (3) constraints on DNA ligations, to more accurately delineate borders of DNA-interacting regions. We further called Hi-C peaks for the matched borders that harbor higher Hi-C reads than expected. The Hi-C peaks are on average 978 b.p. in length and covered 51.4% of the genome. 77% of the CTCF binding sites are covered by the Hi-C peaks, indicating the DNA-interacting sites are associated to enhancer–promoter interactions and insulator sites. Additionally, 86% of the Hi-C peaks are covered with known open chromatin regions, while 61% of the open chromatin regions are covered by Hi-C peaks. This suggests that the DNA-interacting sites have the binding affinity for transcription regulation but not all open chromatin regions are involved in long-range regulation. By applying Fit-Hi-C, a spline-based fitting model calling significant DNA–DNA interactions by the linear genetic distance between two peaks, we recovered 13,812 significant intra-chromosomal interactions. We then identified Hi-C peaks as enhancers with the following criteria: (1) pair with an annotated promoter element; (2) intersect with sites having known enhancer-associated histone modifications (H3K27ac, H3K4me1); and (3) reside in an open chromatin site. We further discovered the motifs that are involved in the DNA interactions and revealed the transcription factor complexes that may suggest the underlying mechanism of long-range regulations.

269

Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissue(s)-of-origin. *M.W. Snyder, M. Kircher, A.J. Hill, J. Shendure.* Genome Sciences, University of Washington, Seattle, WA.

Nucleosomes are the basic unit of packaging of eukaryotic chromatin, and nucleosome positioning can differ substantially between cell types. Previous studies of nucleosome positioning in humans, typically performed by digesting native chromatin with exogenous factors such as micrococcal nuclease, have revealed global, albeit weak, sequence specificity, and have identified stereotyped positioning of nucleosomes around genomic features including insulators and promoters. However, these studies have produced modest numbers of nucleosome calls, typically limited to preferentially open genomic regions, and have suffered from poor concordance of calls between studies. Here, we exploit the origins of circulating, plasma-borne cell-free DNA (cfDNA) in healthy individuals to generate a dense, genome-wide map of *in vivo* nucleosome occupancy. These >11 million (M) nucleosome positions, whose locations and spacings correlate with features of chromatin organization and gene structure, span >2 gigabases (Gb) of the human reference genome. We also show that short cfDNA fragments – poorly recovered by standard protocols – reveal footprints of *in vivo* occupancy by DNA-bound transcription factors including CTCF. We demonstrate that analysis of nucleosome positioning around cell type-specific markers, including DNase-I hypersensitive sites and transcription start sites, can recapitulate existing hypotheses about the tissue-of-origin of the dominant population of cfDNA fragments. Intriguingly, by comparing fragmentation patterns around cell type-specific features in samples derived from healthy individuals to those from a panel of individuals with advanced cancers, we show that the in vivo nucleosome footprint observed in a single individual can be used to infer the tissues or cell types contributing to their circulating cfDNA. These results suggest that analysis of cfDNA fragmentation patterns may represent a new methodology for early and noninvasive detection of some cancers, and possibly for classifying cancers of unknown origin. Furthermore, as this strategy is independent of genotypic differences or DNA methylation profiles, we anticipate that the in vivo footprints of protein-DNA interactions revealed by cfDNA might enable the noninvasive monitoring of a much broader set of clinical conditions than is currently possible.

270

Discovery of dendritic cell sub-populations in human blood by single cell RNA-sequencing. *A.C. Villani[1,2], R. Satija[1,3], C. Ford[1], M. Griesbeck[4], W. Li[1,2], P. De Jager[1,5], A. Regev[1,6], N. Hacohen[1,2].* 1) Broad Institute of MIT and Harvard, Cambridge, MA; 2) Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, MA; 3) New York Genome Center, New York, USA; 4) The Ragon Institute of MGH, MIT and Harvard, Cambridge, MA; 5) Brigham and Women's Hospital, and Harvard Medical School, Boston, MA; 6) Department of Biology, Massachusetts Institute of Technology, and Howard Hughes Medical Institute, Cambridge, MA.

Following-up a decade of successful disease susceptibility loci identification, the next challenge remains translating these findings to biological understanding of disease. Identifying cells in which these loci are expressed and analyzing individual cells' roles in health and disease are critical to this endeavor. Dendritic cells (DC) play a critical role in a host's response to pathogens and in the immune responses characterizing cancer, inflammatory and infectious diseases. DC subsets have historically been defined through surface marker analysis. To discover subtypes more unbiasedly, we used single cell RNA-seq (scRNA-seq) to profile the transcriptome of 1056 single human blood DCs isolated from a healthy individual. While supervised analysis of DC markers effectively classified the 4 known populations (BDCA2$^+$ (*IL3RA*), BDCA1$^+$ (*CD1C*), BDCA3$^+$ (*CLEC9A*), CD16$^+$ (*FCGR3A*)), unsupervised analysis re-discovered all 4 subsets through a 512-discriminative gene signature in addition to highlighting novel heterogeneity within subsets. For examples, the BDCA1$^+$ population clearly subdivides into 2 subsets (MHC class II- and inflammatory-prominent subset), a result clarifying unexplained heterogeneity of BDCA1$^+$ in disease. Cross-referencing the 512-gene set with susceptibility loci identified specific subsets contributing to disease, with, for example, *TCF7L2* (type II diabetes) uniquely expressed in CD16$^+$ and *CARD11* (atopic dermatitis and ulcerative colitis) in BDCA2$^+$. Multi-dimensional classification analysis of the first 384 single cells sequenced identified 26 outlier cells not clustering with any of the 4 known subsets. These outliers displayed a unique expression signature and a shared signature with BDCA2$^+$ and BDCA3$^+$ lineages. Using cell surface markers identified by scRNA-seq, we sorted and validated the existence of these cells in 10 additional healthy donors, showing they represent 0.06% of the PBMCs population in the blood and 2-3% of the DC populations across all 10 donors tested. Isolation, *in vitro* experimentations and profiling of an additional 600 single cell outliers enabled a deeper characterization of their phenotype. These cells show similarity with blastic plasmacytoid dendritic cell neoplasm profile, a rare cancer with undefined ontogenic origin. Together these analyses provide a comprehensive view of the DC landscape in blood, contributing to elucidating the origins of pathogenic cells and nominating relevant cell subsets for functional studies.

271

Mapping Expression Quantitative trait loci to Identify Insulin resistance, Obesity and Type 2 Diabetes genes in African Americans. *S. Das[1,4], S. Sajuthi[2,4], N. Sharma[1], J. Chou[2], J. Calles[1], J. Demons[1], S. Rogers[1], L. Ma[1,4], N. Palmer[3,4], D. McWilliams[2,4], J. Beal[2,4], M. Comeau[2,4], K. Williams[1], L. Menon[1], E. Kouba[1], D. Davis[1], J. Byers[1], M. Burris[1], S. Byerly[1], L. Easter[1], D. Bowden[3], B. Freedman[1,4], C. Langefeld[2,4].* 1) Internal Medicine, Wake Forest School of Medicine, Winston-Salem, NC; 2) Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC; 3) Biochemistry,Wake Forest School of Medicine, Winston-Salem, NC; 4) Center for Public Health Genomics, Wake Forest School of Medicine, Winston-Salem, NC.

Type 2 Diabetes (T2D) and its risk factors, including insulin resistance and obesity are more prevalent in African Americans (AAs) compared to European Americans. Distinct architecture of genetic variants that modulate transcript abundance in insulin responsive tissues may explain the higher prevalence of T2D and population specific characteristics of glucose homeostasis in AAs. However, published eQTL studies in African ancestry populations are restricted to blood cells and lymphoblasts.To address this issue, we integrated quantitative measures of insulin sensitivity (S_I, evaluated by FSIGT), gene expression in adipose and muscle tissue (Illumina HT12-V4), and genotype (IlluminaOmni5+Exome) data in 260 non-diabetic AAs from North Carolina.Expression of many transcripts (n=2212 at FDR<0.01) in adipose and fewer transcripts (n=145) in muscle were associated with S_I. Genes enriched in some pathways (e.g. eIF2, eIF4-p70S6K, mTOR signaling) were modulated with S_I in both tissues, while genes in other pathways showed tissue-specific (e.g. leukocyte extravasation signaling in adipose) or discordant regulation between tissues (e.g. oxidative phosphorylation). Co-expression network analysis also indicated tissue-specific modulation of transcript modules with S_I.The eQTL analysis identified 1971 and 2078 cis-eGenes (associated SNP within ±500Kb at a FDR<0.01) in adipose and muscle, respectively. Cis-eQTLs for 885 transcripts including top cis-eGenes *CHURC1*, *USMG5* and *ERAP2* (FDR<$1x10^{-100}$) were identified in both tissues. Most of the top cis-eSNPs (62.1%) were located within ±50Kb of the transcription start site and 43.1% were intronic. Among these cis-eGenes, 363 and 42 were associated with S_I in adipose and muscle, respectively. The *TINP1/NSA2* (p=$3x10^{-67}$) in adipose and the *SEC61G* (p=$2x10^{-25}$) in muscle were the strongest cis-eGenes among S_I-associated transcripts. Cis-eSNPs for *NINJ1*, *AGA* and *CLEC10A* were also associated with S_I (p<0.001) in this cohort.GWAS database (NHGRI) mining and meta-analysis results from Caucasians and AAs identified association of cis-eSNPs for many genes with T2D (e.g. *PIK3C2A*, *RBMS1*, *UFSP1*), fasting and 2hr-glucose (e.g. *INPP5E*, *SNX17, ERAP2*), HbA1c (e.g. *FN3KRP*), BMI and other obesity traits (e.g. *POMC*, *CPEB4*).In summary, this study delineates molecular mechanisms of insulin resistance and provides a unique map of genetically regulated transcripts in AAs which is critical for identifying the genetic etiology of T2D and related traits.

272

Inferring causal relationships between gene expression and complex traits using Mendelian randomization (MR). *Y. Park[1], I. McDowell[2], G. Gliner[3], B.F. Voight[1,4,5], B.E. Engelhardt[6], C.D. Brown[1], Genotype Tissue Expression (GTEx) Project Consortium.* 1) Department of Genetics, Perelman School of Medicine University of Pennsylvania, Philadelphia, 19104, PA, USA; 2) Department of Computational Biology and Bioinformatics, Duke University, Durham, NC, 27705, USA; 3) Department of Operations Research and Financial Engineering, Princeton University, NJ, 08540, USA; 4) Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, 19104, USA; 5) Institute of Bioinformatics, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, 19104, USA; 6) Department of Computer Science and Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08540, USA.

Over the past decade, genome wide association studies (GWAS) have identified thousands of loci associated with complex traits, including common diseases. Due to inherent limitations of GWAS, such studies typically cannot identify causal variants, nor do they demonstrate whether the variant results in a gain or loss of function, information that is necessary for therapeutic intervention. The majority of these variants are noncoding, leading to the hypothesis that changes in gene expression lead to changes in disease risk. An efficient means of prioritizing disease-associated loci and predicting their mechanism of action would be beneficial to the community. Recently, MR has emerged as a framework to test for causal relationships between biomarkers and disease while reducing the influence of reverse causality or confounding due to unmeasured covariates. However, to date, MR has largely been applied to a limited number of epidemiologically measured biomarkers and diseases. To address this limitation and to leverage thousands of genetic variants associated with gene expression (eQTLs), we have implemented a tool to analyze genome-wide eQTL data and publicly available GWAS summary statistics in an MR framework. As a proof of principle, we have applied MR to a large meta-analysis of blood serum metabolites from the Global Lipid Genetics Consortium using liver eQTLs previously identified by our group. We identified several known and novel associations with LDL-C levels including *SORT1*, *SLC44A2*, *ST3GAL4* and *ANGPTL3*. For example, a one standard deviation change of *SORT1* expression increases LDL-C levels by 6.5 mg/dl ($p=1x10^{-100}$), resulting in an increased risk for coronary heart disease (OR=1.14, $p=1x10^{-9}$), replicating the previously described mechanism of *SORT1*. We have also applied MR to eQTLs identified by the Genotype Tissue Expression (GTEx) Consortium with respect to cardiometabolic traits. The breadth of cell types to be analyzed by GTEx will improve the comprehensiveness of this analysis and assist in prioritizing cell types relevant for disease. Application to GTEx pilot data identified several additional genes meriting further investigation including *APOB* (LDL-C levels, $p=4x10^{-11}$) and *G6PC2* (fasting glucose levels, $p=1x10^{-36}$) in subcutaneous adipose tissues. In conclusion, we propose that GWAS and eQTL data integration through MR is able to prioritize candidate genes and quantify the sensitivity of organismal phenotypes to changes in gene expression.

273

Detection and interpretation of genome structural variation in GTEx samples. *C. Chiang[1], R.M. Layer[2], R.P. Smith[1], A.J. Scott[1], A.B. Wilfert[3], . The GTEx Project Consortium[4], D.F. Conrad[3], I.M. Hall[1,5].* 1) McDonnell Genome Institute, Washington University School of Medicine; 2) Department of Human Genetics, University of Utah; 3) Department of Genetics, Washington University School of Medicine; 4) The Genotype-Tissue Expression (GTEx) Project Consortium; 5) Department of Medicine, Washington University School of Medicine.

Structural variation (SV) is a broad class of genome variation that includes copy number variants (CNVs), balanced rearrangements and mobile element insertions. SV is recognized to be an important source of human genetic diversity – 5,000-10,000 SVs are detectable in the typical human genome using short-read DNA sequencing technologies – but little is known about the mechanisms through which SVs affect gene expression and phenotypic variation. The availability of deep whole genome sequencing (WGS) and RNA expression data in the GTEx cohort offers an unprecedented opportunity to address this question. Here, we describe our work aimed at comprehensive detection and interpretation of structural variation in GTEx samples. We first developed an improved pipeline for SV detection in large WGS cohorts that is fast, accurate, scalable to thousands of genomes, and produces multi-sample callsets that are on par with traditional joint variant calling approaches. This pipeline is loosely based on our SpeedSeq software and is composed of four stages: 1) SV discovery on each individual genome using the LUMPY algorithm, 2) SV integration across all samples to produce a unified, cohort-level VCF of spatially refined breakpoints, 3) SV breakpoint genotyping with SVTyper, and 4) read-depth copy number annotation with CNVnator. We applied these methods to 149 GTEx WGS datasets to generate a unified, cohort-level VCF of 49,730 SVs along with copy number and genotype annotations. Using this dataset, we mapped SV eQTLs in 13 tissues analyzed by RNA-seq, resulting in 1,596 SVs affecting the expression of 1,801 genes, including 506 genes that were not identified by eQTL mapping using SNVs and indels alone. We will present the results of our ongoing analysis of the tissue specificity and directionality of these expression effects with respect to SV class and overlap with known regulatory elements, and our efforts to identify and study causal SVs. We further describe analyses aimed at discerning the contribution of difficult-to-identify SVs that are not typically included in functional studies including rare variants, complex variants, multi-allelic CNVs, and mobile element insertions.

274

Mapping genetic and epigenetic factors influencing human hippocampal gene expression. *A. Hofmann[1], H. Schulz[4], A.-K. Ruppert[4], S. Herms[1,2], K. Pernhorst[5], C. Wolf[6], N. Karbalai[6], D. Czamara[6], A.J. Forstner[1], A. Woitecki[5], B. Pütz[6], A. Hilmer[7], N. Fricker[1], H. Vatter[5], B. Müller-Myhsok[6], M.M. Nöthen[1], T. Sander[4], A. Becker[5], P. Hoffmann[1,2,3], S. Cichon[2,3].* 1) Institute of Human Genetics, Department of Genomics, Life & Brain Center, University of Bonn, Germany; 2) Forschungsgruppe Genomics; Medizinische Genetik; Departement Biomedizin; Universitätsspital Basel, Switzerland; 3) Institute of Neuroscience and Medicine; Research Center Juelich, Germany; 4) Cologne Center for Genomics; University of Cologne, Germany; 5) Department of Neuropathology, University of Bonn Medical Center, Germany; 6) Statistical Genetics, Max Planck Institute of Psychiatry, Munich, Germany; 7) Genome Institute of Singapore, Singapore.

Genome-wide association studies have detected multiple loci associated with psychiatric disorders. The majority of these disease-associated variants are observed in noncoding regions and their functional effects are usually unclear. It is often suspected that at least part of these variants influence the expression of neighboring or distant genes. Novel methods allow to systematically investigate the regulatory effects of genetic variants by screening the genome for correlations between allelic variants and gene expression (expression Quantitative Trait Loci / eQTLs) or DNA methylation (meQTLs) in a tissue of interest. Several studies have investigated the occurrence of eQTLs and meQTLs in human brain tissue. A major problem is the quality of the available brain eQTL and meQTL data as they are usually derived from *post-mortem* brain tissue and the overlap of significant eQTLs and meQTLs between these studies is relatively low.We employed 150 *fresh frozen* hippocampal biopsy samples derived from surgery of patients with chronic pharmaco-resistant temporal lobe epilepsy and performed genome-wide SNP genotyping, expression and methylation profiling. After stringent quality control, 4,250,386 imputed SNPs, 16,023 transcripts and 346,656 CpG islands were correlated in 115 brain samples using a linear regression model implemented in *matrixEQTL*.At a false discovery rate (FDR) threshold of 5%, we detected 2,545 significant eQTLs and 102,506 meQTLs. 34% *cis* eQTLs overlap to *cis* meQTLs. Over 50% of eGenes overlap to those reported by the largest brain eQTL meta-analysis in human cortex, while less than 20% was shared between hippocampus and blood. An enrichment for distinct chromatin state annotations i.e. active enhancers defined by the Epigenomics Roadmap Consortium was observed. GWAS hits from the public NHGRI GWAS catalogue were overrepresented in both eQTLs and meQTLs.In this study, we present an integrative large-scale functional genomic analysis to explore the effects of common DNA sequence variants on DNA methylation and mRNA expression. In contrast to all published studies, our samples were collected from *fresh frozen* and not *post-mortem* brain tissue. Therefore, the identified eQTLs and meQTLs provide an extremely valuable resource for functional annotation of SNPs and will help guiding the interpretation of GWAS hits in genetically complex brain disorders.

275

Identification of Major Genetic Modifiers of Vascular Disease in Marfan Syndrome Mice. *A. Doyle[1,2], J. Doyle[2,3], R. Wardlow[2], N. Wilson[2], D. Bedja[4,5], M. Lindsay[6], J. Habashi[2,7], L. Myers[2], K. Braunstein[8], S. Bachir[2], N. Huso[2], O. Squires[2], B. Rusholme[2], A. George[2], M. Caulfield[1], D. Judge[4], H. Dietz[2,7].* 1) William Harvey Research Institute, Queen Mary University of London, London, UK; 2) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, and Howard Hughes Medical Institute, Baltimore, MD, USA; 3) Wilmer Eye Institute, Johns Hopkins Hospital, Baltimore, MD, USA; 4) Department of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 5) Australian School of Advanced Medicine, Macquarie University, Sydney, Australia; 6) Massachusetts General Hospital Thoracic Aortic Center, Departments of Medicine and Pediatrics, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; 7) Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 8) Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

Many of the manifestations of Marfan syndrome (MFS), including aortic aneurysm, are caused by deranged TGFβ activity, which can be attenuated in MFS mice by systemically administering TGFβ neutralizing antibody, the angiotensin II receptor blocker Losartan or RDEA119 - an inhibitor of the mitogen activated protein kinase (MAPK) ERK. In the present study, we sought to identify genetic modifiers of disease pathology using MFS (*Fbn1*$^{C1039G/+}$) mice backcrossed more than 10 generations onto C57BL/6J (BL6/MFS) and 129S6 (129/MFS) background strains. In comparison to mice from a mixed background, the aortic root size, aortic root growth and rate of aortic dissection were greatly abrogated in BL6/MFS mice, but markedly accentuated in 129/MFS animals in association with increased activation of both the canonical (Smad2/3) and noncanonical (ERK and p38) TGFβ signaling cascades. All aortic disease parameters were rescued with Losartan or RDEA119 treatment. 129/MFS mice also showed significantly worse lung emphysema and spinal kyphosis, indicative of deleterious systemic genetic modification. Wild-type mice from each background strain showed no difference in any parameter, indicating MFS disease-specific modification. We generated a large multi-generational pedigree of intercrossed BL6/MFS and 129/MFS mice. Genome mapping revealed 2 QTLs on chromosomes 5 and 11 that strongly linked with the severe aortic phenotype (LOD=4.76 and 4.78) and reached genome wide significance (p=0.008 for both), with evidence of epistasis between loci (MfLOD=12.8; p=0.0006). We are currently functionally characterizing strain-specific genetic variation within these loci, with early prioritization of a stop-loss mutation in *Mmp17* (rs29636438; p.X579W) and a predicted deleterious missense mutation in *Map2k6* (rs51129320; p.G76E) in the 129S6 background. Informatively, our preliminary studies suggest that the *Map2k6* G76E allele associates with increased steady-state protein levels and function, as assessed by TGFβ-induced MAPK activation. Furthermore, in parallel work, we found that functional variation in MAP3K4, which is immediately upstream of and activates MAP2K6, protectively modifies human MFS. Thus, a confluence of discovery-based and hypothesis-driven methodologies has informed disease pathogenesis for MFS and has potentially defined a pathway of protective disease modification; this provides added confidence regarding the potential of therapeutic targets that leverage nature's success.

276

Novel genetic modifiers of retinitis pigmentosa identified by exploiting natural variation in *Drosophila*. *C.Y. Chow[1,2], M. Wolfner[1], A.G. Clark[1].* 1) Dept Molec Biol & Gen, Cornell Univ, Ithaca, NY; 2) Dept Human Genetics, University of Utah, Salt Lake City, UT.

Retinitis pigmentosa (RP) is characterized by progressive loss of vision due to degeneration of rods and cones in the retina. Autosomal dominant retinitis pigmentosa (ADRP) makes up 30-40% of RP. Dominant mutations in the *rhodopsin* gene (*RHO*) comprise 25% of all ADRP cases and represent the most common cause of RP. Dominant mutations in the *Drosophila* ortholog of *RHO, Rh1*, provide a valuable model for rapidly dissecting the pathophysiology of ADRP. The pathogenicity of the $Rh1^{G69D}$ mutation in *Drosophila* resembles many human mutations in *RHO*. $Rh1^{G69D}$ results in a misfolded protein that is retained in the endoplasmic reticulum (ER), and induces the ER stress response, leading to apoptotic cell death and retinal degeneration. Mutations in genes in the ER stress and apoptosis pathways alter the phenotypic presentation of $Rh1^{G69D}$. However, previous genetic studies relied on loss-of-function (LOF) mutations. In the human population, it is unlikely that severe LOF mutations contribute appreciably to variability in ADRP. We take advantage of natural variation in *Drosophila* to identify modifiers of $Rh1^{G69D}$. We crossed the $Rh1^{G69D}$ mutation into 200 strains from the *Drosophila* Genetic Reference Panel (DGRP). The DGRP is a collection of wild-derived *Drosophila* strains that harbor polymorphisms present in a natural population. To assess the effect of DGRP backgrounds in modulating the phenotypic impact of $Rh1^{G69D}$, we measured eye size to quantify the extent of degeneration. We found that degenerative eye size varied by >3 fold across strains. We performed an association study to identify natural polymorphisms that modify the primary $Rh1^{G69D}$ retinal degeneration. We identified candidate genes that have never been implicated in modifying *Rh1* mutations. These novel candidates include genes involved in ER stress response (*CG2004*), involved in apoptosis (CDK5), are known to modify the Rh1 protein (*HEXA*), and are implicated in other human retinal degenerative diseases (*BBS9* and *CECR1*). Strikingly, we also found enrichment for genes in Notch signaling (*KIRREL, MAP4K1*, and *SHANK1*). We tested all candidate genes by RNAi knockdown and more than half strongly suppressed or enhanced the original $Rh1^{G69D}$ retinal degeneration. The genes identified from this study appear to have excellent biological support to reflect potential human modifiers of ADRP. These results have important implications for identifying modifiers of human ADRP and provide putative targets for therapy.

277

Identification of a novel mutation for Perrault syndrome in the mitochondrial rRNA chaperone ERAL1. *A.S. Plomp[1], I.A. Chatzispyrou[2], S. Guerrero[3], R. Ofman[2], M.A.M.M. Mannens[1], R.J.A. Wanders[2], J.N. Spelbrink[3], R.H.L. Houtkooper[2], M. Alders[1].* 1) Clinical Genetics, Academic Medical Center, Amsterdam, Netherlands; 2) Laboratory Genetic Metabolic Diseases, Academic Medical Center, Amsterdam, Netherlands; 3) Nijmegen Center for Mitochondrial Disorders, Radboud University Medical Center, Nijmegen, Netherlands.

Four unrelated females were diagnosed with Perrault syndrome [MIM 233400], presenting with gonadal dysgenesis and sensorineural deafness. This rare autosomal recessive disorder is genetically heterogeneous, with mutations previously described in five different genes, most of which related to mitochondrial proteostasis. Our patients did not have mutations in these genes, and we therefore set out to identify the genetic cause and underlying pathophysiology in these patients.Using whole exome sequencing, we identified a single homozygous mutation (c.707A>T; p.(Asn236Ile)) in the *ERAL1* gene [MIM 607435] in three out of four patients. The ERAL1 protein is involved in the assembly of the small mitochondrial ribosomal subunit, and therefore represented a likely candidate.We performed cell-based assays on patient skin fibroblasts to demonstrate impaired mitochondrial function. These cells displayed a reduction in ERAL1 protein levels, accompanied by reduced expression of proteins of the small ribosomal subunit. Assembly of the small ribosomal subunit appeared to be slightly disturbed for patients carrying the mutation, as evidenced by ribosome profiling in sucrose density gradients. At the physiological level, we showed that mitochondrial respiration using Seahorse XF96 was markedly decreased in patient fibroblasts.We used *C. elegans* as a model to investigate the importance of functional ERAL1 at an organismal level. Knockdown of the *ERAL1* worm homologue E02H1.2 in *rrf-3(pk1426)* worms almost completely blocked egg production, mimicking the compromised fertility in patients. Our cross-species data in patient cells and worms demonstrate that mutations in *ERAL1* cause Perrault syndrome and are associated with changes in mitochondrial metabolism.

278

The molecular pathology of a large cohort of individuals with inherited retinal disease, determined through whole genome sequencing. *K.J. Carss[1], G. Arno[2], M. Erwood[1], E. Dewhurst[1], J. Stephens[1], K. Stirrups[1], S. Ashford[1], C. Penkett[1], S. Hull[2], S. Lawrence[2], A.T. Moore[2], M. Michaelides[2], W.H. Ouwehand[1], A.R. Webster[2], F.L. Raymond[1], NIHR BioResource-Rare Diseases Consortium.* 1) Department of Haematology, University of Cambridge, Cambridge, United Kingdom; 2) Moorfields Eye Hospital and University College London Institute of Ophthalmology, London, United Kingdom.

Inherited retinal disease, collectively, is the most common cause of blind registration in the working age population in the UK. There are over 200 known associated genes, which can exhibit autosomal dominant (AD), autosomal recessive (AR), X-linked (XL), and mitochondrial inheritance. We are evaluating the utility of whole genome sequencing (WGS) to determine molecular pathology in a heterogeneous cohort of 450 unrelated individuals with retinal disease, as part of the UK National Institute for Health Research (NIHR) BioResource-Rare Diseases study. Specific phenotypes in our cohort include retinitis pigmentosa (RP), cone and cone-rod dystrophy, Usher syndrome, and macular dystrophy. Probands with either syndromic or non-syndromic inherited retinal disease were recruited from specialized clinics, when i) the molecular diagnosis was unknown and ii) the phenotype was not clearly associated with a single specific gene, unless prior sequencing of that gene had been negative.Of 208 likely pathogenic alleles identified in 222 individuals analysed so far, 184 are exonic single nucleotide variants (SNVs) or indels, 12 are intronic SNVs within 10 base pairs of an exon boundary, 11 are structural variants including deletions and tandem duplications, and 1 is a deep intronic SNV. Together these explain the molecular pathology in 129/222 (58%) individuals, in whom 105 are AR, 17 AD, and 7 XL. The most frequently implicated genes are *USH2A* (27/129, 21%) and *ABCA4* (14/129, 11%). Likely pathogenic variants were found in 54 distinct genes in our cohort.WGS has several advantages over exome or targeted sequencing, including the ability to identify non-coding variants, the inclusion of regions otherwise refractory to enrichment techniques, and more accurate characterisation of structural variants. However, some clinically relevant repetitive regions remain intractable, including *RPGR* ORF15 – the most common cause of X-linked RP. On-going analyses on the remaining patients include the identification of novel non-coding pathogenic variants, and pathogenic variants in novel disease associated genes.

279

Vibration-induced urticaria due to aberrant mast cell degranulation caused by a mutation in *ADGRE2*. *S.E. Boyden[1], A. Desai[2], G. Cruse[2], M.L. Young[3], H.C. Bolan[2], L.M. Scott[2], A.R. Eisch[2], R.D. Long[4], C.R. Lee[5], C.L. Satorius[1], A.J. Pakstis[6], A. Olivera[2], E. Chouery[7], A. Mégarbané[7,8], M. Medlej-Hashim[9], K.K. Kidd[6], D.L. Kastner[1], D.D. Metcalfe[2], H.D. Komarow[2].* 1) Inflammatory Disease Section, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; 2) Mast Cell Biology Section, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892; 3) Clinical Research Directorate/Clinical Monitoring Research Program, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702; 4) Veterinary Pathology Section, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT 59840; 5) Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892; 6) Department of Genetics, Yale University School of Medicine, New Haven, CT 06520; 7) Medical Genetics Unit, Faculty of Medicine, Saint Joseph University, Beirut, Lebanon; 8) Institut Jérôme Lejeune, Paris, France; 9) Department of Life and Earth Sciences, Faculty of Sciences II, Lebanese University, Fanar, Lebanon.

Vibratory urticaria (VU) is a rare condition in which sustained vibration against the skin induces both a localized hive and systemic manifestations such as facial flushing. We ascertained two large Lebanese kindreds in which VU segregates as an autosomal dominant trait. In affected family members, acute onset of symptoms, concurrent peripheral histamine release, and increased tryptase staining in post-vibration skin samples compared to controls implicated mast cell degranulation in the pathogenesis. Through linkage analysis and exome sequencing we identified the missense mutation p.C492Y in *ADGRE2* (formerly known as *EMR2*) as the only rare nonsynonymous or splice variant co-segregating with VU in these kindreds. Poorly covered exons within the linkage interval were Sanger sequenced to ensure no alternative variants had been missed, and the p.C492Y mutation was absent from variant databases and 200 ancestry-matched controls. *ADGRE2* encodes an adhesion G-protein coupled receptor that undergoes autocatalytic cleavage, producing an N-terminal extracellular alpha subunit that remains non-covalently bound to a C-terminal transmembrane beta subunit. ADGRE2 was highly expressed in human mast cells, and its alpha subunit binds to dermatan sulfate, which is abundant in skin. Patient-derived primary mast cells, when adhered with either dermatan sulfate or an antibody that ligates the ADGRE2 alpha subunit, degranulated when subjected to vibration, whereas mast cells from unaffected subjects showed no response. Likewise, human LAD2 mast cells expressing ADGRE2 with the p.C492Y mutation in the alpha subunit showed greater degranulation in response to vibration than control cells expressing non-mutant ADGRE2. This activity was cleavage-dependent, suggesting the subunit interaction must be non-covalent to permit vibration-induced degranulation. Furthermore, LAD2 cells expressing an ADGRE2 truncation mutant encoding only the beta subunit showed constitutive degranulation, indicating the alpha subunit is likely auto-inhibitory. Our data suggest a pathogenic mechanism whereby the p.C492Y mutation destabilizes this inhibitory subunit interaction, sensitizing dermal mast cells to vibration-induced hyperactivation of beta subunit-mediated signaling. We describe vibration as a novel IgE-independent mechanism for mast cell degranulation and provide the first genetic basis for a mechanically induced urticaria.

280

Mutations in the unfolded protein response regulator *ATF6* cause the cone dysfunction disorder achrom. *S. Kohl[1], D. Zobor[1], W-C. Chiang[2], N. Weisschuh[1], I. Gonzalez Menendez[1], S. Chang[3,4], S.C. Beck[1], M. Garcia Garrido[1], V. Sothilingam[1], M.W. Seeliger[1], F. Stanzial[5], E. Heon[6], A. Vincent[6], J. Beis[7], T.M. Strom[8,9], G. Rudolph[10], S. Roosing[11], A.I. den Hollander[11,12], F.P.M. Cremers[11], I. Lopez[13], H. Ren[13], A.T. Moore[14,15,16], A.R. Webster[14,15], M. Michaelides[14,15], R.K. Koenekoop[13], E. Zrenner[1,17], R.J. Kaufman[18], S.H. Tsang[3,19,20,21,22], B. Wissinger[1], J. Lin[2,23].* 1) Centre for Ophthalmology, Institute for Ophthalmic Research, University Tuebingen, Tuebingen, Germany; 2) Department of Pathology, University of California San Diego, La Jolla, California, USA; 3) Department of Ophthalmology, Columbia University, New York, New York, USA; 4) Edward Harkness Eye Institute, New York Presbyterian Hospital, New York, NY, USA; 5) Clinical Genetics Service, Regional Hospital Bozen, Italy; 6) Department of Ophthalmology and Vision Sciences, Programme of Genetics and Genomic Medicine, The Hospital for Sick Children, University of Toronto, Toronto, Canada; 7) Medical Genetics, IWK Health Centre, Halifax, Canada; 8) Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany; 9) Institute of Human Genetics, Technische Universität München, Munich, Germany; 10) University Eye Hospital, Ludwig-Maximilians-University, Munich, Germany; 11) Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands; 12) Department of Ophthalmology, Radboud University Medical Center, Nijmegen, the Netherlands; 13) McGill Ocular Genetics Centre, McGill University Health Centre, Montreal, Quebec, Canada; 14) University College London Institute of Ophthalmology, University College London, London, UK; 15) Moorfields Eye Hospital, London, UK; 16) Ophthalmology Department, University of California San Francisco Medical School, San Francisco, California, USA; 17) Werner Reichardt Center for Integrative Neuroscience, University of Tuebingen, Germany; 18) Degenerative Diseases Program, Sanford-Burnham Medical Research Institute, La Jolla, California, USA; 19) Jonas Laboratory of Stem Cell and Regenerative Medicine, Columbia University, New York, New York, USA; 20) Brown Glaucoma Laboratory, Columbia University, New York, New York, USA; 21) Institute of Human Nutrition, Columbia University, New York, New York, USA; 22) Department of Pathology and Cell Biology, Columbia University, New York, New York, USA; 23) Department of Ophthalmology, University of California, San Diego, La Jolla, California, USA.

Achromatopsia (ACHM; rod monochromatism, total colorblindness) is an autosomal recessive eye disorder characterized by low vision, lack of color discrimination, photophobia and nystagmus. Electroretinographic recordings show absence or severely reduced cone photoreceptor function in these patients. To date mutations in five genes – all of them encoding for essential components of the cone phototransduction cascade – have been shown to be associated with this disorder, and account for about 75% of cases in our patient cohort of more than 1,000 patients. By applying autozygosity mapping and whole exome sequencing in an Irish family with three affected siblings, we identified a homozygous missense mutation in *ATF6* that segregated with the disease in this family. Sanger sequencing of *ATF6* in a cohort of 301 unsolved ACHM patients resulted in the identification of nine further families with either homozygous or compound-heterozygous mutations, including a second missense variant, three frameshifting indel mutations and three splice site mutations, the latter verified by cDNA analysis. Patients with *ATF6* mutations presented with visual deficits typical for ACHM. Retinal imaging revealed foveal hypoplasia with an essentially absent foveal pit and a variable degree of disruption of the cone photoreceptor layer at the macula. There was no evidence for extraocular manifestation of the disease. In contrast to the known ACHM genes, ATF6 has no specific or exclusive function in phototransduction but encodes the ubiquitously expressed Activating Transcription Factor 6 that is known as a key regulator of the Unfolded Protein Response (UPR) and cellular endoplasmic reticulum (ER) homeostasis. Functional analysis showed that disease-associated *ATF6* variants lead to attenuated ATF6 transcriptional activity in response to ER stress in patient fibroblasts. *Atf6*$^{-/-}$ knockout mice present with normal retinal morphology and function in young, but both rod and cone dysfunction at older ages. Our study demonstrates that mutations in *ATF6* are a rare cause of ACHM (1% in our patient cohort) and suggests a crucial and unexpected role of ATF6 in human foveal development and/or cone photoreceptor function.

281

Mutations in the *MET* proto-oncogene cause osteofibrous dysplasia and alter the regulation of periosteal osteogenesis. *C.A. Wise[1,4], M.J. Gray[2], P. Kannu[3], S. Sharma[1], S.P. Robertson[2], The International Genetics of Osteofibrous Dysplasia Research Group.* 1) Seay Center for Musculoskeletal Research, Texas Scottish Rite Hosp, Dallas, TX; 2) Department of Women's and Children's Health, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand; 3) Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada; 4) Departments of Orthopaedic Surgery, Pediatrics, McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas, USA.

Osteofibrous dysplasia (OFD) (OMIM 607278) is a congenital bone dysplasia usually affecting the tibia, causing deformity and pathologic fractures in children. OFD typically occurs sporadically and is marked by radiolucent lesions of the diaphyseal cortex that may be pathogenetically related to differentiated adamantinoma, a primary malignant tumor of bone (OMIM 102660). To define the genetic basis of OFD we performed linkage mapping and exome sequencing in three families segregating an autosomal dominant form of the disease, and in a fourth sporadic case, identifying germline mutations in the proto-oncogenic *MET* gene encoding hepatocyte growth factor receptor tyrosine kinase. All OFD-associated mutations abolished the splice inclusion of exon 14 in *MET* transcripts, producing receptors (MET$^{\Delta 14}$) with an in-frame exclusion of a 47 amino acid cytoplasmic juxta-membrane domain (JMD). Exclusion of the MET JMD is known to attenuate receptor internalization and stabilize its ligand-dependent signalling. RNA *in situ* hybridization revealed prominent expression of both splice forms at E15 in the juxta-diaphyseal periosteum and also in dissected periosteum from a 3-week-old mouse. Furthermore, induction of endogenous *Met*$^{\Delta 15}$ transcripts in the mouse pre-osteoblastic cell line MC3T3 led to a block in progression to late stages of osteo-differentiation, suggesting that loss of the JMD subverts core functions of the mature receptor in the regulation of osteogenesis within the periosteum. To investigate sporadic OFD we Sanger-sequenced exon 14 in DNA from lesional tissue samples but did not detect mutations. Exome sequencing an additional lesional sample revealed a missense mutation in *MET* exon 14, c.3008A>C (p.Y1003S) that was clonally derived. This substitution affects a phosphorylation site that regulates JMD-mediated receptor degradation mediated by the ubiquitin ligase, CBL. Accordingly in transfected cell lines we found that the Y1003S substitution substantially reduced the sensitivity of the Met β-chain to CBL-mediated degradation. Together, these data suggest that mutations that effectively prolong ligand-dependent MET activity cause familial OFD and also explain some sporadic cases. These results support a central and heretofore unappreciated role for MET in periosteal osteogenesis, and suggest new avenues to therapeutically alter this process under circumstances when bone repair and maintenance need to be enhanced or sustained.

282

***MIPEP* Mutations Cause Autosomal Recessive Mitochondrial Dysfunction With Left Ventricular Non-Compaction, Hypotonia And Infantile Death.** *M.K. Eldomery[1], Z.C. Akdemir[1], J.A. Rosenfeld[1], R. Medikonda[10], L.C. Burrage[1,5], A.A. Shamsi[7], S. Penney[1], T. Gambin[1], S.N. Jhangiani[2], H.H. Zimmerman[9], D.M. Muzny[2], X. Wang[1,6], P. Ramachandran[10], L.J. Wong[1,6], E. Boerwinkle[2,3], R.A. Gibbs[1,2], S.E. Plon[1,4], A.L. Beaudet[1,6], C.M. Eng[1,6], J.R. Lupski[1,2,4,5], S.R. Lalani[1,5], J. Hertecant[7], R.J. Rodenburg[8], O.A. Abdul-Rahman[9], Y. Yang[1,6], F. Xia[1,6], M.C. Wang[1,10], V.R. Sutton[1,5].* 1) Molecular & Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; 3) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; 4) Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA; 5) Texas Children's Hospital, Houston, TX 77030, USA; 6) Baylor Miraca Genetics Laboratories, Baylor College of Medicine, Houston, TX 77030, USA; 7) Tawam Hospital, Al Ain, UAE; 8) Nijmegen Center for Mitochondrial Disorders, Department of Pediatrics, RadboudUMC, Nijmegen, Netherlands; 9) Department of Pediatrics, University of Mississippi Medical Center, 2500 N State St, Jackson, MS 39216, USA; 10) Huffington Center on Aging, Baylor College of Medicine, Houston, TX 77030, USA.

Mitochondrial peptidases perform fundamental roles in the processing of nuclear encoded proteins that are imported into the mitochondria. Although a recent report implicates one of the mitochondrial peptidases, the α subunit of the mitochondrial processing peptidase (MPP) encoded by *PMPCA* in non-progressive cerebellar ataxia, there remains a gap of knowledge with regards to the roles of other mitochondrial peptidases in human disease. Using whole exome sequencing (WES) we identified mutations in the *MIPEP* gene, which encodes the mitochondrial intermediate peptidase (MIP), in four unrelated individuals with early-infantile left ventricular non-compaction (LVNC), developmental delay (DD) and infantile death. Two individuals had compound heterozygous mutations (p.Leu582Arg/p.Leu71Gln and p.Glu602X/p.Leu306Phe) in the *trans* configuration and one individual from a consanguineous family had a homozygous mutation (p.Lys343Glu). The *MIPEP* gene discovery was identified through a coordinated effort between the clinical diagnostic laboratory (Baylor Miraca Medical Genetics Labs) and research efforts of the Baylor Hopkins Center for Mendelian Genomics. The fourth family segregating the same phenotype was identified through the GeneMatcher tool - a part of the Matchmaker Exchange Project. In this latter family the proband was found to have inherited a paternal SNV (p.His512Asp) and a maternal CNV (1.4 Mb deletion of 13q12.12 that includes the *MIPEP* gene). We show that RNA interference of the *MIPEP* orthologue, *Y67H2A.7*, in the model organism *C. elegans*, induces stress in the mitochondria, likely via the unfolded protein response (UPR) as evidenced by greatly increased production of heat shock protein 60 (HSP60). Our findings further define the role of mitochondrial peptidases in human disease and specifically implicate impaired MIP activity in LVNC. Moreover, our approach highlights the power of both data exchange and the importance of an interrelationship between clinical/research efforts for new gene discovery.

283

Identification of *RCBTB1* as novel disease gene for retinal ciliopathy. *F. Coppieters[1], G. Ascari[1], M. Karlstetter[2], M. Bauwens[1], N. De Rocker[1], A. Boel[1], K. Vleminckx[3], M. Van der Eecken[1], B.P. Leroy[1,4,5], F. Meire[6], T. Langmann[2], E. De Baere[1].* 1) Center for Medical Genetics Ghent, Ghent University, Ghent, Belgium; 2) Dept of Ophthalmology, Cologne University, Cologne, Germany; 3) Department of Biomedical Molecular Biology, Ghent University, Ghent (Zwijnaarde), Belgium; 4) Dept of Ophthalmology, Ghent University Hospital, Ghent, Belgium; 5) Division of Ophthalmology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States; 6) Dept of Ophthalmology, Queen Fabiola Children's Hospital (Huderf), Brussels, Belgium.

Purpose: To identify and functionally study a novel disease gene mutated in a Turkish consanguineous family with a severe retinal ciliopathy, characterized by retinitis pigmentosa (RP), hypothyroidism, hypogonadism (amenorrhea), short stature, intellectual disability and facial dysmorphism.**Methods:** Genome-wide SNP arrays were used for homozygosity mapping in three affected and one healthy sibling of a consanguineous family. Two affected individuals underwent whole exome sequencing (WES) (HiSeq2000, Illumina; CLC bio). Segregation analysis of variants was done using Sanger sequencing. *RCBTB1* expression was assessed in human cDNAs and in zebrafish tissues of different developmental stages. Localization studies were carried out in mouse retina. *Rcbtb1 in situ* hybridization was performed in zebrafish. Knockdown and RNA rescue experiments in zebrafish are ongoing. **Results:** Homozygosity mapping revealed a single 11 Mb homozygous region on chromosome 13 shared by the three affected individuals. WES identified a novel missense variant, c.973C>T p.(His325Tyr) (rs200826424), in *RCBTB1* (NM_018191.3). This variant was found to be homozygous in all affected individuals and heterozygous in the healthy sibling and parents of the affected persons. Strong conservation and different prediction tools point toward an effect of the variant on protein function. *RCBTB1* expression was demonstrated in human retina, consistent with retinal expression in databases, also showing thyroid expression, and with *rcbtb1 in situ* hybridization in zebrafish embryos. In addition, *rcbtb1* expression was demonstrated in ovaries, eye and brain tissues from adult zebrafish. Immunostaining in mouse retina showed a ciliary staining in inner segments. **Conclusions:** By combining homozygosity mapping and WES, a missense variant was identified in the *RCBTB1* gene, encoding a regulator of chromosome condensation (RCC1) and BTB (POZ) domain containing protein , in a family with a retinal ciliopathy. Interestingly, RCC1-like domains are also present in NEK8 and RPGR, which are known ciliary proteins implicated in nephronophthisis and X-linked RP, respectively. Further functional characterization of RCBTB1 will provide more insights in its role in the pathogenesis of retinal ciliopathies.

284

THE INVESTICATE project: Identification of New Variation, Establishment of Stem cells, and TIssue Collection Advancing Treatment Efforts. *C. Ernst[1], W. Al-Hertani[2], N. Mechawar[1].* 1) Psychiatry, McGill University, Montreal, Quebec, Canada; 2) Medical Genetics, University of Calgary, Calgary, Alberta, Canada.

Neurodevelopmental disorders (NDDs) are a large and complex group of disorders with varied etiologies. Combining sequencing, induced stem cell, and small molecule screening technologies allow for the development of personalized treatment for NDDs. Patients are enrolled if they have a similarly affected sibling and negative genetic tests, or a *de novo* balanced chromosomal rearrangement (BCR). We use Next-Generation Sequencing tools to find variation and structural variant breakpoints. Fibroblasts from patients undergo rapid induced pluripotent stem cell (iPSC) to neural progenitor cell (NPC) differentiation, CRISPR-guided mutation correction of small variants, and cell phenotyping. Where feasible, we collect brains from cases with reduced life expectancy. Patient-derived NPCs undergo high-throughput small molecule screening to reverse cell phenotypes associated with disease. INVESTICATE has to date recruited six families, and we have identified never before observed mutations in genes not previously associated with NDDs. We identified a stop codon altering base deletion in one family, a 51-basepair promoter deletion in another family, and a gene truncating translocation implicating chromatin remodelling, netrins, and maintenance of brain pH in NDDs. Functional assays using iPSC-NPCs of these rare variants support their role in disease. INVESTICATE is a rapid bedside-to-bench and back again pipeline capable of finding variants missed using standard methodology, and complements variant detection with a full battery of cell phenotyping assays, brain collection, and high-throughput screening.

285

The Koolen-de Vries syndrome: A phenotypic comparison of microdeletion and point mutation patients. *D.A. Koolen[1], R. Pfundt[1], B.P. Coe[2], J. Gecz[3], C. Romano[4], E.E. Eichler[2,5], B.B.A de Vries[1], DDD Study, KdVS research group.* 1) Department of Human Genetics, Radboud University Medical Center, Nijmegen, Gelderland, Netherlands; 2) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; 3) School of Paediatrics and Reproductive Health and Robinson Research Institute, The University of Adelaide at the Women's and Children's Hospital, North Adelaide, Adelaide, Australia; 4) Pediatrics and Medical Genetics, I.R.C.C.S. Associazione Oasi Maria Santissima, Troina, Italy; 5) Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

The Koolen-de Vries syndrome (KdVS; OMIM #610443), also known as the 17q21.31 microdeletion syndrome, is a clinically heterogeneous disorder characterized by (neonatal) hypotonia, developmental delay, moderate intellectual disability, and characteristic facial dysmorphism. Other frequently reported features include epilepsy, musculoskeletal anomalies, congenital heart defects, urogenital malformations, and ectodermal anomalies. Expressive language development is particularly affected compared to receptive language or motor skills. Moreover, many individuals with KdVS display a social and friendly behavior.The syndrome is either caused by a truncating mutation in the KAT8 regulatory NSL complex unit 1 (*KANSL1*) gene or by a 17q21.31 microdeletion encompassing *KANSL1*.We collected clinical information on a unique cohort of 45 individuals with KdVS, of whom 33 have a 17q21.31 microdeletion, and 12 have a mutation in *KANSL1* (19 males, 26 females, age range 7 months to 50 years). We show detailed phenotypic information, including neuropsychological features, that contribute to the broad phenotypic spectrum of the syndrome. Importantly, comparison of the phenotypes of both the microdeletion and single-nucleotide variants patients does not show differences of clinical importance, stressing that haploinsufficiency of *KANSL1* is sufficient to cause the full KdVS phenotype. In addition, we provide practical information on the molecular and clinical interpretation of *KANSL1* mutations and genomic copy number variation in the complex 17q21.31 region.

286

Exome Sequencing Suggests Aicardi Syndrome is Genetically Heterogeneous and not Exclusive to Females. *I. Schrauwen[1,2,3], S. Szelinger[1,2], A.L. Siniard[1,2], J.J. Corneveaux[1,2], A.M. Claasen[1,2], R.F. Richholt[1,2], M. De Both[1,2], B. Hjelm[1,2], S. Rangasamy[1,2], N. Kulkarni[4], S. Bernes[4], J. Buchhalter[5], M. Russell[1,2], A.L. Courtright[1,2], K. Ramsey[1,2], D.W. Craig[1,2], V. Narayanan[1,2], M. Huentelman[1,2].* 1) Center for Rare Childhood Disorders, Translational Genomics Research Institute, Phoenix, AZ, USA; 2) Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ, USA; 3) Department of Medical Genetics, University of Antwerp, Antwerp, Belgium; 4) Phoenix Children's Hospital, Phoenix, AZ, USA; 5) Alberta Children's Hospital, University of Calgary, Alberta, Canada.

Aicardi Syndrome (AIC), a rare female neurodevelopmental disorder affecting the brain and retina, has captured the attention of geneticists for some time since it was strongly presumed to be X-linked, however, no gene on the X-chromosome has ever been conclusively associated with the disease. By performing exome and/or genome sequencing in 10 trios with AIC, we identified a *de novo* mutation in the Hippo pathway gene *TEAD1* in a patient with chorioretinal lacunae, infantile spasms, a posterior fossa cyst and periventricular heterotopias. Mutations in *TEAD1* have previously been linked to Sveinsson's chorioretinal atrophy (SCRA). The Hippo pathway is a highly conserved signaling pathway that regulates cell number by modulating cell proliferation, cell death, and cell differentiation. Selective dysregulation in tissues expressing high levels of *TEAD1* during development (i.e. brain and eye) explain the specific targeted lesions in both AIC and SCRA. In addition, by performing RNA-sequencing on RNA extracted from blood, we found that altered expression of genes associated with synaptic plasticity, neuronal development, retinal development, and cell cycle control/apoptosis is an important underlying potential pathogenic mechanism shared among cases compared to parents and age-matched controls. Further research aimed to identify causal mechanisms will include methylation profiling in cases and parents. The finding of a mutation on an autosomal gene in a case with AIC is of clinical importance, because it suggests AIC can develop in boys. Current diagnostic criteria for AIC include female gender, and this finding, once replicated, could change clinical practice and identify significant numbers of boys with this disease and most of these families may either have no clinical diagnosis as yet or may be diagnosed with a different disorder. The publication of two recent reports of males with AIC (46, XY) also supports this notion. Lastly, we are beginning to recognize that AIC is a spectrum disease – this is demonstrated by the phenotypic differences in our cohort and further solidified by our genetic findings. In conclusion, in this study, we expand the phenotype of *TEAD1* mutations, demonstrate its importance in chorioretinal complications, and propose the first putative pathogenic mechanisms underlying AIC. Our data suggest that AIC is a genetically heterogeneous disease and is not restricted to the X-chromosome, and that *TEAD1* mutations may be present in males.

287

Targeted sequencing of 15 genes in a cohort of 169 patients with unexplained lissencephaly detects mutations in 37% of patients. *N. Di Donato[1,2], AE. Timms[3], S. Collins[1], C. Adams[1], GM. Mirzaa[1,4], WB. Dobyns[1,4].* 1) Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA; 2) Institute for Clinical Genetics, Technical University Dresden, Dresden, Germany; 3) Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, Seattle, WA; 4) Department of Pediatrics and Department of Neurology, University of Washington, Seattle, WA.

We have collected DNA samples on more than 700 children with lissencephaly (LIS) over ~30 years. Many have been tested for deletion 17p13.3 and mutations of LIS1 and DCX, but few for other genes. We therefore designed a targeted sequencing panel of 15 genes including ACTB, ACTG1, DCX, LIS1, TUBA1A, TUBA8, TUBB2B, TUBB, TUBB3, TUBG1, KIF2A, KIF5C, DYNC1H1, RELN and VLDLR, using single molecule molecular inversion probes (smMIPs). In our first run, we found mutations in 63 of 169 (37.3%) patients ascertained between 1998 and 2015, with several additional variants still under review. LIS1 remains the most common causative gene (N=14, most not previously tested), followed by DYNC1H1 (N=13), the largest known LIS gene. All but one of the mutations in DYNC1H1 are novel, and include the first splice site mutation. Two patients presented with bilateral congenital cataracts, expanding the spectrum of DYNC1H1 associated malformations. We also report a novel recurrent mutation of TUBG1 (p.S259L) in three patients with posterior LIS and normal head size; most reported patients have had severe microcephaly. We also found a novel recurrent mutation in TUBB3 (p.M388T) in two patients with severe diffuse LIS and cerebellar hypoplasia; the phenotype resembles a single fetus reported with p.M388V. We detected 5 mutations in ACTG1, all with mild features of Baraitser-Winter (cerebrofrontofacial) syndrome noted in retrospect. We found two mutations in KIF5C and none in KIF2A, which suggests that these are rare causes of LIS. Finally, we identified two children with homozygous truncating mutations of RELN. Both had severe diffuse LIS rather than the mild frontal predominant LIS seen in the few currently reported splicing and missense variants, a striking expansion of the RELN-associated phenotype.Despite deep sequencing of most known LIS genes, we found no mutations in 106 of 169 (62.7%) patients. This strongly suggests that several and possibly many additional LIS genes remain to be discovered. The majority of the unsolved patients presented with posterior predominant LIS grade 4 with or without additional features suggestive for tubulinopathies, subcortical band heterotopia or mild anterior predominant PGY. We report several novel mutations and define key clinical features of DYNC1H1-, TUBG1-, TUBB3-, and RELN-associated phenotypes.

288

Integration of functional "omics" data uncovers mitochondrial deficiency in Smith-Magenis syndrome. *J.T. Alaimo, S.V. Mullegama, L. Chen, R. Masand, T. Donti, A. Besse, P.E. Bonnen, B.H. Graham, S.H. Elsea.* Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

Discerning the multifaceted cellular defects that contribute to the pathogenesis of neurodevelopmental disorders (NDDs) remains an important challenge. We devised an innovative combinatorial functional "omics" approach to dissect the pathological and physiological cellular states of Smith-Magenis syndrome (SMS), a NDD caused by reduced gene dosage of *RAI1*. We employed transcriptional profiling on neuronal cells with targeted knockdown of *RAI1* in conjunction with a small molecule metabolomics screen targeting >1000 molecules in plasma samples from individuals with SMS. Enrichment analysis of differentially expressed transcripts within *RAI1* haploinsufficient neuronal cells identified a significant enrichment of 60 mitochondria-associated genes. The metabolomics screen revealed significantly increased mitochondria-associated metabolites including pyruvate and lactate which are known to be altered amongst individuals with a NDD. The convergence of both transcriptomic and functional metabolomic profiles prompted the biochemical evaluation of mitochondrial function in SMS patient fibroblast cell lines. Our battery of mitochondrial tests identified diminished mitochondrial membrane potential and increased cellular respiration and oxygen consumption rates which are indicative of compromised mitochondrial function. Further analysis uncovered elevated mitochondrial DNA content, elevated citrate synthase protein levels and activity and elevated *PGC1A* transcripts which suggests mitochondrial biogenesis and proliferation are occurring. This may represent a compensatory response to compromised function. Lastly, fluorescence microscopy revealed an unusual perinuclear distribution of mitochondria in SMS cell lines. Taken together, our results demonstrate that mitochondrial function and integrity are compromised in SMS patients. The SMS mitochondrial deficiencies observed in this study resemble those observed in other NDDs including ASD, Down syndrome, fragile X syndrome, and Rett syndrome. This finding suggests a model in which mitochondrial dysfunction elicits a similar pattern of clinical manifestations of intellectual disability, developmental delay, hypotonia, seizures and gastrointestinal symptoms across several genetically distinct NDDs. Overall, our integrative functional "omics" approach has pinpointed cellular defects in SMS patients similar to those observed in other NDDs and highlights potential new avenues of exploration for therapeutic interventions for SMS.

289

Mutations in DDX3X are a common cause of unexplained intellectual disability with gender-specific effects on Wnt signaling. *L. Snijders Blok[1], E. Madsen[2], J. Juusola[3], C. Gilissen[1], D. Baralle[4], M.R.F. Reijnders[1], H. Venselaar[5], C. Helsmoortel[6], M.T. Cho[3], A. Hoischen[1], L. Vissers[1], T.S. Koemans[1], W. Wissink-Lindhout[1], E.E. Eichler[7,8], C. Romano[9], H. Van Esch[10], C. Stumpel[11], M. Vreeburg[11], E. Smeets[11], K. Oberndorff[12], B.W.M. van Bon[1,13], M. Shaw[13], J. Gecz[13], E. Haan[13,14], M. Bienek[15], C. Jensen[15], B.L. Loeys[6], A. Van Dijck[6], A.M. Innes[16], H. Racher[16], S. Vermeer[17], N. Di Donato[18], A. Rump[18], K. Tatton-Brown[19], M. J. Parker[20], A. Henderson[21], S. A. Lynch[22], A. Fryer[23], A. Ross[24], P. Vasudevan[25], U. Kini[26], R. Newbury-Ecob[27], K. Chandler[28], the DDD study[29], S. Dijkstra[30], J. Schieving[31], J. Giltay[32], K. L.I. van Gassen[32], J. Schuurs-Hoeijmakers[1], P. L. Tan[2], I. Pediaditakis[2], S. A. Haas[33], K. Retterer[3], P. Reed[3], K. G. Monaghan[3], E. Haverfield[3], M. Natowicz[34], A. Myers[35], M. C. Kruer[35], Q. Stein[35], K. A. Strauss[36], K. W. Brigatti[36], K. Keating[37], B. K. Burton[37], K. H. Kim[37], J. Charrow[37], J. Norman[38], A. Foster-Barber[39], A. D. Kline[40], A. Kimball[40], E. Zackai[41], M. Harr[41], J. Fox[42], J. McLaughlin[42], K. Lindstrom[43], K. M. Haude[44], K. van Roozendaal[11], H. Brunner[1,11], W. K. Chung[45], R. F. Kooy[6], R. Pfundt[1], V. Kalscheuer[15], S. G. Mehta[46], N. Katsanis[2], T. Kleefstra[1].* 1) Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands; 2) Center for Human Disease Modeling, Department of Cell Biology, Duke University Medical Center, Durham, NC 27710, United States; 3) GeneDx, Gaithersburg, Maryland, 20877, United States; 4) Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, United Kingdom; 5) Nijmegen Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Medical Center, 6500 HB Nijmegen, The Netherlands; 6) Department of Medical Genetics, University of Antwerp and University Hospital Antwerp, 2650 Antwerp, Belgium; 7) Department Genome Sciences, University of Washington, Seattle, WA, United States; 8) Howard Hughes Medical Institute, Seattle, WA, United States; 9) Pediatrics and Medical Genetics, IRCCS Associazione Oasi Maria Santissima, 94018 Troina, Italy; 10) Center for Human Genetics, University Hospitals Leuven, 3000 Leuven, Belgium; 11) Department of Clinical Genetics and School for Oncology & Developmental Biology (GROW), Maastricht UMC+, 6202 AZ Maastricht, Netherlands; 12) Department of Pediatrics, Atrium-Orbis Medical Center, 6162 BG, Sittard, The Netherlands; 13) School of Paediatrics and Reproductive Health and Robinson Research Institute, The University of Adelaide, Adelaide, South Australia 5006, Australia; 14) South Australian Clinical Genetics Service, SA Pathology, Adelaide, South Australia 5006, Australia; 15) Department of Human Genetics, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; 16) Department of Medical Genetics and Alberta Children's Hospital Research Institute for Child and Maternal Health, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada; 17) Department of Genetics, University Medical Center Groningen, 9713 GZ Groningen, the Netherlands; 18) Faculty of Medicine Carl Gustav Carus TU Dresden, 01307 Dresden, Germany; 19) St George's University of London, London, United Kingdom, SW170RE; 20) Sheffield Clinical Genetics Service, Sheffield Children's Hospital, Western Bank, Sheffield S10 2TH, United Kingdom; 21) Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom; 22) National Centre for Medical Genetics, Temple street Children's Hospital, Dublin, Ireland; 23) Department of Clinical Genetics, Liverpool Women's Hospital and Alder Hey Children's Hospital, Liverpool, L8 7SS, United Kingdom; 24) North of Scotland Regional Genetics Service, Clinical Genetics Centre, Aberdeen, United Kingdom; 25) Department of Clinical Genetics, University Hospitals of Leicester, Leicester Royal Infirmary, Leicester, United Kingdom; 26) Department of Clinical Genetics, Oxford University Hospitals NHS Trust, Oxford, United Kingdom; 27) Department of Clinical Genetics, University Hospitals, Bristol, United Kingdom; 28) Manchester Centre for Genomic Medicine, St. Mary's Hospital, Manchester Academic Health Sciences Centre (MAHSC), Manchester, UK; 29) Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 30) ORO, Organisation for people with Intellectual Disabilities, 5751 PH Deurne, Netherlands; 31) Department of Child Neurology, Radboud University Medical Center, 6500 HB Nijmegen, Netherlands; 32) Department of Medical Genetics, University Medical Center Utrecht, 3508 AB Utrecht, Netherlands; 33) Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; 34) Pathology & Laboratory Medicine and Genomic Medicine Institutes, Cleveland Clinic, Ohio, USA; 35) Barrow Neurological Institute and Ronald A. Matricaria Institute of Molecular Medicine, Phoenix Children's Hospital, Phoenix, AZ 57117, USA; 36) Clinic for Special Children, Franklin & Marshall College, Pennsylvania, USA; 37) Division of Genetics, Birth Defects & Metabolism, Ann & Robert H. Lurie Children's Hospital Of Chicago, Illinois, 60611, USA; 38) Integris Pediatric Neurology, Oklahoma City, Oklahoma, 73112, USA; 39) Child Neurology and Palliative Care, Benioff Children's Hospital San Francisco, CA 94925, USA; 40) The Harvey Institute for Human Genetics, Greater Baltimore Medical Center, Maryland, USA; 41) Department of Pediatrics, Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA; 42) Division of Medical Genetics, North Shore-LIJ, Manhasset, New York, 11040, USA; 43) Phoenix Children's Hospital, Division of Genetics and Metabolism, AZ, 85006, USA; 44) University of Rochester Medical Center, New York, 14642, USA; 45) Departments of Pediatrics & Medicine, Columbia University Medical Center, 10032 New York, USA; 46) East Anglian Regional Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Addenbrooke's Hospital, Cambridge, CB2 0QQ, United Kingdom.

Intellectual disability (ID) affects approximately 1% of humans with a gender bias towards males. Previous studies have identified mutations in over 100 genes on the X chromosome in males with ID, but there is less evidence for de novo mutations on the X chromosome causing ID in females. By whole exome sequencing we identified 34 unique deleterious de novo mutations in DDX3X in 37 females with ID and various other features including hypotonia, movement disorders, behavior problems, corpus callosum hypoplasia and epilepsy. *DDX3X* is among the most intolerant genes, normal variation in this gene is extremely rare. Based on our findings, mutations in DDX3X are one of the more common causes of ID accounting for 1-2% of unexplained ID in females.Although no de novo DDX3X mutations were identified in males, we present three families with segregating missense mutations in DDX3X, suggestive of an X-linked recessive inheritance pattern. In these families males with the DDX3X variant all had ID, while carrier females were unaffected. To explore the pathogenic mechanisms accounting for the differences in disease transmission and phenotype between affected females and affected males with DDX3X missense variants, we used canonical Wnt defects in zebrafish as a surrogate measure of DDX3X function in vivo. We demonstrate a consistent loss of function effect of all tested de novo mutations on the Wnt-pathway, and we further show a differential effect by gender. The differential activity possibly indicates a dose dependent effect of DDX3X expression in the context of functional mosaic females versus one-copy males, which reflects the complex biological nature of DDX3X mutations.

290

Mutations in *TKT* gene are a novel cause of short stature, developmental delay, and congenital heart defects. *A.H. Begtrup[1], L. Boyle[2], M.M.C. Wamelink[3], G.S. Salomons[3], B. Roos[3], M.T. Cho[1], A. Dauber[4], J. Douglas[5], M. Feingold[5], S. Saitta[6], N. Kramer[6], J. Wynn[7], W.K. Chung[8].* 1) GeneDx, Gaithersburg, MD; 2) Physicians and Surgeons, Columbia University, New York, NY; 3) Metabolic Unit, Department of Clinical Chemistry, VU University Medical Center, Amsterdam, The Netherlands; 4) Division of Endocrinology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 5) Boston Children's Hospital, Boston, MA; 6) Cedars-Sinai Medical Center, Los Angeles, CA; 7) Department of Pediatrics, Columbia University, New York, NY; 8) Department of Medicine, Columbia University, New York, NY.

The use of whole-exome sequencing (WES) is increasing in clinical practice to diagnose patients with undiagnosed disorders, particularly those that are familial and likely to have an inherited basis. Developmental delay and short stature are common clinical indications for WES. We performed WES on three proband-parent trios and two additional affected siblings. Variants were evaluated using a custom analysis platform encompassing alignment, variant calling and annotation, and interactive filtering. We describe the identification of a novel syndrome due to an autosomal recessively inherited deficiency of transketolase encoded by *TKT* on chromosome 3p21. Our series includes three families with a total of five affected individuals, four females and one male ranging in age from 4 to 25 years, all of European ancestry. Two families of Ashkenazi Jewish ancestry were homozygous for an 18 base pair inframe insertion in *TKT*, indicating a possible founder mutation. The third family was compound heterozygous for nonsense and missense mutations in *TKT*. All patients were small for gestational age, had short stature, and were developmentally delayed. Congenital heart defects were noted in 4 of the 5 affected individuals, and a history of chronic diarrhea and cataracts were noted in the older individuals with the homozygous insertion mutation. Enzymatic testing confirmed significantly reduced transketolase activity in four cases. Elevated urinary excretion of erythritol, arabitol, ribitol, and pent(ul)ose-5-phosphates was detected as well as elevated erythritol, arabitol and ribitol levels in plasma of these patients.Transketolase (TK) is a reversible, thiamine-dependent enzyme in the pentose phosphate pathway (PPP) necessary for NADPH synthesis, nucleic acid synthesis, and cell division. Mice completely deficient for the enzyme are not viable, suggesting that TK is an essential enzyme. Transketolase deficiency is one of a growing list of inborn errors of metabolism in the non-oxidative portion of the pentose phosphate pathway. With the increase in utilization of WES and other genetic and metabolic testing, we anticipate that more patients may be identified with mutations in this pathway. Confirmation by metabolite and enzymatic testing is essential to make the diagnosis.

291

Deciphering phenotypic variability of genomic disorders using the 16p11.2 syndromes as a paradigm. *K. Männik[1, 2], A.M. Maillard[3], K. Popadin[1], L. Hippolyte[3], A. Pain[3], S. Martin-Brevet[3], A. Alfaiz[1, 4], E. Migliavacca[1, 4], J. Kosmicki[5, 6], S. Lebon[7], B. Kolk[2, 8], M. Noukas[2, 8], A. Metspalu[2, 8], M.M. van Haelst[9], M.J. Daly[5, 6], N. Katsanis[10], J.S. Beckmann[3, 4], S. Jacquemont[3], A. Reymond[1], 16p11.2 European Consortium, Simons VIP Consortium.* 1) Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; 2) Estonian Genome Center, University of Tartu, Tartu, Estonia; 3) Department of Medical Genetics, Lausanne University Hospital, Lausanne, Switzerland; 4) Swiss Institute of Bioinformatics, Lausanne, Switzerland; 5) Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA; 6) Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 7) Pediatric Neurology Unit, Department of Pediatrics, Lausanne University Hospital, Lausanne, Switzerland; 8) Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; 9) Department of Medical Genetics, University Medical Centre Utrecht, Utrecht, the Netherlands; 10) Center for Human Disease Modeling and Department of Cell biology, Duke University, Durham, NC, USA.

The reciprocal CNVs in 16p11.2 are one of the most frequent genetic lesions in neurodevelopmental disorders. They impact cognition, behavior, head size and BMI in a dosage-dependent manner, and akin to several genomic disorders, the 16p11.2 deletion (OMIM #611913) and duplication (#614671) syndromes are characterized by considerable variance in expressivity. This suggests that yet unexplained modifying factors may contribute to the patients' phenotypic outcome. Our recent findings indicated a possible contribution of ciliary dysfunction to the clinical phenotypes of the 16p11.2 CNVs. Intriguingly, ciliopathies (e.g. Bardet-Biedl syndrome) share phenotypic overlap with 16p11.2 syndromes and are known for extensive variance in the phenotype.To potentially identify both driver and modifier genes of the 16p11.2 patients' phenotype we have sequenced the exomes and the transcriptomes of 200 deeply phenotyped individuals from 16p11.2 families. To avoid ascertainment bias and better represent the phenotypic spectrum, we recruited carriers identified among unselected population cohorts, as well as in clinical cytogenetic setting and will complement this set by 306 exomes of 16p11.2 trios from the Simons VIP Consortium. We are cataloging potentially deleterious variants at ciliopathy loci, in 16p11.2-altered pathways, and performing correlative analyses between quantitative traits and transcript levels.Although systematic analysis of this data is in progress, during the pilot phase we discovered an individual carrying a *de novo* 16p11.2 deletion and a heterozygous null *CEP290* (#610142) allele, and an individual carrying a paternally inherited 16p11.2 duplication and a *de novo* non-synonymous mutation in *PTPN11* (#176876). Compatible with our "two-hit" hypothesis these patients present alleviated and aggravated phenotype spectra, respectively. Corroboratingly, correlation analyses of transcript levels within the 16p11.2 interval suggest that *KCTD13* (#60894), *MVP* (#605088) and *MAPK3* (#601795), three genes with epistatic effect on zebrafish head neuroanatomy, show co-regulated expression and significant association with BMI in adult human 16p11.2 carriers.Our study has potential implications for precise clinical management of 16p11.2 CNV carriers, and sheds light to the mechanisms contributing to the complex etiologies of genomic disorders.

292

Modeling Microcephaly using DNA Repair Defective Induced Pluripotent Stem Cells and Cerebral Organoids. *F. Pirozzi[1], K. Plona[1], B. Ward[1], J. Ngo[1], T.H. Kim[1], E. Gilmore[2], A. Wynshaw-Boris[1].* 1) Department of Genetics and Genome Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH 44106; 2) Department of Pediatrics Division of Neurology, University Hospitals Case Medical Center, Cleveland, OH 44106.

Microcephaly is found in isolated or syndromic forms of neurodevelopmental diseases and may be associated with neurological defects, brain structural abnormalities, intellectual disabilities and seizures. Mutations in DNA repair genes lead to microcephaly, demonstrating that the maintenance of genomic stability is crucial for proper brain development and size. Microcephaly caused by mutations in DNA repair genes could be due to abnormal proliferation and/or increased apoptosis during neurogenesis, but the pathogenesis is poorly understood. We have generated patient-derived induced Pluripotent Stem Cell (iPSC) with mutations in the DNA repair pathway genes *LIG4*, *PNKP* or *NBN*. As controls, we derived iPSCs from unaffected individuals; isogenic lines in which the mutations are corrected using CRISPR/Cas9 genome editing; and iPSCs from patients with *ATM* mutations who lack microcephaly but have defects in the DNA damage response. We used these iPSCs to generate neuronal precursor cells (NPCs), cortical neurons, and 3D cerebral organoids, which will allow us to study proliferation, apoptosis and differentiation as well as early self-arranged neuronal structures in the organoids. Preliminary results examining the differentiation of iPSCs with a *LIG4* mutation into cortical neurons by transduction of neurogenin-2 (*NGN2*) demonstrated clear neuronal morphology after 6 days and more mature neurons within 2 weeks. Interestingly, these LIG4 iPSCs displayed higher transduction efficiency and differentiated faster into neurons compared to control NGN2+ cells. However, 2 weeks after transduction, the LIG4 neurons displayed increased cell death. In addition, LIG4 cerebral organoids were 2 times smaller than the control organoids during the first 4 weeks of formation. Furthermore, one of the LIG4 clones failed to develop after the first 35 days of treatment in approximately 30% of cases. Immunostaining of LIG4 organoid sections showed an increase in cleaved caspase-3 compared to the controls, supporting a role for apoptosis in the microcephaly phenotype in these patients.We are producing NPCs that will be examined for proliferation and differentiation phenotypes, and we are expanding all of the above mentioned experiments to the other DNA repair-deficient iPSC lines. We believe that this will allow us to further dissect the importance of DNA damage repair underlying the pathogenesis of DNA repair-related microcephaly.

293

3q29 deletion syndrome is associated with feeding problems, reduced birth weight, and a range of neuropsychiatric phenotypes: Results from the 3q29 registry. *J.G. Mulle[1,2], M. Glassford[2], J.A. Rosenfeld[3], A. Freedman[1], E. McGarry[4], M.E. Zwick[2,5], Unique Rare Chromosome Disorder Support Group.* 1) Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA; 2) Department of Human Genetics, Emory University School of Medicine, Atlanta GA; 3) Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX; 4) Marcus Autism Center, Children's Healthcare of Atlanta and Emory University School of Medicine, Atlanta, GA; 5) Department of Pediatrics, Emory University School of Medicine, Atlanta GA.

Recurrent ~1.6 Mb interstitial deletions on chromosome 3q29 were first described in 2005, in six patients with mild to moderate intellectual disability. The 3q29 deletion has since been associated with a range of neurodevelopmental and neuropsychiatric phenotypes, including autism, schizophrenia, and bipolar disorder in addition to intellectual disability and developmental delay. Notably, risk for schizophrenia is particularly high; recent data suggests a greater than 40-fold increase in risk for deletion carriers. However, many aspects of the phenotype are not well described. To better understand the range of manifestations associated with 3q29 deletion syndrome, we have created an internet-based registry and research study (3q29deletion.org) to determine the medical, behavioral, and biological consequences of the deletion, consistent with the "genetics first" approach utilized in other CNV studies. Results from the first year of data collection for the registry (n = 36 participants) indicate a high prevalence of neuropsychiatric phenotypes, including anxiety disorder, panic attacks, depression, bipolar disorder, and schizophrenia. Overall, 28% of registry participants have a psychiatric condition, which is striking given that the average age of participants in the registry is only 11.5 years old (range < 1 yr to 75 yrs). Other findings include a high prevalence of feeding disorders in the first year of life, and an observation of reduced weight at birth for 3q29 deletion carriers (average reduction 13.6 oz, adjusted for gestational age and sex, p = 6.74e-06). The latter suggests a possible metabolic imbalance and reduced capacity for energy harvest, a phenotype that may exist prenatally. These results are clinically actionable toward improving patient care for 3q29 deletion carriers, highlight the utility of internet-based registries for the study of rare variants, and emphasize the importance of the 3q29 deletion as a potential molecular handle on neurodevelopmental pathways relevant to neuropsychiatric phenotypes.

294

Schizophrenia risk gene MIR137 modulates neurodevelopment and behavior. *Y. Cheng[1,2], W. Tan[2], Z. Teng[2], B. Bai[3], L. Lin[1], Y. Kang[1], Y. Li[1], B. Yao[1], X. Li[1], N. Xie[1], J. Peng[3], D. Chen[2], P. Jin[1].* 1) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA; 2) State Key Laboratory of Reproductive Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P.R. China; 3) Departments of Structural Biology and Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

Schizophrenia is an early adult-onset neurodevelopmental disorder characterized by a constellation of symptoms including hallucinations and delusions. Recent genome-wide association studies have shown the SNPs in the vicinity of MIR137 gene (encoding the microRNA miR-137) and other four miR-137 target genes are associated with schizophrenia, which highlights MIR137-mediated dysregulation that may act as an unknown etiologic mechanism in schizophrenia. To systematically investigate the role of miR-137 in neurodevelopment, we have generated the *miR-137* knockout mice. Mature miR-137 level in brain tissues significantly reduced in both heterozygotes (*miR-137*$^{+/-}$) and homozygotes (*miR-137*$^{-/-}$). The complete loss of miR-137 (*miR-137*$^{-/-}$) leads to postnatal lethality (up to postnatal 21 days, P21), while the *miR-137*$^{+/-}$ mice exhibit normal morphology and fertility. Using a battery of behavioral tests, we have found that *miR-137*$^{+/-}$ mice exhibited impaired social memory, anxiety-like behavior and decreased locomotor activity. Analyses of *miR-137*$^{+/+}$ and *miR-137*$^{-/-}$ mice (P18) brain tissues suggested that the loss of miR-137 led to synaptic defects and increased cell proliferation in brain. Furthermore RNA-seq analyses of the hippocampus from wildtypes, heterozygotes and homozygotes identified *miR-137*$^{+/-}$- and *miR-137*$^{-/-}$-specific differentially expressed genes. Gene ontology analyses indicate that *miR-137*$^{+/-}$-specific differentially expressed genes are enriched with the genes involved in immune system process and glutamate receptor signaling pathway. *miR-137*$^{-/-}$-specific differentially expressed genes are significantly associated with developmental process and regulation of renal sodium excretion. These results together suggest that miR-137 plays important role(s) in neurodevelopment and the dysregulation of MIR137 could contribute to neuropsychiatric disorders in human.

295

A meta-analysis of >16,000 exomes reveals a dominant, highly penetrant subtype of schizophrenia comorbid with intellectual disability. *T. Singh, J.C. Barrett, The UK10K Consortium, The DDD Study.* Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom.

Recent exome sequencing studies have demonstrated that very rare, damaging variants likely play an important role in schizophrenia (SCZ) risk. A significant burden of such variants has been observed in certain neurological pathways, but no individual gene has yet been implicated at genome-wide levels of statistical significance. We have exome sequenced 1,735 SCZ cases (from the UK and Finland, as part of the UK10K project) and 6,789 ancestry-matched controls, and performed joint variant calling and analysis. Consistent with previous studies, we replicate a burden of rare loss-of-function (LoF) variants in specific gene sets, but do not identify any exome-wide significant genes in our data alone.To increase power, we combined our novel discovery set with eight published SCZ exome studies: 2,519 cases and 2,554 controls from Sweden, and 1,077 SCZ trios. We performed a meta-analysis of *de novo* mutations and rare case-control burden in >16,000 exomes using a similar approach as a recently published autism analysis. We found that constrained genes (OR 1.4, $P < 5 \times 10^{-7}$) and targets of FMRP had the strongest enrichment of private LoF variants among tested genesets and pathways. Furthermore, we discovered that LoF variants in a single gene, *KMT2F,* to be significantly associated with SCZ risk ($P = 4.5 \times 10^{-9}$). The number of *KMT2F* LoF variants in the 60,706 ExAC exomes suggests that *KMT2F* is among the most constrained genes in the genome.Phenotypic information in patients carrying *KMT2F* LoF variants revealed comorbid intellectual disability (ID) in addition to classic symptoms of SCZ. This prompted us to query *KMT2F* in 4,281 children with diverse, severe, undiagnosed developmental disorders (DD) exome sequenced as part of the DDD project. We further identified five LoF variants in *KMT2F*, and all five children have DD and ID. Combined, our observations suggest that *KMT2F* LoF mutations confer risk for a dominant, highly penetrant subtype of SCZ comorbid with ID, and implicate epigenetic regulation as an important mechanism in the pathogenesis of SCZ.Using our data and published data, we demonstrate that the excess of *de novo* damaging variants is dramatically stronger in DD than ASD, and stronger in ASD than SCZ, where it is not significantly different than in controls. This result has implications on future studies of neurodevelopmental disorders, as some, like SCZ, may have only a small fraction of cases explained by variants of very large effect.

296

Schizophrenia associated variation in *DPYSL2* perturbs mTOR signaling and produces cellular phenotypes. *X. Pham, L. Liu, S. Guang, R. Wang, H. Zhu, A. Pulver, D. Valle, D. Avramopoulos.* Johns Hopkins University, Baltimore, MD.

Located in a schizophrenia (SZ) susceptibility locus on chr8p21, *DPYSL2* has been implicated in schizophrenia by numerous linkage and association studies; however research on its biology in relation to the risk for SZ has been sparse. *DPYSL2* encodes CRMP2, a protein that functions in axon growth and maintenance. We previously showed that a polymorphic CT di-nucleotide repeat (DNR) variant located in the 5'-untranslated region of *DPYSL2* is associated with SZ risk and disrupts regulation of the gene by mTOR. In dual luciferase assays, the 13-repeat (DNR13) risk allele consistently showed a 3-fold decrease in luciferase activity compared to the 11-repeat (DNR11) common allele in HEK293 cells and mouse primary cortical neurons. The alleles differentially responded to mTOR inhibition in a dose-dependent manner. We have now investigated proteins that bind to each DNR allele to mediate these differences. Using a microarray of >4,000 human transcription factors and proteins, we identified five that preferentially bind to the DNR11 allele. Among these is a ribosomal binding protein, *HuD/ELAVL4,* which is also involved in mTOR signaling and neuronal differentiation. We confirmed *HuD/ELAVL4* binding to DNR11 in an electrophoretic mobility shift assay. Next, we targeted the DNR13 variant mutation into HEK293 cell lines using CRISPR/Cas9 to study the effects of this variant on the cellular phenotype. Our gene-editing scheme resulted in four separate clones homozygous for the DNR13 allele and multiple clones that were targeted (GFP positive) but not modified at the repeat site (DNR11). The *DPYSL2* transcript levels of the DNR13 cell lines were > 2fold decreased compared to 8 different targeted DNR11 cell lines ($p = 1.1 \times 10^{-10}$). Furthermore, the mutant DNR13 cell lines display striking morphological differences. The homozygous DNR13 cells display a clumping phenotype with individual cells showing decreased projection length compared to targeted non-mutant cells and to non-targeted cells. In summary we show significant effects of a naturally occurring and disease associated human DNA variant on the *DPYSL2* transcription and the phenotype of living cells. We also show *in vitro* an effect of the variant on the binding of an mTOR-associated protein, *ELAVL4*, further implicating the mTOR pathway in *DPYSL2* regulation. Follow up work is ongoing to determine if these isogenic cell lines also produce different CRMP2 protein levels and if they differ in mTOR signaling.

297

Somatic mutations in the *MTOR* gene cause focal cortical dysplasia type IIb. *M. Nakashima[1], H. Saitsu[1], N. Takei[2], J. Tohyama[3], M. Kato[4], H. Kitaura[5], M. Shiina[6], H. Sirouzu[7], H. Masuda[7], K. Watanabe[8], C. Ohba[1], Y. Tsurusaki[1], N. Miyake[1], Y. Zheng[5], T. Sato[9], H. Takebayashi[8], K. Ogata[6], S. Kameyama[7], A. Kakita[5], N. Matsumoto[1].* 1) Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan; 2) Department of Molecular Neurobiology, Brain Research Institute, Niigata University, Niigata, Japan; 3) Department of Child Neurology, Nishi-Niigata Chuo National Hospital, Niigata, Japan; 4) Department of Pediatrics, Showa University School of Medicine, Tokyo 142-8555, Japan; 5) Department of Pathology, Brain Research Institute, University of Niigata, Niigata, Japan; 6) Department of Biochemistry, Yokohama City University Graduate School of Medicine, Yokohama, Japan; 7) Department of Functional Neurosurgery, Epilepsy Center, Nishi-Niigata Chuo National Hospital, Niigata, Japan; 8) Division of Neurobiology and Anatomy, Graduate School of Medical and Dental Sciences, Niigata University, Niigata, Japan; 9) Division of Biochemistry, School of Pharmaceutical Sciences, Kitasato University, Tokyo, Japan.

Focal cortical dysplasia (FCD) Type IIb is a cortical malformation characterized by cortical architectural abnormalities, dysmorphic neurons, and balloon cells. It has been suggested that FCDs are caused by somatic mutations in cells in the developing brain. Here, we explore the possible involvement of somatic mutations in FCD Type IIb. We collected a total of 13 blood-brain paired samples with FCD Type IIb. We performed whole exome sequencing using paired samples from nine of the FCD Type IIb subjects and further investigated using all 13 paired samples by deep sequencing. We identified four lesion-specific somatic *MTOR* mutations in six of 13 (46%) individuals with FCD Type IIb showing mutant allele rates of 1.11–9.31%. Functional analyses showed that phosphorylation of ribosomal protein S6 in FCD Type IIb brain tissues with *MTOR* mutations was clearly elevated compared with control samples. Transfection of any of the four *MTOR* mutants into HEK293T cells led to elevated phosphorylation of 4EBP, the direct target of mTOR kinase. These findings suggest that mutations in *MTOR* are likely to cause hyperactivation of the mTOR signaling pathway and induce dysregulation of growth of neurons and glia, or presumably of their progenitors during brain development.

298

Interrogating the mechanisms of schizophrenia genetic risk in the fully characterized human brain transcriptome. *A.E. Jaffe[1,2], J. Shin[1], R.E. Straub[1], R. Tao[1], Y. Gao[1], Y. Jia[1], L. Collado-Torres[1,2], J.T. Leek[2], T.M. Hyde[1,2], J.E. Kleinman[1,2], D.R. Weinberger[1,2].* 1) Lieber Institute for Brain Development, Baltimore, MD; 2) Johns Hopkins University, Baltimore, MD.

Genetic risk for schizophrenia has begun to emerge through large genome-wide association studies (GWAS) in hundreds of thousands of individuals. However, the exact gene(s) and/or transcript(s) that are being regulated by these risk SNPs are largely uncharacterized due to the difficulty in obtaining expression and genotype data in large samples of postmortem human tissue. Advances in RNA sequencing (RNA-seq) have further permitted flexible and largely unbiased characterization of high-resolution transcriptomes, but the incomplete annotation of the human brain transcriptome can potentially affect the ability to use existing tools that rely on complete gene structure information. We have therefore sequenced the transcriptomes of the dorsolateral prefrontal cortex (DLPFC) from 320 non-psychiatric controls across the lifespan at deep coverage, including 50 second trimester fetal samples, and 175 samples from patients with schizophrenia, and characterized their expression profiles across five summarizations that capture elements of transcription – genes, exons, junctions, transcripts, and expressed regions. We show that annotation-agnostic approaches like junction and expressed-region analysis may outperform gene-, exon- and transcript-based approaches when the annotation is incomplete. We further conducted global expression quantitative trait loci (eQTL) analyses across the five expression summarizations in the adult control samples (age > 13, N=237), and identify hundreds of thousands of expression features that associate with local genetic variation, including extensive genetic regulation of previously unannotated sequence. The eQTLs in junction-level data (N= 53,497 unique junctions annotated to 16,481 genes at FDR < 0.01) showed the largest effect sizes (fold change per allele copy) and identified SNPs as eQTLs with the lowest minor allele frequencies (18.1% versus 23.1-24.2%). We lastly identified eQTLs to specific transcript elements in individual genes in over half of the genome-significant genetic variants for schizophrenia identified genome-wide association studies (GWAS), illuminating potential mechanisms of risk for many of these genetic variants. Leveraging human postmortem brain data can therefore fine map the functional effects of genetic risk variation for schizophrenia identified in large GWAS, and can identify novel targets for drug discovery and more focused biological assays.

299

Large-scale exome chip association analysis identifies novel type 2 diabetes susceptibility loci and highlights candidate effector genes. *A. Mahajan on behalf of the ExT2D Exome Chip Consortium, for PROMIS, CHARGE and T2D-GENES/GoT2D.* Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.

To evaluate the contribution of low-frequency and rare coding variants to type 2 diabetes (T2D) risk, we combined exome array data from 232,621 individuals (56,597 cases and 176,024 controls) in 50 studies from five ancestry groups (European, South Asian, African American, East Asian, and Hispanic). Within each study, we tested single variants for association with T2D, with/without body-mass index (BMI) adjustment, using a linear mixed model to account for relatedness and population structure. We combined association summary statistics across studies in a fixed-effects meta-analysis.A total of 44 coding variants, mapping to 26 loci, were associated with T2D at exome-wide significance ($P<5x10^{-7}$) in ancestry-specific or trans-ethnic meta-analysis. All but three were common, with minor allele frequency (MAF)>5%. Thirteen variants were located outside established T2D loci. These included common variants in *ZZEF1* (L1972P, $P=1.9x10^{-9}$), *POC5* (H36R, $P=1.1x10^{-7}$), *PNPLA3* (I148M, $P=1.6x10^{-9}$), and a low-frequency variant (MAF=1% in Europeans) in *FAM63A* (Y285N, $P=6.5x10^{-9}$). Of these, *POC5* maps to a known BMI locus and the *PNPLA3* variant has been implicated in fatty liver disease. Within established T2D loci, nine variants (in *SLC30A8*, *MACF1*, *GCKR*, *PPARG, KCNJ11-ABCC8*, and *PAM-PPIP5K2*) were confirmatory of previously-reported coding alleles that drive association signals. However, the 22 remaining variants have not been directly implicated in T2D susceptibility, so we investigated their relationship with previously reported non-coding lead SNPs through conditional analyses. At the *CILP2* locus, a coding variant in *TM6SF2* (E167K; $P=3.6x10^{-12}$) was indistinguishable from the previously reported lead SNP ($P_{cond}=0.052$), suggesting that the association signal is mediated through this gene. Conversely, the association for *GIPR* E354Q ($P=1.1x10^{-8}$) was not eliminated after conditioning on the previously reported inter-genic lead SNP ($P_{cond}=4.0x10^{-6}$), implying these signals to be distinct. Our results indicate that low-frequency and rare coding variants of large effect do not make a major contribution to T2D risk. However, these analyses implicate several novel genes in T2D pathogenesis, and provide direct insight into the underlying biology of the disease.

300

Large scale exome array meta-analyses identify numerous novel common, low-frequency and rare coding variant associations with glycaemia. *S. Willems on behalf of the Meta-Analyses of Glucose and Insulin-related traits Consortium.* MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, United Kingdom.

Protein coding single nucleotide variants (SNVs) have been hypothesized to contribute to variation in complex traits and diseases. Furthermore, relative to non-coding or intergenic SNVs, coding variants may facilitate identification of effector transcripts and thus facilitate biological inference. To explore the role of coding variants in glycaemic traits, we investigated the association of 241,320 common (minor allele frequency (MAF)>5%), low frequency (MAF 1-5%) and rare (MAF<1%) exome array variants with fasting glucose (FG), fasting insulin (FI), 2-h glucose (2hGlu) and HbA1c levels. We included up to 137,949 non-diabetic individuals of European (88.8%, N=122,464), African American (5.9%, N=8,135), South Asian (1.7%, N=2,377), East Asian (1.9%, N=2,644) and Hispanic (1.7%, N=2,327) ancestry from up to 65 studies. We combined single variant results from linear mixed models by fixed-effect meta-analyses. 45 loci contained exome-wide significant ($P<2.07x10^{-7}$) associations with FG, including 13 newly associated loci. Among 59 loci containing associations with HbA1c, 18 were novel. For FI and 2hGlu we found 18 and 14 associated loci, respectively, including two novel loci associated with FI and one with 2hGlu. Among the lead SNVs at newly associated loci, 18 were missense variants, including two low-frequency and five rare SNVs. These included a rare variant (polymorphic in 16 studies, overall MAF 0.2%) in *AKT2* (previously implicated in monogenic insulin resistance), associated with FI levels (P50T, β =0.12 log-pmol l^{-1}, $P=3.1x10^{-10}$) and rare variants in *ANKH* and *MAP3K15* associated with FG (*ANKH*: R187Q, MAF=0.4%, β= -0.09 mmol l^{-1}, $P=6.7x10^{-10}$; *MAP3K15*: G838S, MAF=0.3%, β=-0.09 mmol l^{-1}, $P=1.5x10^{-11}$). Common missense SNVs in novel loci associated with HbA1c included M708I in *EGF* (MAF=40.9%, β=-0.007%, $P=7.7x10^{-8}$) and A265V in *MLXIPL* (MAF=11.8%, β =0.01%, $P=5.5x10^{-8}$). *MLXIPL* maps to a known lipid-associated locus. We did not identify missense variants among lead SNVs in novel loci associated with 2hGlu, likely attributable to the smaller sample size (N=57,878). In conclusion, we identified coding variants across the allele frequency spectrum in loci associated with glycaemic traits, further increasing insight into genetic architecture and novel biology underlying these traits.

301

Causal mechanisms and balancing selection inferred from genetic associations with the Polycystic Ovary Syndrome. *F. Day[1], D.A. Hinds[2], J.Y. Tung[2], L. Stolk[3], U. Styrkarsdottir[4], R. Saxena[5], A. Bjonnes[5], L. Broer[3], D.B. Dunger[6], B.V. Halldorsson[4, 7], D.A. Lawlor[8, 9], G. Laval[10], I. Mathieson[11], W.L. McCardle[9], Y. Louwers[12], C. Meun[12], S. Ring[8, 9], R.A. Scott[1], P. Sulem[4], A.G. Uitterlinden[3], N.J. Wareham[1], U. Thorsteinsdottir[4, 13], C. Welt[14], K. Stefansson[4, 13], J.S.E. Laven[12], K.K. Ong[1, 6], J.R.B. Perry[1].* 1) MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom; 2) 23andMe Inc., Mountain View, California, USA; 3) Department of Internal Medicine, Erasmus MC, 3015 GE Rotterdam, the Netherlands; 4) deCODE Genetics / Amgen, Sturlugata 8, IS-101 Reykjavik, Iceland; 5) Department of Anaesthesia and Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA; 6) Department of Paediatrics, University of Cambridge School of Clinical Medicine, Box 181, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK; 7) Institute of Biomedical and Neural Engineering, School of Science and Engineering, Reykjavìk University, Menntavegur 1, 101 Reykjavìk, Iceland; 8) MRC Integrative Epidemiology Unit at the University of Bristol, Bristol BS8 2BN, UK; 9) School of Social and Community Medicine, University of Bristol, Oakfield House, Bristol BS8 2BN, UK; 10) Human Evolutionary Genetics, CNRS URA3012 Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris Cedex 15, France; 11) Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; 12) Division of Reproductive Medicine, Department of Obstetrics & Gynaecology, Erasmus MC, 3015 GE Rotterdam, the Netherlands; 13) Faculty of Medicine, University of Iceland, IS-101 Reykjavik, Iceland; 14) Division of Endocrinology, Metabolism and Diabetes, University of Utah School of Medicine, Salt Lake City, UT 84112, USA.

Polycystic Ovary Syndrome (PCOS) is the most common reproductive disorder in women, yet there is little consensus regarding its aetiology. Furthermore, the high prevalence of a heritable condition characterised by impaired fertility presents an evolutionary paradox. To understand the biological and genetic basis of PCOS, we performed a dense genome-wide association study of ~9M 1000 Genomes reference haplotype imputed variants in up to 5,184 self-reported PCOS cases of white European ancestry and 82,759 controls, with follow-up in a further ~2000 clinically-validated cases and ~100,000 controls. We identified six signals for PCOS at genome-wide statistical significance ($P<5x10^{-8}$), in/near genes *ERBB4/HER4*, *YAP1*, *THADA*, *FSHB*, *RAD50* and *KRR1*. Effect sizes ranged from per-allele odds ratios 1.12–1.37 in the follow-up studies, with no significant heterogeneity by PCOS case definition. Variants in/near three of the four epidermal growth factor receptor genes (*ERBB2/HER-2, ERBB3 and ERBB4)*, which encode targets of cancer chemotherapy, were associated with PCOS at or near genome-wide significance, highlighting potentially novel pharmaceutical targets. Mendelian randomisation analyses indicated causal roles in PCOS for higher BMI ($P=2.5x10^{-9}$), higher insulin resistance ($P=6x10^{-4}$) and lower serum sex hormone binding globulin levels ($P=5x10^{-4}$). Furthermore, genetic susceptibility to later menopause was associated with higher PCOS risk ($P=1.6x10^{-8}$) and PCOS risk-increasing alleles were associated with higher serum anti-Müllerian hormone levels in girls ($P=8.9x10^{-5}$), although no evidence could be demonstrated for positive selection of these alleles. This first large-scale genetic study of PCOS in white Europeans implicates an aetiological role of the epidermal growth factor receptors, infers causal mechanisms relevant to clinical management and prevention, and suggests balancing selection mechanisms involved in PCOS risk.

302

Genetic contributions to long-term severe obesity and leanness defined by electronic medical record phenotyping – a genome-wide association study. *C. Schurmann, N.S. Abul-Husn, E.P. Bottinger, R.J.F. Loos.* Icahn School of Medicine at Mount Sinai, New York, NY.

Genetic factors explain 40-70% of inter-individual variation in obesity susceptibility. The heritability for the more extreme forms of obesity has been estimated to be even higher. However, only a few genome-wide association studies (GWAS) have focused on these phenotypes so far and it remains unclear whether *severe* and *common* obesity share the same genetic architecture.In an effort to characterize genetic variation contributing to severe obesity risk, we used electronic medical records (EMRs) of the Bio*Me* Biobank to define long-term severe obese cases (BMI≥35kg/m^2) and long-term lean controls (BMI≥18.5 kg/m^2 and <23kg/m^2). Specifically, individuals with three or more BMI measurements (median: n=17, range: 3-391) across at least one year (median: 5yrs, range: 1-17yrs) within the permissible range for cases and controls were included (European ancestry (EA), n=222, 30% cases; African ancestry (AA), n=582, 81% cases; Hispanic ancestry (HA), n=573, 73% cases). We then performed a GWAS of up to 16M genotyped and imputed variants (MAF>1%) comparing cases and controls, adjusted for age, sex, and PCs, stratified by ancestry.We identified 5, 29 and 3 loci harboring genome-wide significant variants (p<5e-8) for EA, AA, and HA, respectively. These included variants overlapping known adiposity loci, as well as several novel loci. The novel loci include common variants near the leucine-rich repeat kinase 1 gene (*LRRK1*, rs12903795, MAF=15.6%, p=1.6e-9) and an intronic low-frequency variant in the glutamate receptor, metabotropic 7 gene (*GRM7*, rs139984586, MAF=1.8%, p=4.6e-8) in EA. *LRRK1* is a homolog of *LRRK2*, a gene harboring known disease-causing variants for Parkinson's disease. *GRM7* encodes a neurotransmitter that modifies synaptic transmission and neuronal excitation. In HA, we identified variants near the synuclein alpha interacting protein coding gene (*SNCAIP*, rs114698063, MAF=1.4%, p=1.5e-8) associated with long-term leanness. Overexpression of the human *SNCAIP* gene has been shown to result in increased body weight in mice and obesity in flies.Using EMR data, we defined long-term extreme body weight phenotypes. Despite the small sample size, our association analyses were successful in identifying novel, biologically-plausible loci, which point to the involvement of the central nervous system in extreme body weight phenotypes. Ongoing analyses include an expansion of the sample size, replication of the findings and functional follow-up.

303

Integrative Personal Omics Profiling During Periods of Disease, Weight Gain and Loss. *M. Snyder[1,4], B. Piening[1,4], W. Zhou[1], K. Contrepois[1], G. Gu[1], S. Leopold[3], K. Kukurba[1], T. Mishra[1], C. Craig[2], D. Perleman[2], E. Sodergren[3], B. Leopold[3], T. McLaughlin[2], G. Weinstock[3].* 1) Genetics, Stanford University, Stanford, CA; 2) Endocrinology, Stanford University, Stanford, CA; 3) The Jackson Laboratory for Genomic Medicine, Farmington, CT; 4) Stanford Cardiovascular Institute, Stanford, CA.

While significant genetic and environmental risk factors are known that contribute to the development of Type 2 Diabetes (T2D), overall our ability to predict which individuals will eventually develop T2D and when this will occur is woefully inadequate. To better understand these factors, we present a longitudinal multi-omic personalized medicine pipeline for the comprehensive molecular profiling of blood- and microbiome-based analytes that we apply to track the progression to T2D in a cohort of 75 individuals over periods of health, illness and weight gain and loss. Multi-omic profiling (transcriptome, DNA methylome, proteome, metabolome etc.) revealed significant differences in multiple 'omes between prediabetics and healthy controls at steady state, implicating pathways related to chronic inflammation and insulin regulation as well as novel connections to T2D. A subset of participants was then placed on a short-term high caloric diet, followed by additional multi-omic profiling. The dietary perturbation was associated with a wealth of biomolecular expression changes concomitant with weight gain and spanning multiple 'omes including the microbiome, and the omic response to weight gain differed between prediabetics and healthy controls. For another subset of participants who went through respiratory viral infections, their multi-omic profiling, including the microbiome, responded distinctly to different illness stages during the infection. Overall, the multi-omic profiles of individuals are unique compared to others regardless diet or illness perturbations. In total, these large-scale longitudinal data offer a novel and comprehensive view of the dysfunction in cellular networks associated with the progression to T2D and may offer new strategies for predicting and preventing the disease.

304

Causal FTO Obesity Variant Represses Adipocyte Browning in Humans. *M. Kellis.* MIT and Broad Institute, Cambridge, MA.

Genome-wide association studies can uncover disease-relevant genomic regions, but interpretation is challenging. The *FTO* region harbors the strongest genetic association with obesity, yet its mechanistic basis remains elusive.Here, we use epigenomics, allelic activity, motif conservation, and directed perturbations in patient samples and in mice, and endogenous CRISPR/Cas9 genome editing in patients, aiming to dissect the regulatory circuitry and mechanistic basis of the *FTO* region association with obesity.Our data showed that the rs1421085 T-to-C single-nucleotide polymorphism (44% frequency in Europeans) in the *FTO* obesity risk locus, was associated with the disruption of a conserved ARID5B repressor motif in risk-allele carriers, resulting in de-repression of a potent preadipocyte enhancer, which doubled *IRX3* and *IRX5* expression during early adipocyte differentiation. This de-repression resulted in a cell-autonomous developmental shift from energy-dissipating beige/brite adipocytes to energy-storing white adipocytes, with 4-fold reduced mitochondrial thermogenesis and increased lipid storage. Adipose inhibition of *Irx3* in mice reduced body weight and increased energy dissipation, with unchanged physical activity or appetite. Knockdown of *IRX3* or *IRX5* in primary adipocytes restored 7-fold higher thermogenesis in risk-allele participants, and overexpression led to its 8-fold disruption in protective-allele participants. Repair of the ARID5B motif by CRISPR/Cas9 editing of rs1421085 in primary adipocytes from a risk-allele patient restored *IRX3* and *IRX5* repression, activated browning expression programs, and restored 7-fold higher thermogenesis.The present results point to a pathway for adipocyte thermogenesis regulation involving ARID5B, rs1421085, IRX3, and IRX5, whose manipulations showed pronounced pro-obesity and anti-obesity effects.

305

ExomeChip meta-analysis of 526,508 individuals from five ancestries identifies novel coding variation associated with body mass index. *H.M. Highland[1,2], V. Turcot[3,4], Y. Lu[5], C. Schurmann[5], A.E. Justice[1], K.L. Young[1], J. Wang[6], P. Lenzini[6], M. Graff[1], A.L. Cupples[7], T.M. Frayling[8], J.N. Hirschhorn[9,10,11], G. Lettre[3,4], C.M. Lindgren[12], K.E. North[1,13], I.B. Borecki[6], R.J.F. Loos[5] for the BBMRI, GoT2D, CHARGE, and GIANT Consortia.* 1) Dept of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Human Genetics Center, University of Texas Health Science Center, Houston, TX, USA; 3) Montreal Heart Institute, Montréal, Québec, Canada; 4) Faculty of Medicine, Université de Montréal, Montréal, Québec, Canada; 5) The Genetics of Obesity and Related Metabolic Traits Program, The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 6) Department of Genetics Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA; 7) Boston University School of Public Health, Boston, MA, USA; 8) University of Exeter Medical School, Exeter, UK; 9) Divisions of Endocrinology and Genetics and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA; 10) Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA; 11) Department of Genetics, Harvard Medical School, Boston, MA, USA; 12) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 13) Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

High body mass index (BMI) increases risk of chronic diseases including type 2 diabetes (T2D), heart disease and stroke. To date, 97 genetic variants, mostly common (MAF>5%) and non-coding, have been identified for BMI through genome-wide association studies (GWAS). We investigated the role of coding variants with putative functional effects (missense, nonsense, splicing, stop gain), using ExomeChip data of 526,508 individuals from 163 strata, predominantly of European ancestry. Individual studies performed association analyses using Rvtest or RareMetalWorker; results were subsequently combined in single variant meta-analyses and gene-based tests (SKAT, VT) stratified by ancestry using RareMETALS. Gene-based tests were grouped based on functional annotation and *in silico* predicted impact on proteins for variants with a MAF<5%. Among the 245,497 variants analyzed, 68 coding variants outside (≥1Mb) the 97 GWAS-identified BMI loci, reached array-wide significance (P<2E-07) including a low frequency variant in *ACHE* (rs1799805, p.H353N, p=8.01E-10, EAF=4%, β(SE)=0.032(0.005) SD/allele) and a common nonsynonymous variant in *PSMD2* (rs11545169, p.E183D, p=1.16E-17, EAF=15%, β(SE)=-0.017(0.003) SD/allele). *PSMD2* is required for ubiquitin dependent Insig1 degradation. Eight gene-based associations were significant (p<2.5E-06), including *GIPR* (p=7.17E-09), whose effects were independent of the common GWAS proxy in the gene (rs1800437, p. E354Q, p=7.91E-30, MAF=20%, β(SE)=-0.029(0.003) SD/allele). *GIPR* is part of the *GLP1* pathway responsible for the secretion of postprandial insulin. Other gene-based associations were largely driven by a single variant, such as for *KSR2* (p=7.15E-09), *RAPGEF3* (p=8.91E-15) and *PRKAG1* (p=2.75E-12). *Ksr2* knockout mice are obese. Multiple variants in *KSR2* have been shown to associate with severe early onset obesity. Mutations in *KSR2* disrupt the Raf-MEKERK pathway, resulting in impaired fatty acid oxidation and glucose oxidation. In the case of *KSR2*, aggregating rare and low frequency variants in a gene-based test enabled detection of the association with BMI. *RAPGEF3* is involved in the *GLP1* pathway and regulates Akt2 response to insulin. *PRKAG1* is involved in AMPK signaling and has previously been associated with T2D and weight gain on antipsychotic medication. Taken together, by interrogating coding variants in a large sample set, we have identified both additional genes and variants associated with BMI, some of which may be causal.

306

Clustering of exome variants from 6272 individuals with Type 2 diabetes identifies etiological convergence amongst four distinct populations and with other T2D genetic risk factors. *C. Sandor[1], N. Beer[2,3], J. Fernandez[2,3], F. Honti[4], J. Steinberg[5], M. McCarthy[2,3], C. Webber[1], T2D-GENES Consortium, GoT2D Consortium.* 1) MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK; 2) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 3) Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK; 4) University of Lausanne, Lausanne, Switzerland; 5) Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge, UK.

Type 2 Diabetes (**T2D**) is a global health burden. Efforts to uncover predisposing genetic variation have been considerable, yet our knowledge of the underlying heritability and complex aetiology remains poor. To identify genes that influence the same phenotype, multiple sources of information about gene function (e.g. co-expression, protein interactions, etc) can be integrated into a single measure of functional similarity between genes, forming a phenotypic-linkage network (**PLN**). Here, we constructed a **T2D-PLN**, by integrating only gene functional information that is informative for identifying genes that influence T2D-relevant phenotypes. This new approach identified a convergent set of biological pathways that were perturbed within each of four independent T2D case/control ethnic sets of ~2K exomes each, and that the same pathways were over-represented among both known monogenic or syndromic diabetes genes and genes within T2D-associated common risk loci.Compared to our standard PLN, the T2D-PLN was more sensitive and specific in identifying functional associations between 32 monogenic or syndromic diabetes diabetes genes and both (i) genes mapping to 72 established T2D Genome-wide association study (**GWAS**) loci and (ii) genes within 614 nominally-associated ($p < 10^{-3}$) T2D GWAS loci. Amongst the protein-truncating variant (**PTV**) possessing genes found to be most associated with T2D in the exomes of 12,940 individuals from five ethnic sets, the T2D-PLN identified an excess of functionally-convergent genes amongst four ancestral exome sets and were found to have roles in lipid and glucose metabolism ($q = 0.02$; functional convergence was identified amongst the Hispanic, African-American, South Asian and East Asian sets but not European). Despite little overlap in individual genes carrying PTVs between the different exome ancestral sets, 3/5 sets' constituent genes demonstrated significant inter-set functional clustering within theT2D-PLN ($q < 0.02$), and were found to functionally cluster with genes mapping to T2D GWAS loci. The GTEx tissue-specific expression patterns and other orthogonal functional information were consistent between genes identified across these different genetic risk associations, strikingly so between the monogenic/syndromic diabetes genes and the exome PTV-possessing genes.Taken together, the T2D-PLN was able to identify convergent pathways perturbed in T2D by non-coding and coding variants from multiple independent samples.

307

Genetic study identifies common variation in *PHACTR1* to associate with fibromuscular dysplasia. *N. Bouatia-Naji[1,2], SR. Kiando[1,2], N. Tucker[3], A. Katz[4], C. Treard[1,2], V. D'Escamard[5], LJ. Castro-Vega[1,2], C. Ye[6], E. Smith[6], C. Austin[6], C. Barlasina[7], D. Cusi[7], P. Galan[8], JP. Empana[1,2], X. Jouven[1,2,9], P. Bruneval[1,2], JW. Olin[5], HL. Gornik[10], PF. Plouin For ARCADIA[1,2,11], IJ. Kullo[6], DJ. Milan[3], SK. Ganesh[4], P. Boutouyrie[1,2,12], J. Kovacic[5], X. Jeunemaitre[1,2,13].* 1) Paris Cardiovascular Research Center, INSERM UMR970 PARCC, Paris, France; 2) Faculty of medicine, Paris-Descartes University, Sorbonne Paris Cité, Paris, FRANCE; 3) Cardiovascular research Center, Massachusetts General Hospital, Charlestown, MA, USA; 4) Department of Internal Medicine and Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA; 5) Zena and Michael A. Wiener Cardiovascular Institute & Marie-Josée and Henry R. Kravis Center for Cardiovascular Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 6) Department of Medicine, Division of Cardiovascular Diseases, Mayo Clinic, Rochester, Minnesota, USA; 7) Dept. of Health Sciences, Genomic and Bioinformatics Unit, School of Nephrology, University of Milano, Institute of Biomedical Technologies, Italian National Centre of Research, ITALY; 8) Nutritional Epidemiology Research Group, Sorbonne-Paris-Cité, UMR University of Paris 13/Inserm U-557/INRA U-1125/CNAM, Bobigny, France F-93017, Bobigny, FRANCE; 9) AP-HP, Department of Cardiology, Hôpital Européen Georges Pompidou, Paris, FRANCE; 10) Cleveland Clinic Heart and Vascular Institute, Cleveland, OH, USA; 11) AP-HP, Department of Hypertension, Hôpital Européen Georges Pompidou, Paris, FRANCE; 12) AP-HP, Department of Pharmacology, Hôpital Européen Georges Pompidou, Paris, FRANCE; 13) AP-HP, Refferal Center for Rare Vascular Diseases, Hôpital Européen Georges Pompidou, Paris, FRANCE.

Fibromuscular dysplasia (FMD) is a non-atherosclerotic vascular disease leading to arterial stenosis, aneurysm and dissection. The renal and cerebrovascular arteries are most commonly involved. FMD has higher prevalence in females (80-90%) and is associated with hypertension and stroke. The pathophysiology of FMD is unclear and a genetic origin is suspected.We performed a three-stage genetic association study in cases and population-based controls of European ancestry. The discovery stage included 249 FMD patients (66% renal) and 689 controls, in which we analyzed ~26K common variants (MAF>0.05) using an exome-chip array. We followed-up 13 independent loci with suggestive association with FMD ($P<10^{-4}$) in 393 cases and 2537 controls replicated a signal on Chr6. Three additional case-control studies (combined N_{cases}=512, $N_{controls}$=669) confirmed this association, with an overall odds ratio of 1.39, 95% confidence interval=1.25-1.54, $P=7.4\times10^{-10}$ in a global sample of N_{cases}=1154, $N_{controls}$=3895.The FMD risk variant is intronic to the phosphatase and actin regulator 1 gene (*PHACTR1*), involved in angiogenesis and cell migration. *PHACTR1* is a risk locus for coronary artery disease, migraine, and cervical artery dissection, which may occur in FMD patients. We found a significant association between the FMD risk allele and higher central pulse pressure (P=0.0009), increased intima media thickness (P=0.001) and wall cross-sectional area (P=0.003) of carotids assessed by echotracking in 3800 population-based individuals. The correlation of genotypes with the expression of *PHACTR1* in primary cultured human fibroblasts showed higher expression in FMD risk allele carriers, compared to non-carriers (N=57, P=0.02). Antibodies detected PHACTR1 in paraffin-fixed sections of renal arteries, both from healthy and FMD patients. Finally, *Phactr1* knockdown of zebrafish showed significantly dilated vessels (9 morphants vs. 9 controls, P=0.003) indicating impaired vascular development.In conclusion, we report the first risk locus for FMD with the largest genetic association study conducted so far. Our data reveal a common genetic variant at *PHACTR1* supporting a complex genetic pattern of inheritance and indices of shared pathophysiology between FMD and other cardiovascular and neurovascular diseases.

308

A rare synonymous variant in *GFI1B* associated with lower platelet counts reveals a role for alternative *GFI1B* splice forms in human hematopoiesis. *P. Auer[1], R. Khajuria[2, 3], J. Huang[4], U. Schick[5, 6], J. Johnsen[7, 8], N. Soranzo[4], A. Reiner[9], V. Sankaran[2, 3].* 1) Department of Biostatistics, University of Wisconsin-Milwaukee, Milwaukee, WI; 2) Division of Hematology and Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts; 3) Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts; 4) Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; 5) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY; 6) The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, NY; 7) Bloodworks Research Institute - Puget Sound, Seattle, WA; 8) Department of Medicine, University of Washington, Seattle, WA; 9) Department of Epidemiology, University of Washington, Seattle, WA.

Whole genome sequencing (WGS) followed by imputation can be used to interrogate the association of low-frequency and rare variants with complex hematologic traits in a comprehensive manner. By imputing low-depth (8X) WGS data from the UK10K project into a total sample of approximately 35,000 samples from 10 European and American studies, we identified a rare synonymous variant in *GFI1B* (rs150813342, allele frequency = 0.8%) associated with 30,000/uL lower platelet count ($P=5 \times 10^{-18}$) per allele, but normal red cell and white cell parameters. *GFI1B* encodes a transcription factor involved in the regulation of hematopoiesis, which is required for normal red blood cell (erythroid) and platelet (megakaryocyte) production. Recently, rare loss-of-function mutations of *GFI1B* have been identified in patients with an autosomal dominant form of Gray platelet syndrome, a bleeding disorder characterized by macrothromobocytopenia and platelet alpha-granule deficiency. This synonymous variant is located within a putative SRSF1 exonic splicing enhancer within an alternatively spliced exon that results in production of two *GFI1B* transcript isoforms, long and short. We performed CRISPR/Cas9 genome editing to create isogenic K562 cell lines harboring either the synonymous variant or wild type alleles. K562 cells have both erythroid and megakaryocyte potential. We found that cells expressing the rs150813342 synonymous variant had suppressed formation of the long isoform of *GFI1B* and increased expression of the short isoform. Surprisingly, based on surface phenotypes, hemoglobinization, and marker gene expression, erythroid differentiation was preserved in the presence of the synonymous variant. In contrast, megakaryocyte differentiation was impaired with an over 10-fold reduction in formation of CD41a positive cells and concomitantly reduced megkaryocyte marker gene expression. Taken together, these results suggest that the long isoform plays a key role in platelet formation and megakaryopoiesis, while being dispensable for erythropoiesis. These results demonstrate how rare human genetic variation - in particular a synonymous variant that otherwise may have been considered unlikely to be functional - can provide important and unanticipated insight into human hematopoiesis and understanding the genetic basis of complex traits.

309

Meta-analysis of rare and common exome chip variants identifies and replicates *S1PR4* and other novel genes influencing blood cell traits. *N. Pankratz on behalf of the CHARGE Hematology Working Group.* University of Minnesota, Saint Paul, MN.

Hematologic measures such as hemoglobin, hematocrit and white blood cell (WBC) count are heritable traits that are clinically important indicators of hematologic, inflammatory, and infectious diseases, as well as general health status. Using the Illumina HumanExome BeadChip, we analyzed erythrocyte and WBC phenotypes in 52,531 individuals (37,775 of European ancestry; 11,589 African Americans; 3,167 Hispanic Americans) from 16 population-based cohorts in our discovery set and 18,018 European American women in our replication set. We identified and replicated four novel erythrocyte trait-locus associations (*CEP89*, *SHROOM3*, *FADS2*, and *APOE*) and six novel WBC loci for neutrophil count (*S1PR4*), monocyte count (*BTBD8*, *NLRP12,* and *IL17RA*), eosinophil count (*IRF1*), and total WBC (*MYB*). In addition, rare missense variants within genes previously linked to erythrocyte traits (*ANK1* and *ITFG3*) were significantly associated with the same traits but were independent of the common variation identified through genome-wide association studies (GWAS). The four variants with novel erythrocyte associations were all common variants that had previously been associated with either kidney traits (*CEP89*, *SHROOM3*) or fatty acid traits (*FADS2*, *APOE*) in prior GWAS, demonstrating pleiotropic effects for these loci. The missense variant in *S1PR4* was the only novel association that was rare (minor allele frequency <1%), so we performed additional follow-up of this gene in two model organisms. Loss-of-function experiments of *S1pr4* in mouse and zebrafish demonstrated phenotypes consistent with the association observed in humans (20-30% decrease in neutrophil counts) as well as altered kinetics of neutrophil recruitment and resolution in response to tissue injury. These findings support the role of sphingosine-1-phosphate signaling in leukocyte trafficking and circulating neutrophil counts and demonstrate the utility of the exome array genotyping approach to identify novel genetic determinants of hematologic traits.

310

Parent of Origin Genome-Wide Association Studies (GWAS) of Cardiovascular Disease (CVD) Associated Phenotypes in the Hutterites. *S.V. Mozaffari[1], J. DeCara[2], S. Shah[3], C. Herman[1], R. Lang[2], D. Nicolae[2,4], C. Ober[1].* 1) Department of Human Genetics, University of Chicago, Chicago, IL; 2) Department of Medicine, University of Chicago, Chicago, IL; 3) Department of Medicine, Northwest University Feinberg School of Medicine, Chicago, IL; 4) Department of Statistics, University of Chicago, Chicago, IL.

GWAS typically treat maternal and paternal alleles as equivalent, although it is possible that sequence variants can affect traits differently depending on whether they are inherited from the father or the mother. To explore this possibility, we performed parent of origin GWAS of CVD associated traits in a large Hutterite pedigree using ~5M variants (MAF>5%) imputed from Hutterite whole genome sequences. The Hutterites are a founder population of European descent that live on communal farms in the northern plains states and western Canada. The Hutterites of S. Dakota, the subjects of our studies, are descendants of 64 founders. We performed a parent of origin GWAS of systolic blood pressure (SBP) and diastolic pressure (DBP) (N=807) and carotid intima media thickness (CIMT) (N=547) using linear mixed models (GEMMA), correcting for the relatedness among individuals by including kinship as a random effect. CIMT is a biomarker for atherosclerosis, which is a leading cause of heart attacks, stroke, and peripheral vascular disease. Our studies revealed a significant association between SBP and a paternally-inherited allele on chromosome 13 upstream of an uncharacterized long intergenic noncoding RNA, *LINC01055* (rs1536182, P=3x10^{-8}). This same paternally-inherited allele was associated with DBP at near genome-wide level of significance (P=1.5x10^{-6}). The maternally-inherited allele at this SNP or at any SNPs within 100kb of rs1536182 were not associated with either SBP or DBP (rs1536182 SBP P= 0.56, DBP P=0.18). In contrast, the maternally-inherited allele at an intronic variant in another noncoding RNA, *LINC00607*, on chromosome 2 was associated with CIMT (rs4077567, P=7.9x10^{-8}); the paternally-inherited allele at this SNP or at any SNPs within 100kb were not associated with CIMT (rs1536182 P=0.57). These findings are particularly intriguing because long intergenic noncoding RNAs are commonly found at imprinted loci, which show parent of origin effects, and suggest that these parent of origin specific associations with blood pressure and CIMT may be due to the presence of imprinted genes at these loci. Such loci would not be detected in GWAS that consider maternal and paternal alleles as equivalent and raise the possibility that previously unknown imprinted loci may account for a portion of the missing heritability of these phenotypes and contribute toward genetic risk of CVD.

311

Rare genetic variants inform on blood pressure pathophysiology. *P. Surendran[1], F. Drenos[2,3], R. Young[1], H. Warren[4,5], J.P. Cook[6,7], A.K. Manning[8,9,10], N. Grarup[11], X. Sim[12,13], D. Saleheen[1,16,18], F.W. Asselbergs[14,19,20], C.M. Lindgren[9,15,21], J. Danesh[1,17], L.V. Wain[6], A.S. Butterworth[1], J.M.M. Howson[1], P.B. Munroe[4,5], CHARGE+, T2D-GENES, GoT2DGenes, ExomeBP, CHD Exome+ Consortium.* 1) Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 2) MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK; 3) Centre for Cardiovascular Genetics, Institute of Cardiovascular Science, Rayne Building University College London, London, WC1E 6JF, UK; 4) Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, EC1M 6BQ,UK; 5) NIHR Barts Cardiovascular Biomedical Research Unit, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, EC1M 6BQ, UK; 6) Department of Health Sciences, University of Leicester, Leicester, LE1 7RH, UK; 7) Department of Biostatistics, University of Liverpool, Liverpool, L69 3GA, UK; 8) Department of Genetics, Harvard Medical School, Boston, MA 02138, USA; 9) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 10) Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA; 11) The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; 12) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA; 13) Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, Singapore; 14) Department of Cardiology, University Medical Center Utrecht, Utrecht, Netherlands; 15) Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK; 16) Department of Biostatistics and Epidemiology, University of Pennsylvania, USA; 17) MRC Human Genetics Unit, MRC Institute of G 107 Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU; 18) Center for Non-Communicable Diseases, Karachi, Pakistan; 19) Durrer Center for Cardiogenetic Research, ICIN-Netherlands Heart Institute, Utrecht, The Netherlands; 20) Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, United Kingdom; 21) The Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7BN, UK.

High Blood Pressure (BP) is a major risk factor for cardiovascular disease and premature death. However, there is limited evidence about the specific causal genes and variants involved in blood pressure regulation. To better understand the genetic basis of blood pressure, we analysed ~250,000 rare, low frequency and common single nucleotide variants (SNVs) in the largest Pan-ethnic BP genetic study to date of up to 350,000 individuals. We discovered and validated 32 novel genetic regions associated with systolic BP, diastolic BP, pulse pressure or hypertension through SNV and gene-based association tests. Conditional analyses identified multiple independent signals in novel and known BP associated regions including a nonsense SNV in *ENPEP* associated with increased BP. Gene expression data combined with functional evidence, implicated *CERS5* as potentially causal for BP with the SNVs in this region showing a strong association with expression of *CERS5* in multiple tissues.For the first time, rare missense SNVs (MAF<0.01) with strong effects on BP were identified, providing compelling support for candidate genes in these regions. We describe the value of these findings in discovering biological pathways that link blood pressure genes to vascular remodelling and cardiac abnormalities suggesting a shared genetic aetiology. The rare nonsense SNV in the highly conserved region of *ENPEP* is predicted to cause nonsense mediated decay of the transcript and the protein, aminopeptidase A (APA) that converts Angiotensin II (AngII) to AngIII. Association of the rare nonsense *ENPEP* SNV with increased blood pressure indicates activation of AT1 receptor by AngII in the peripheral vascular system. We discuss the potential therapeutic implications of the *ENPEP* BP association for the use of APA inhibitors as blood pressure lowering drugs. Together, our results pinpoint causal genes and highlight the role of rare SNVs in blood pressure regulation and hypertension.

312

Meta-analysis of gene-smoking interactions in blood pressure using 1000 Genomes imputed data from four ethinc groups. *Y. Sung[1], T. Winkler[2], A. Bentley[3], M. Brown[4], T. Bartz[5], D. Chasman[6], R. Dorajoo[7], M. Fornage[8], N. Franceschini[9], X. Guo[10], C. Hayward[11], S. Kardia[12], K. Lohman[13], R. Loos[14], J. Marten[11], B. Tayo[15], C. van Duijn[16], W. Xu[17], I. Borecki[18], L. Cupples[19], D. Rao[1], A. Morrison[4] on behalf of the CHARGE Gene-Lifestyle Interactions Working Group.* 1) Div Biostatistics, Washington Univ, St Louis, St Louis, MO; 2) Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; 3) Center for Research on Genomics and Global Health, NHGRI, NIH, Bethesda, MD, USA; 4) Department of Epidemiology, Human Genetics, and Environmental Sciences, University of Texas Health Science Center, Houston TX, USA; 5) Cardiovascular Health Research Unit, Department of Biostatistics and Medicine, University of Washington, Seattle, WA, USA; 6) Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA; 7) Genome Institute of Singapore, Agency for Science Technology and Research, Singapore; 8) Brown Foundation Institute of Molecular Medicine, University of Texas Health Science Center, Houston, TX, USA; 9) Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA; 10) Division of Genomic Outcomes, Department of Pediatrics, LABioMed at Harbor-UCLA Medical Center, Torrance, CA, USA; 11) MRC Human Genetics Unit, IGMM, University of Edinburgh, Edinburgh, UK; 12) Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA; 13) Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA; 14) The Charles Bronfman Institute for Personalized Medicine, The Icahn School of Medicine at Mount Sinai, New York, NY, USA; 15) Stritch School of Medicine, Loyola University Chicago, Chicago, IL, USA; 16) Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, Netherlands; 17) Life Sciences Institute, National University of Singapore, Singapore, Singapore; 18) Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA; 19) Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.

Introduction: The management of cardiovascular disease and associated risk factors including blood pressure (BP) are a major public health problem. Accounting for gene-environment interactions are important for identifying novel loci not detected through main-effects-only analyses.**Methods**: We investigated the role of smoking in the genetic and environmental architecture of BP, using 1000 Genomes imputed data from 4 ethnic groups: 21 European-ancestry (EA) cohorts (N=80,551), 14 African-ancestry (AA) cohorts (N=25,887), 9 Asian cohorts (N=13,438) and 3 Hispanic American (HA) cohorts (N=8,805) with an overall total of N=128,687. For each ethnic group, cohort-specific results were combined to perform the 2 degree of freedom joint test of main and interaction effects. These ethnic-specific results were then combined through meta-analysis using inverse-variance weighting.**Results:** Trans-ethnic analysis identified 88 genome-wide significant loci ($p<5\bullet10^{-8}$). On chromosome 12, an InDel variant near *KERA* showed the strongest association (overall $p=5.3\bullet10^{-27}$; EA $p=1.7\bullet10^{-19}$; AA p=0.015; Asian $p=4.2\bullet10^{-7}$; HA p=0.005; overall MAF=17%). On chromosome 1, intronic SNPs clustered in *CLCN6* showed the next strongest association (overall $p=1.2\bullet10^{-14}$; EA $p=2.3\bullet10^{-12}$; AA p=0.11; Asian p=0.03; HA p=0.009; overall MAF=14%). Most identified low frequency variants (1% <MAF<5%) showed associations that were significant only for one ethnic group: 6 loci from EA data, 43 loci from AA data, 2 loci from Asian data, and 8 loci from HA data. **Conclusion:** Our gene-smoking interaction analyses validated several known loci and identified multiple novel loci. It is not clear whether our findings were driven by accounting for gene-smoking interaction effects on BP or by allowing for the influence of smoking on BP traits. Considerable levels of heterogeneity were found across multi-ethnic results in both main and interaction effects. Systematic examination of these effects across ethic groups is on the way. Our findings highlight advantages and challenges with multi-ethnic studies.

313

Genome-wide study for blood metabolites identifies 62 loci and connects *LPA* with lipoprotein metabolism from a new perspective.

J. Kettunen[1,2,3] on behalf of the MAGNETIC consortium. 1) Computational Medicine, Institute of Health Sciences, University of Oulu, Oulu, Finland; 2) National Institute for Health and Welfare, Helsinki, Finland; 3) NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland.

Metabolic phenotypes are highly heritable and have great potential in providing insight into genetic variation influencing both metabolism and complex diseases. We present the largest evaluation of genetic variance in human metabolism so far. We combined 123 metabolic phenotypes from blood samples of 24,925 individuals from fourteen European cohorts and associated 62 loci with blood metabolite concentrations. 8 of the loci were new. For 15 loci the lead SNP was low frequency, for 8 a coding variant and 22 involved transcription factor binding sites. We showed that two loci, known for Mendelian amino acid metabolism disease with severe neurological manifestations, also harbor low-frequency variants affecting the same circulating amino acid in healthy population samples. Moreover, we show new evidence that the biosynthesis of lipoprotein(a) (Lp(a)), a known coronary heart disease (CHD) biomarker, is associated with very-low-density lipoprotein metabolism and other lipoprotein particles. Causality of metabolite associations was shown with a genetic risk score ($GRS_{Lp(a)}$) for Lp(a), which composed of 18 SNPs independently associated with circulating Lp(a) levels at genome-wide significance in a discovery cohort. $GRS_{Lp(a)}$ SNPs were confirmed in a replication cohort, where $GRS_{Lp(a)}$ explained 45% of Lp(a) variance. Linking $GRS_{Lp(a)}$ to electronic health records of over 17 000 population-based persons showed that the risk score was associated with CHD outcomes but not with any other disease-category. (ICD10:I20-I25 category, $P=6.4\times10^{-10}$, $N_{events}=1251$, OR=1.28 per one unit increment in $GRS_{Lp(a)}$)). We present a new hypothesis for the biosynthesis of Lp(a) and our results together with previous findings reinforce the observation that *LPA*-targeting treatment has great potential for CHD risk reduction in humans.

314

Characterisation of the metabolic impact of rare genetic variation within *APOC3*: Proton NMR based analysis of rare variant gene effects.

N.J. Timpson[1], F. Drenos[1], J. Kettunen[2,3,6], P. Wurtz[2], P. Soininen[2,3], A.J. Kangas[2], A. Hingorani[8], T. Gaunt[1], J.P. Casas[7], M. Ala-Korpela[2,3,1,4,5], G. Davey Smith[1]. 1) MRC IEU, University of Bristol, Bristol, Bristol, United Kingdom; 2) Computational Medicine, Institute of Health Sciences, University of Oulu, Oulu, Finland; 3) NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland; 4) Computational Medicine, School of Social and Community Medicine, University of Bristol, Bristol, UK; 5) Computational Medicine, Oulu University Hospital, Oulu, Finland; 6) Public Health Genomics Unit, Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland; 7) London School of Hygiene and Tropical Medicine; 8) Genetic Epidemiology Group, University College London, UK.

Plasma triglyceride levels (TG) have been implicated in atherosclerosis and coronary heart disease (CHD). APOC3 plays a key role in the hydrolysis of TG rich lipoproteins to remnant particles by lipoprotein lipase (LPL) and uptake by the liver. We previously reported rare variant in *APOC3* (rs138326449(splice donor)) associated with plasma TG levels (-1.43SD(s.e.=0.27),$p=8.0\times10^{-8}$) discovered initially using low read-depth whole genome sequence. This represents one of the first examples of a rare, large effect variant identified from sequencing at a population scale and here we aimed to characterise in greater detail the impact of variation at this locus by assessing the association profile of this variant across a broad metabolomic data set. A high-throughput serum nuclear magnetic resonance (Proton NMR) metabolomics platform was used to quantify =>225 metabolic measures in >10000 participants from the Avon Longitudinal Study of Parents and children and >3500 from the British Women's Heart and Health Study. Using genotyping information of 12,812 individuals we analysed the effect of the *APOC3* variant on the measured metabolites and used the common *LPL* rs12678919 polymorphism to test for LPL independent effects. In testing all 225 metabolites measured in all samples for association with rs138326449, 142 showed evidence of association ($p<0.05$). rs138326449 was associated with TG ($p = 6.7\times10^{-6}$) and HDL ($p = 1.1\times10^{-11}$). We also identified additional associations with VLDL and HDL composition, other total cholesterol measures and fatty acids. The greater resolution provided by the detailed measurement of the lipoprotein sub-classes showed that the effects of rs138326449 on VLDL and HDL are seen in almost the entire spectrum of their particle size and are not specific. Our comparison of the *APOC3* and *LPL* association revealed that of the 225 metabolites tested, 3 had no overlapping effects between rs138326449_*APOC3* effects and those predicted by rs12678919_*LPL*. Specifically, the composition of medium and very large VLDL is not predicted by the action of APOC3 through LPL where the ratio of TG in the particles is much lower than expected. We have characterised the effects of the newly discovered *APOC3* rs138326449 loss of function mutation in lipoprotein metabolism. Results are consistent with recent clinical trials of APOC3 inhibition and should be used to inform future Mendelian randomisation and recall by genotype studies.

315

Analysis of more than 800,000 genotypes from individuals born in the United States reveals trends in increasing genetic diversity during the 20th century. *A.R. Kermany, M. Barber, J. Byrnes, P. Carbonetto, R. Curtis, J. Granka, E. Han, N. Myres, K. Noto, Y. Wang, C. Ball, K. Chahine.* AncestryDNA, San Francisco, CA.

In recent years there has been much attention to changing demographic landscape of the United States. Understanding temporal trends in genetic diversity, in addition to being of sociological interest, also has important ramifications in our approach towards public health and clinical studies. To our knowledge, there has been no study to understand recent changes in human genetic diversity. Here, we analyze a massive cohort of more than 800,000 genotypes from AncestryDNA customers born in the U.S. between 1920 and 2010.

The AncestryDNA database is composed of more than 850K individuals, genotyped over ~700,000 SNPs using Illumina OmniExpress array. All individuals are annotated with birth locations and birth dates. For each individual, admixture proportions are calculated using ADMIXTURE (based on a reference panel of 26 regions). We used the GERMLINE algorithm to identify DNA matches. We divided customers based on their year of birth from 1920 to 2010 and quantified genetic diversity using different criteria (*e.g.,* entropy in admixture proportions) within each age group. In addition, we conducted a separate study of duo data (parent-child relationships identified based on DNA matching) and calculated parent-child ethnicity divergence for different parental age groups.

Our analysis unfolds a detailed picture of increasing genetic diversity in the U.S. and also implies different rates of change over different time periods.

316

Reconstructing the population history of New York City. *G.M. Belbin[1,5], D. Ruderfer[2,3,4], E.A. Stahl[2,3,5,6], J. Jeff[5], Y. Lu[5], R.J.F. Loos[5,7], E.P. Bottinger[1], N.S. Abul-Husn[1,5], A. Auton[8], E.E. Kenny[1,3,5,6].* 1) Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York , NY; 2) Broad Institute, Cambridge, MA; 3) Division of Psychiatric Genomics, Icahn School of Medicine at Mt Sinai, New York, NY; 4) Center for Statistical Genetics, Icahn School of Medicine at Mt Sinai, New York, NY; 5) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mt Sinai, New York, NY; 6) Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mt Sinai, New York, NY; 7) The Mindich Child Health and Development Institute, Icahn School of Medicine at Mt Sinai, New York, NY; 8) Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, New York, NY.

New York City (NYC) has historically been a significant point of entry for immigration into the United States for the past 500 years and as a consequence has become a highly structured and ethnically diverse population. Census ethnic labeling reveals some of this diversity, but does not fully capture the variety of cultural groups, with complex and diverse demographic origins, foods and traditions, living in New York. Using genome-wide data, it is possible to detect such population structure. We have combined genetic data along with detailed ancestry information and ZIP code information derived from an Electronic Health Record (EHR) database for a population of over 31,000 patients enrolled in the Icahn School of Medicine Bio*Me* Biobank Cohort (Bio*Me*). The combined analysis of this data has allowed us to capture complex patterns of migration and population-structure at the ultra-fine resolution of the NYC neighbourhood. Specifically, we present an array of approaches for the analysis of fine-scale population structure across ~13,500 genotyped on Illumnia Omni Express and Exome Chip data (~900K SNPs) and an additional ~10,000 individuals on the Illumina MEGA array (~1.8M SNPs) in conjunction with self-reported ancestry information that includes individual country of birth as well as both parental and grandparental country of origin. We combined our genetic data with data generated from the 1000 genomes project and an additional unique database of genomic variation in over 400 populations representing diversity from Europe, Middle East, Native American and Oceanian groups, Africa, East and South Asia. Proportion continental and sub-continental genetic ancestry was calculated using standard Principal Component Analysis (PCA) and ADMIXTURE. In addition, patterns of diaspora, migration and endogamy in the past 10-15 generations are revealed via analysis of patterns of local ancestry and identity-by-descent (IBD) haplotypes among and between our data and reference populations. Combining this with time-stamped ZIP code information from the EHR, we are able to create a temporal map of population structure and migration within NYC, revealing ultra fine-scale patterns of demography. Recognition of fine-scale population structure such as highlighted here will become increasingly important for the genetic-analysis of diverse, urban populations.

317

Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *F. Gao, A. Keinan.* Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY.

Several recent studies utilized the site frequency spectrum to show that human populations have undergone a recent epoch of fast growth in effective population size. Inference in these studies has been based on the assumption that the rate of population growth is exponential. However, the observation of an unexplained excess amount of extremely rare, previously unknown variants in several large sequencing studies suggests that human populations might have experienced a recent growth with speed faster than exponential. Recent studies introduced a generalized growth model in which the growth speed can be faster or slower than exponential. We aim to accurately describe human genetic history by using a richer family of models that allow for multiple epochs of such generalized growth. However, only simulation approaches are currently available for obtaining the site frequency spectrum of generalized growth models, which renders inference for human populations computationally intractable. In this study, we derived numerically stable formulae of the expected time to coalescent under multi-epoch generalized growth models within a coalescent theory framework. This facilitates efficient evaluation of the site frequency spectrum and other summary statistics. We further implemented the formulae in a publicly available software, EGGS (http://keinanlab.cb.bscb.cornell.edu), that calculates the summary statistics with high accuracy and orders of magnitude more efficiently compared to simulation approaches. We tested the software on simulated datasets and investigated the power to infer deviation from exponential growth in models that approximate the history of human populations. We observed that decent sample sizes facilitate accurate inference, e.g. a sample size of 3,000 individuals allows observing a growth with speed that deviates by 10% or more from that of exponential. Applying our inference framework to Europeans from the Exome Sequencing Project with focus on synonymous variants we showed that recent epoch of generalized growth provides a better fit to the observed site frequency spectrum than exponential growth ($P = 7.0\times10^{-4}$), with the former estimating growth speed modestly (11%) faster than exponential ($P = 3.7\times10^{-51}$), consistent with historical records. With increasing amount of high-quality, whole-genome sequencing data of large sample sizes, the application of generalized models holds the promise of refining our understanding of human demographic history.

318

Genome-wide data on 34 ancient Anatolians identifies the founding population of the European Neolithic. *I. Lazaridis[1,2], D. Fernandes[3], N. Rohland[1,2], S. Mallick[1,2,4], K. Stewardson[1,4], S. Alpaslan[5], N. Patterson[2], R. Pinhasi*[3], D. Reich*[1,2,4].* 1) Department of Genetics, Harvard Medical School, Boston, MA USA; 2) Broad Institute of MIT and Harvard, Cambridge, MA USA; 3) Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland; 4) Howard Hughes Medical Institute, Harvard Medical School, Boston, MA USA; 5) Independent physical anthropologist, Netherlands.

It has hitherto been difficult to obtain genome-wide data from the Near East. By targeting the inner ear region of the petrous bone for extraction [Pinhasi et al., PLoS One 2015] and using a genome-wide capture technology [Haak et al., Nature, 2015] we achieved unprecedented success in obtaining genome-wide data on more than 1.2 million single nucleotide polymorphism targets from 34 Neolithic individuals from Northwestern Anatolia (~6,300 years BCE), including 18 at greater than 1× coverage. Our analysis reveals a homogeneous population that is genetically a plausible source for the first farmers of Europe in the sense of (i) having a high frequency of Y-chromosome haplogroup G2a, and (ii) low Fst distances from early farmers of Germany (0.004 ± 0.0004) and Spain (0.014 ± 0.0009). Model-free principal components and model-based admixture analyses confirm a strong genetic relationship between Anatolian and European farmers. We model early European farmers as mixtures of Neolithic Anatolians and Mesolithic European hunter-gatherers, revealing very limited admixture with indigenous hunter-gatherers during the initial spread of Neolithic farmers into Europe. Our results therefore provide an overwhelming support to the migration of Near Eastern/Anatolian farmers into southeast and Central Europe around 7,000-6,500 BCE [Ammerman & Cavalli Sforza, 1984, Pinhasi et al., PLoS Biology, 2005]. Our results also show differences between early Anatolians and all present-day populations from the Near East, Anatolia, and Caucasus, showing that the early Anatolian farmers, just as their European relatives, were later demographically replaced to a substantial degree.

319

The evolutionary impact of Denisovan ancestry in Australo-Melanesians. *S. Sankararaman[1,2], S. Mallick[1,2,3], N. Patterson[2], D. Reich[1,2,3] for The Simons Genome Diversity Project.* 1) Department of Genetics, Harvard Medical School, Boston, MA USA; 2) Broad Institute of Harvard and MIT, Cambridge, MA USA; 3) Howard Hughes Medical Institute, Harvard Medical School, Boston, MA USA.

Analyses of genome sequences from archaic and modern humans have documented major admixture events between the ancestors of Neanderthals and non-Africans as well as between the Denisovans (a sister-group of the Neanderthals) and populations in island south-east Asia. Understanding the impact of these ancient admixture events on evolution and phenotypes is a central goal in human population genomics. While a number of recent studies have made progress towards understanding the structure and impact of Neanderthal admixture [Sankaramanan et al. Nature 2014; Vernot and Akey Science 2014], the Denisovan admixture event remains poorly understood. To this end, we adapted a statistical method previously developed for inferring Neanderthal ancestry to infer Neanderthal and Denisovan local ancestries in Melanesian populations. We applied this method to a dataset of high-coverage whole-genome sequences from 11 Melanesian individuals (2 Aboriginal Australians, 1 Bougainville Islander, 8 Papua New Guineans) that were sequenced as part of the Simons Genome Diversity Project to infer maps of Denisovan and Neanderthal ancestry in these populations. Power to confidently infer Denisovan ancestry is estimated to be about half that of Neanderthal ancestry – a consequence of the greater divergence of the sequenced Denisovan genome from the ancestral population. Nevertheless, our statistical method identifies around 38,000 Neanderthal-derived alleles and around 25,000 Denisovan-derived alleles. Using the confidently inferred ancestries across multiple individuals, we can reconstruct about 150 Mb of the genome of the introgressing Denisovan. We observe that the proportion of both Denisovan and Neanderthal local ancestry is reduced in regions of the genome with strong background selection. This observation is consistent with a model in which Neanderthal and Denisovan alleles are subject to strong purifying selection in the admixed Melanesian populations analogous to the previous observation of strong purifying selection against Neanderthal alleles in non-Africans. In addition, we document a number of regions with elevated proportions of archaic ancestry (including a previously reported example at the STAT2 locus) which represent putative candidates for adaptive introgression.

320

IBD sharing in the 1000 Genomes Project Phase 3 data reveals relationships from Neandertals to present day families. *G. Povysil, S. Hochreiter.* Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria.

The 1000 Genomes Project data harbor information about a great variety of relationships which can be recovered using identity by descent (IBD) analysis. Short IBD segments convey information about events far back in time because the shorter IBD segments are, the older they are assumed to be. At the same time longer IBD segments can be used to detect more recent relationships as they occur in families. The identification of short IBD segments becomes possible through next generation sequencing (NGS), which offers high variant density and reports variants of all frequencies. However, only recently HapFABIA has been proposed as the first method for detecting very short IBD segments in NGS data. HapFABIA utilizes rare variants to identify IBD segments with a low false discovery rate. We applied HapFABIA to the 1000 Genomes Phase 3 whole genome sequencing data to identify IBD segments which are shared within and between populations as well as with the genomes of Neandertal and Denisova. Using the proportion of IBD segments an individual shares with any other individual in the data set, we were able to discover first degree relatives that we consequently removed from further analyses. Not only are most IBD segments found in Africans, but also each African individual has about ten times more IBD segments than any East Asian, South Asian, or European individual. Furthermore, the number of IBD segments of an individual correlates with his degree of African ancestry as reported by other methods. IBD segments can be used to recover the population of origin of an individual and find individuals with wrong population labels. By comparing the rare variants that tag an IBD segment with the genome of Neandertal and Denisova, we were able to find IBD segments shared with these ancient genomes. We extracted two types of very old IBD segments that are shared with Neandertals/Denisovans: (1) longer segments primarily found in East Asians, South Asians, and Europeans that indicate introgression events outside of Africa; (2) shorter segments mainly shared by Africans that may indicate events involving ancestors of humans and other ancient hominins within Africa. Our results from the autosomes are further supported by an analysis of chromosome X, on which segments that are shared by Africans and match the Neandertal and/or Denisova genome were even more prominent.

321

Computational reconstruction of a haploid genome from the 18th century. *A. Jagadeesan[1,2], E.D. Gunnarsdóttir[1], S.S. Ebenesersdóttir[1,2], V.B. Guðmundsdóttir[1,2], E.L. Þórðardóttir[1,2], M.S. Einarsdóttir[1,2], J. Dugoujon[9], C. Fortes-Lima[9], F. Migot-Nabias[10], A. Massougbodji[10], G. Bellis[11], P. Triska[4,5,6], V. Cerny[7], L. Pereira[4,5,8], A. Kong[1], A. Helgason[1,2], K. Stefánsson[1,3].* 1) deCODE genetics, Sturlugata 8, Reykjavik 101, Iceland; 2) Department of Anthropology,University of Iceland, Sæmundargötu 2 101 Reykjavík, Iceland; 3) Faculty of Medicine, University of Iceland, Sæmundargötu 2 101 Reykjavík, Iceland; 4) Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto 4200-135, Portugal; 5) Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto 4200-465, Portugal; 6) Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto (ICBAS), Porto 4050-313, Portugal; 7) Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague, Czech Republic; 8) Faculdade de Medicina da Universidade do Porto, Porto 4200-319, Portugal; 9) Evolutionary Medicine Group, Laboratoire 'Anthropologie Moléculaire et Imagerie de Synthèse, UMR 5288, Centre National de la Recherche Scientifique (CNRS), Université Toulouse 3-Paul-Sabatier, Toulouse, France; 10) Centre d'Etude et de Recherche sur le Paludisme Associé à la Grossesse et à l'Enfance, Faculté des Sciences de la Santé, Université d'Abomey-Calavi, Cotonou, Benin; 11) Institut National d'Etudes Démographiques, Paris, France.

Icelandic history contains an unusually tractable case study for the problem of ancestral genome reconstruction. The case is that of Hans Jonatan (HJ), born in 1784 on the Caribbean island St. Croix to an African enslaved mother and a supposed Danish father. Following a complex series of events, HJ eventually settled in Iceland in 1805 and married an Icelandic woman, with whom he had two children. In all, HJ has 636 contemporary descendants in the Icelandic population, of whom we have Illumina microarray SNP genotypes for 114. Using these data, we identified African fragments using two different methods. The first, "AdmixIBD", was devised by us and is based on detection of identity-by-descent (IBD) sharing of long contiguous chromosome fragments with a set of reference samples from Europe and Africa. The second is a published method called Hapmix. The African fragments identified by both methods were highly consistent and were therefore combined and then further filtered based on correspondence between the parental origin of alleles in HJ descendants and the known genealogy linking them to HJ. In this way, we were able to identify and reconstruct 32% of HJ's maternal haploid African genome. We then attempted to determine the geographical origin of this reconstructed genome. The transatlantic slave trade resulted in mass uprootment of millions of Africans to the New World. Using reference samples from 48 different African populations, we were able to use multi-dimensional scaling analysis and a statistical test of IBD values to rule out East and Southern Africa as places of origin for HJ's mother. Within West Africa, we found statistical support for a closer relationship to populations from Nigeria and Benin than to other populations.

322

***Polly*: A novel approach for estimating local and global admixture proportion based on rich haplotype models.** *K. Noto, Y. Wang, M. Barber, J. Byrnes, P. Carbonetto, R.E. Curtis, E. Han, A. Kermany, N. Myres, C.A. Ball, K. Chahine.* AncestryDNA, 153 Townsend St. Suite 800. San Francisco, CA. 94107.

We present Polly, a novel method for local and global admixture estimation based on the *Beagle* (Browning and Browning, 2007) haplotype models that are learned from hundreds of thousands of haplotypes and that we have previously shown to work well for phasing (Noto *et al.,* ASHG 2014). These haplotype models each cover a small section (about one 1,000th) of the genome. Polly works by annotating the haplotype clusters that make up the models according to their agreement with a large labeled reference panel of single-origin individuals. Given a new unlabeled genotype, in a process similar to the one used for phasing, we consider the model's haplotype clusters that may explain each of the genotype's two constituent haplotypes, and measure the likelihood of the population assignment to each of them based on the annotations of those clusters. The result of these measurements is a likelihood distribution over population assignments for each haplotype, for each section of the genome. Polly then finds the most parsimonious population assignments to each of these sections, taking into account both the individual likelihood distributions and the fact that changes to population assignments require recent recombination and should be relatively rare at the fine scale of these small sections of the genotype. Over 20-fold cross-validation on a set of 8,507 single-origin individuals from 26 distinct regions all over the world, we observe that the agreement between the Polly-estimated global assignments and the correct assignment averages 11% higher per individual than it does for the estimates based on the semi-supervised approach ADMIXTURE (Alexander *et al.,* 2009) using the same train/test data. When we use the test data to simulate individuals with precisely-known genetic admixture proportions, Polly's agreement with the correct global assignment averages 15% to 61% higher than does ADMIXTURE's, depending on the amount of admixture in the test set individual. Additionally, in these experiments Polly runs in about one-fifth the time of ADMIXTURE with the same compute resources, and is easily parallelizable.

323

An integrative model for predicting the regulatory impact of rare non-coding variants on the human transcriptome. *Y. Kim[1], X. Lee[2], F. Damani[1], Z. Zappala[3], J. Davis[3], E. Tsang[3], S.B. Montgomery[2,3], A.J. Battle[1], The GTEx Consortium.* 1) Department of Computer Science, Johns Hopkins University, Baltimore, MD, United States; 2) Department of Pathology, Stanford University, Stanford, CA, United States; 3) Department of Genetics, Stanford University, Stanford, CA, United States.

The enormous increase in availability of full human genomic sequences presents great opportunity for understanding the impact of rare genetic variants in human disease. Based on current knowledge, however, we are still limited in our ability to interpret or predict consequences of rare variants in the non-coding regions of the genome. Understanding the effects of private regulatory genetic variants on gene expression would allow us to take an essential step toward personalized medicine. Diverse genomic annotations such as epigenetic data and have been shown to be informative for prioritizing deleterious genetic variants and have demonstrated enrichment among common expression quantitative trait loci, and are now beginning to be evaluated in the context of rare variation. The availability of RNA-seq and other molecular phenotypes for the same individuals with genome sequencing offers a new avenue for integrated methods for prioritizing rare functional variants. By considering informative genomic annotations along with molecular phenotyping, we are able to identify the regulatory impact of individual-specific genetic variants largely excluded from previous analyses. Specifically, we have developed a Bayesian statistical learning approach that integrates whole genome sequencing with RNA-seq data from the same individual, leveraging gene expression levels and allele-specific expression along with diverse genomic annotations and performing joint inference to identify likely rare regulatory variants. We have applied this model to 157 individual whole genome sequences and corresponding RNA-sequence in 54 different tissues samples from the GTEx project and report the most likely functional rare regulatory variants for each individual. We show that specific genomic features including location, conservation scores, transcription factor binding affinities, and epigenetic data associated with rare and private variants are predictive of impact on expression of nearby genes. Further, we demonstrate that correctly controlling for hidden confounders can greatly increase the power to identify functional rare variants. Our probabilistic model of the regulatory impact on human tissues from rare genetic variants offers great potential for differentiating causal variants from neutral mutations, assists better understanding of functionality of non-coding regions of the genome, and helps pave the way to ultimately predict phenotypic consequences of rare regulatory variation.

324

Genetic and epigenetic factors affecting regulatory elements underlie lactose intolerance and lactase persistence. *R. Jeremian[1], V. Labrie[1,2], O. Buske[3], E. Oh[1], C. Ptak[1], G. Gasiunas[4], A. Maleckas[5], R. Petereit[6], A. Zvirbliene[6,7], K. Adamonis[6], E. Kriukienė[8], K. Koncevicius[9], J. Gordevičius[9], A. Nair[1], A. Zhang[1], S. Ebrahimi[1], G. Oh[1], V. Siksnys[4], L. Kupcinskas[6,7], M. Brudno[3], A. Petronis[1].* 1) Krembil Family Epigenetics Laboratory, Centre for Addiction and Mental Health, Toronto, ON, Canada; 2) Department of Psychiatry, University of Toronto, Toronto, ON, Canada; 3) Department of Computer Science, University of Toronto, Toronto, ON, Canada; 4) Department of Protein-DNA Interactions, Institute of Biotechnology, Vilnius University, Vilnius, Lithuania; 5) Department of Surgery, Lithuanian University of Health Sciences, Kaunas, Lithuania; 6) Department of Gastroenterology, Lithuanian University of Health Sciences, Kaunas, Lithuania; 7) Institute for Digestive Research, Lithuanian University of Health Sciences, Kaunas, Lithuania; 8) Department of Biological DNA Modification, Institute of Biotechnology, Vilnius University, Vilnius, Lithuania; 9) Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania.

Lactose intolerance is mediated by the lactase (*LCT*) gene and is a widespread heritable trait affecting humans and other mammals, apart from certain human populations that exhibit lactase persistence. Despite progress in mapping DNA sequence-based factors associated with this trait, it remains unclear what accounts for the age-dependent downregulation in *LCT* transcription, which results in lactose intolerance in adulthood. Why is this gene very transcriptionally active in the intestine of newborns but later declines dramatically in most (~65%), but not all, adults worldwide? To address this question, we investigated epigenetic regulation of *LCT* in the human intestine using chromosome–wide DNA modification profiling and targeted bisulfite sequencing. We identified 7 sites exhibiting major DNA modification differences that direct *LCT* mRNA levels in lactose intolerant and lactase persistent adults. These sites account for the age-dependent, cell-type specific, and inter-individual differences in *LCT* mRNA. The discovered regions contained chromatin signatures of enhancers, and mapped to introns of *LCT*, the upstream *MCM6* gene and to the promoter of a previously uncharacterized long non-coding RNA. Similar analysis in cohorts of newborn and adult mice corroborated the human findings, revealing the evolutionary conservation of specific lactase regulatory elements between humans and mice. Next, we investigated the causal roles of the discovered enhancers by deleting them via CRISPR-Cas9 mutagenesis. Induced deletions of the non-coding *LCT* regulatory regions resulted in significant loss in lactase expression in mice and a human intestinal cell line. Furthermore, downregulation of the *LCT*-associated long non-coding RNA by RNA interference resulted in a concomitant reduction of *LCT* mRNA. Integration of genetic analyses with our epigenetic investigations revealed that different DNA haplotypes show differential epigenetic aging patterns, signifying that DNA haplotypes can affect this age-dependent trait by influencing the epigenetic control of gene regulatory elements. In conclusion, lactose intolerance is produced by the synergistic effects of the inherited haplotype and acquired age-dependent epigenetic modifications. The lessons learned from this study can be directly applied to etiopathogenic studies of common, complex human diseases, such as cancer, diabetes and Alzheimer's disease.

325

Characterizing *de novo* balanced cytogenetic abnormalities through sequencing in 147 subjects with multiple congenital anomalies. *C. Redin[1,2,3], H. Brand[1,2,3], R.L. Collins[1], V. Pillalamarri[1], C. Hanscom[1], T. Kammin[4], S. Pereira[4], B.B. Currall[4], Z. Ordulu[4], S. Althari[5], J. Shen[5], A. Ragavendran[1], E.C. Liao[6,7,8], E. Mitchell[9], J.C. Hodge[9,10], C.C. Morton[3,4,5,11], J.F. Gusella[1,3,12], M.E. Talkowski[1,2,3].* 1) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 2) Department of Neurology, Massachusetts General Hospital, Boston, MA; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge; 4) Department of Obstetrics, Gynecology, and Reproductive Biology, Brigham and Women's Hospital, Boston, MA; 5) Department of Pathology, Brigham and Women's Hospital, Boston, MA; 6) Center for Regenerative Medicine, Massachusetts General Hospital, Harvard Medical School, Boston; 7) Division of Plastic and Reconstructive Surgery, MGH, Boston, MA; 8) Harvard Stem Cell Institute, Boston, MA; 9) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; 10) Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, California; 11) School of Psychological Sciences, University of Manchester, England; 12) Department of Genetics, Harvard Medical School, Boston, MA.

De novo balanced chromosomal abnormalities (BCAs) represent an important component of the genetic etiology of human congenital anomalies, yet they are invisible to first-tier screening tools such as microarray or exome sequencing. We characterized at sequence resolution a large cohort of *de novo* BCAs initially detected by karyotyping in 147 subjects, 111 (76%) of which presented with neurodevelopmental abnormalities. We delineated 414 breakpoints from these 147 BCAs, revealing 117 simple, balanced exchanges and 30 that sequencing showed to be complex. Consistent with our previous findings, 80% of complex BCAs involved at least one inverted segment, and karyotyping always misassigned one or more breakpoints by at least one sub-band. Comparing these breakpoints to a random distribution of 10,000 sets of simulated breakpoints throughout the genome revealed that cytogenetically detectable *de novo* BCAs are: 1) significantly more likely to occur near telomeres rather than centromeres, 2) overlap with known recombination hotspots, and 3) occur more often between chromosomal regions of concordant chromatin states as indicated when overlaying Hi-C data. We also found that these *de novo* BCAs are significantly enriched for genes associated with autism, evolutionary constrained genes, and those containing RFBOX RNA-binding sites. The integration of expression profiles from Brainspan and GTEx revealed that these BCA breakpoints were enriched for genes expressed early during brain development, but not those expressed later in infancy or in tissues other than brain. Among 147 BCAs studied, sequencing revealed the disruption of 33 previously defined pathogenic loci (e.g., *AUTS2, CHD8, MBD5, MTAP, FGFR1*) and 24 novel or recently described risk loci for congenital anomalies (e.g., *CUL3, CTNND2, MYT1L, NFIA, PSD3*). This collectively represents a yield of 39% of *de novo* BCAs likely contributing to the phenotype in the proband, while 51 additional subjects harbored single gene disruptions of unknown consequence that warrant further study. Notably, 9 subjects with similar phenotypes harbored BCAs that did not disrupt genes but all clustered within the cytogenetic band 5q14.3. These data suggest that *de novo* BCAs represent a relatively uncharacterized source of variations with high functional impact, and that the amalgamation of large datasets may improve risk prediction for anomalies that occur early in development.

326

Adipose- and maternal- specific regulatory variants at *KLF14* influence Type 2 Diabetes risk in women via a female-specific effect on adipocyte physiology and body composition. *K. Small[1], M. Todorcevic[2], M. Civelek[3], J. El-Sayed Moustafa[1], A. Mahajan[4], M. Horikoshi[4], A. Hough[5], C. Glastonbury[1], G. Thorleifsson[6], L. Quaye[1], J. Fernandez[4], A. Buil[7], A. Vinuela[1], M. Yon[5], M. Simon[5], S. Sethi[5], J. Bell[1], B. Sharifi[8], U, Thorsteinsdottir[6,9], A.L. Gloyn[2], R. Cox[5], A. Lusis[3], F. Karpe[2], M. McCarthy[2,4].* 1) Department of Twin Research and Genetic Epidemiology, King's College London, London, UK; 2) Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford, UK; 3) School of Medicine, University of California, Los Angeles, CA; 4) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK,; 5) MRC Harwell, Oxford, UK; 6) deCODE genetics/Amgen, Reykjavik, Iceland; 7) Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; 8) Division of Cardiology, Cedars-Sinai Heart Institute, Los Angeles, CA; 9) Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland.

Variants upstream of the maternally-expressed transcription factor *KLF14* are associated with risk of Type 2 Diabetes (T2D). We utilized an integrated *in silico, in vitro* and *in vivo* strategy to demonstrate the molecular, cellular, and whole-body effects of the *KLF14* T2D variant. In population-level RNAseq (N= ~800) and epigenetic (Illumina 450K, N= 640) data from multiple tissues from the TwinsUK cohort the T2D risk allele is associated with an adipose-specific decrease in *KLF14* expression and concurrent increase in methylation at a probe mapping ~ 3KB upstream of *KLF14*. This probe maps to an adipose-specific stretch enhancer and we demonstrate the five-SNP risk haplotype covering the enhancer is sufficient to modulate expression of *KLF14 in vitro.* The *KLF14 cis*-eQTL has a widespread effect on the adipose transcriptome, regulating the expression of 385 genes in *trans* (FDR 5%). The *trans*-genes are enriched for KLF binding sites, indicating they are directly regulated by *KLF14*. The *cis* and *trans*-effects are robust and adipose specific, replicating in three adipose cohorts but not present in any other tissue. Consistent with the known imprinted status of *KLF14*, the T2D, *cis* and *trans* effects are limited to the maternal allele. We show that *KLF14* regulatory changes effect adipocyte cellular physiology and manifest at a whole-body level in a reduction in hip circumference and concurrent increase in T2D disease susceptibility and lipid levels. Mouse knockout models further support a causal role for *KLF14* expression in metabolic phenotypes. While the molecular *cis* and *trans* regulatory effects are present in both sexes, the cellular and whole body effects (T2D, Lipids, Hip Circumference) are female-specific. The disconnect between female-specific trait associations at the whole-body level and the presence of the *cis* & *trans* regulatory effects in both sexes may be due to higher baseline expression levels of *KLF14* observed in females or to balancing effects of male-specific body composition GWAS signals (VAT/SAT, Hip Circumference) located at several *trans*-genes, or both. While the *KLF14* T2D association was discovered in a large GWAS, its remarkable specificity- parent-of-origin-specific, female-specific, European-specific, tissue-specific- implies that there is value in further GWAS efforts in sub-populations both for discovery and for refinement of risk prediction models.

327

Single variant resolution association mapping of inflammatory bowel disease loci. *H. Huang[1,2], L. Jostins[3], M. Fang[4], M.U. Mirkov[5], M. Georges[4], J. Barrett[5], M.J. Daly[1,2], the International IBD Genetics Consortium.* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Broad Institute of MIT and Harvard, Cambridge MA; 3) Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK; 4) Unit of Animal Genomics, Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA-R) and Faculty of Veterinary Medicine, University of Liège (B34), Liège, Belgium; 5) Wellcome Trust Sanger Institute, Hinxton, UK.

Inflammatory bowel disease (IBD), a chronic gastrointestinal (GI) inflammatory disorder, affects millions worldwide. Genomewide association studies (GWAS) have identified nearly 200 IBD-associated loci, but only a handful have been conclusively resolved to specific functional variants, limiting the ability to pursue insightful functional investigation. Here we report a fine-mapping study to comprehensively define the number of independent association signals in each IBD locus, the disease form (Crohn's or ulcerative colitis) of each signal and a minimal set of variants that contains the causal variants.We genotyped 66,849 European-derived samples using a high-density genotyping array (Immunochip). After stringent quality control to remove genotyping errors and batch effects, the high-density data was imputed into the 1000 Genomes reference panel. By applying and integrating three Bayesian fine-mapping algorithms, we fine-mapped 139 independent associations in 97 IBD loci, accounting for about 90% of the total variance explained by these loci. We mapped 116 associations to ≤ 50 variants. Among them, 18 associations are mapped to a single causal variant with >95% certainty, including 4 coding in NOD2, 1 coding in IL23R and a splice variant in CARD9, all of which have been previously reported and thereby provide confidence in the statistical procedures. The remaining 12 causal variants are novel and include variants implicating type 1 diabetes, reducing the protein binding affinity and regulating immune cell populations. We found causal variants are enriched with coding variants and variants that disrupt transcript factor binding motifs. Similar enrichment was observed in H3K4me1 marks in immune related cell types and H3K27ac marks in sigmoid colon and rectal mucosal tissues. In contrast, we only observed limited enrichment of causal variants in the cis-eQTLs in immune cell lines and GI tissues, and no enrichment in whole blood. This suggests the number of genetic associations driven by baseline eQTL may be limited, with further studies in appropriate cell types, stimuli and sufficient sample size needed. This also underscores the incompleteness of our knowledge regarding the function of non-coding DNA and its role in disease.Our results suggest that high-resolution, large-sample mapping can convert many GWAS discoveries into statistically convincing causal variants providing a powerful substrate for experimental elucidation of disease mechanisms.

328

Genome sequencing of autism families reveals disruption in non-coding regulatory DNA. *T.N. Turner[1], F. Hormozdiari[1], M. Duyzend[1], I. Iossifov[2], A. Raja[1], C. Baker[1], K. Hoekzema[1], H.A. Stessman[1], M.C. Zody[3], B. Nelson[1], J. Huddleston[1], R. Sandstrom[1], J. Smith[1], D. Hanna[1], M. Bamshad[1], J. Stamatoyannopoulos[1], D.A. Nickerson[1], R. Darnell[3], E.E. Eichler[1,4].* 1) Department of Genome Sciences, University of Washington, Seattle, WA; 2) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; 3) New York Genome Center, New York, NY; 4) Howard Hughes Medical Institute, University of Washington, Seattle, WA.

We performed whole-genome sequencing (WGS) of 208 genomes from 53 simplex autism families, the majority of which had no copy number variant (CNV) or candidate *de novo* gene disruptive single nucleotide variant (SNV) by microarray or whole-exome sequencing (WES). We integrated multiple CNV and SNV analyses with extensive experimental validation to identify additional high-risk mutations in eight families. We report a modest (p=0.03) enrichment of *de novo* and private disruptive mutations within fetal central nervous system (CNS) DNase I hypersensitive sites mapping within 50 kbp of autism genes when comparing probands to controls. The effect is observed near genes where dosage sensitivity has already been established by recurrent disruptive *de novo* protein coding mutations (*ARID1B, SCN2A, TRIO, NR3C2, PRKCA, DSCAM*). Functional testing of these events are currently underway. In addition, we provide evidence of gene-disruptive CNVs (*DISC1*, *WNT7A*, *RBFOX1* and *MBD5*) as well as smaller *de novo* CNVs and SNV exon-specific mutations in neurodevelopmental genes missed by exome sequencing (eg. *CANX, SAE1* and *PIK3CA*). While WGS is superior with respect to uniformity, increased CNV detection and access to GC-rich regions, the greatest sensitivity for single-nucleotide variants (SNVs) was obtained by combining exome and genome sequencing data (93.9% shared, 3.1% exome-only and 3.0% genome-only; n = 176,131) due to increased sequence coverage afforded by WES. Our results suggest that the detection of smaller often multiple CNVs affecting single exons and regulatory elements in different genes will help explain an additional ~15% of unexplained simplex autism risk.

329

Fine mapping of psoriasis susceptibility loci: Enrichment of pro-inflammatory genomic marks in lymphocytes and keratinocytes. *L.C. Tsoi[1], S.L. Spain[2], E. Ellinghaus[3], P.E. Stuart[4], T. Esko[5], T. Pers[5], X. Wen[1], F. Capon[6], J. Knight[7], T. Tejasvi[4], H.M. Kang[1], M.H. Allen[6], S. Weidinger[8], J.E. Gudjonsson[4], S. Koks[9], K. Kingo[10], A. Metspalu[11], G.G. Krueger[12], J.J. Voorhees[4], V. Chandran[13], C.F. Rosen[14], P. Rahman[15], D.D. Gladman[13], A. Reis[16], R.P. Nair[4], A. Franke[3], J.NWN Barker[6], G.R. Abecasis[1], R.C. Trembath[17], J.T. Elder[4].* 1) Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA; 2) Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; 3) Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany; 4) Department of Dermatology, University of Michigan, Ann Arbor, Michigan 48109, USA; 5) Department of Genetics, Harvard Medical School, Boston, MA, USA; 6) Division of Genetics and Molecular Medicine, King's College London, London WC2R 2LS, UK; 7) Neuroscience Research, Centre for Addiction and Mental Health, Toronto, Ontario, Canada M5T 1R8; 8) Department of Dermatology, University Hospital, Schleswig-Holstein, Christian-Albrechts-University, 24105 Kiel, Germany; 9) Department of Pathophysiology, Centre of Translational Medicine and Centre for Translational Genomics, University of Tartu, 50409 Tartu, Estonia; 10) Department of Dermatology and Venereology, University of Tartu, 50409 Tartu, Estonia; 11) Estonian Genome Center, University of Tartu, 51010 Tartu, Estonia; 12) Department of Dermatology, University of Utah, Salt Lake City, Utah 84132, USA; 13) Department of Medicine, Division of Rheumatology, University of Toronto, Toronto Western Hospital, Toronto, Ontario, Canada M5T 2S8; 14) Department of Medicine, Division of Dermatology, University of Toronto, Toronto Western Hospital, Toronto, Ontario, Canada M5T 2S8; 15) Department of Medicine, Memorial University, St John's, Newfoundland, Canada A1C 5B8; 16) Institute of Human Genetics, University of Erlangen-Nuremberg, Erlangen 91054, Germany; 17) Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AD, UK.

Psoriasis is an immune-mediated disorder of skin with complex genetic architecture. Previous studies have identified close to 70 psoriasis susceptibility regions. However, fewer than 25% of them harbor or are in strong linkage disequilibrium with protein-altering variants. Therefore, translating disease-associated genetic signals to biologic effects remains a major challenge. Using 30,000 samples of European ancestry from three GWAS and one Immunochip datasets, we performed a fine-mapping meta-analysis of psoriasis through refined genotyping, imputation, and enrichment analysis focused on genomic regulatory elements. We aimed to identify disease-relevant cell types and the corresponding functional elements that are enriched among independent psoriasis genetic signals. We first performed step-wise conditional analysis, revealing 23 significant ($p<1\times10^{-6}$) independent secondary signals among nine non-MHC psoriasis susceptibility loci. Our results show that these signals could increase the variance explained in psoriasis risk from 12.3% to 17.1%, compared to using only primary signals. We implemented an approach called GEGA (Genomic feature Enrichment analysis for Genetic Association study) to examine the enrichment of multiple genomic elements across eight different ENCODE cell types. We show that chromatin marks in lymphoblastoid cell lines ($p = 1 \times 10^{-4}$; average observed-to-expected ratio (OE) = 1.7) and keratinocytes ($p = 4 \times 10^{-3}$; OE = 1.4) are significantly enriched among the independent signals. We then performed enrichment screening for 87 histone modifications and transcription factor binding sites using ENCODE ChipSeq data for these two cell lines, and identified 22 significant genomic features. Notably, binding sites for IRF4 ($p = 1 \times 10^{-4}$; OE = 3.9) and NFκB ($p = 3 \times 10^{-5}$; OE = 4.3) are enriched among the independent signals, as are histone H3K27ac marks ($p = 3 \times 10^{-5}$; OE = 1.8), which are indicative of enhancer activity. We further show that psoriasis signals overlapping with the enriched features contain a significantly higher proportion of lymphoblastoid cis-eQTLs when compared to other psoriasis signals ($p = 6 \times 10^{-3}$; OE = 1.3). Through conditional analysis on known loci, our results highlight how genomic elements can facilitate fine-mapping analysis for complex diseases.

330

Colocalization of eQTLs at WHRadjBMI GWAS loci with multiple association signals highlighted candidate functional genes for body fat distribution. *Y. Wu[1], M. Civelek[2], C.K. Raulerson[1], A. He[3], C. Tilford[3], C. Fuchsberger[4], A.E. Locke[4], H.M. Stringham[4], A.U. Jackson[4], N.K. Saleem[5], N. Narisu[6], P.S. Chines[6], P. Gargalovic[3], T. Kirchgessner[3], F.S. Collins[6], M. Boehnke[4], M. Laakso[5], A.J. Lusis[2], K.L. Mohlke[1].* 1) Genetics, University of North Carolina, Chapel Hill, Chapel Hill, NC; 2) 2Department of Medicine, University of California, Los Angeles, CA; 3) Bristol-Myers Squibb, Pennington, NJ; 4) Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI; 5) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 6) National Institutes of Health, Bethesda, MD.

The Genetic Investigation of Anthropometric Traits (GIANT) consortium recently reported 49 loci associated with waist-to-hip ratio adjusted for body mass index (WHRadjBMI), including nine loci that contained more than one association signal. The identity and number of target genes at these multi-signal loci are not fully defined. WHRadjBMI loci are enriched for genes expressed in adipose tissue. To identify GWAS signal(s) that are coincident with expression quantitative trait loci (eQTLs), we used a Bayesian test of colocalization to integrate GIANT GWAS results with new subcutaneous adipose eQTL data in 1,381 Finns from the METabolic Syndrome In Men (METSIM) study. To assess whether the multiple GWAS signals correspond to the same or distinct eQTL signals, we performed conditional eQTL analyses. Among the 25 GWAS SNPs at nine multi-signal loci, 17 were associated with expression of ≥1 gene at $p<1.4\times10^{-4}$ (FDR<1%, 21 GWAS-eQTL pairs). The Bayesian test provided evidence of colocalization for five of these GWAS-eQTL pairs (posterior probability of a shared signal >0.8) including *SNX10* and *CBX3* eQTLs at the *SNX10* locus, *GNL3* at the *PBRM1* locus, *NT5DC2* at the *PBRM1* locus and *HOXC4* at the *HOXC* locus. At the *HOXC* gene cluster, which contained three GWAS signals for WHRadjBMI, eQTL mapping for nine local *HOXC* genes revealed significant association of rs2071449 with *HOXC4* ($p=3.3\times10^{-17}$; lead eSNP $p=1.6\times10^{-17}$) and *HOXC8* ($p=2.6\times10^{-20}$; lead eSNP $p=3.4\times10^{-29}$). Consistent with these results, the colocalization test prioritized *HOXC4* as a better candidate gene at this signal with posterior probabilities of 0.95 for *HOXC4* and 0 for *HOXC8*. At the *PBRM1* GWAS locus, the two moderately correlated GWAS SNPs (LD $r^2=0.53$) were both associated with *GNL3* (rs2276824 $p=2\times10^{-10}$; rs12489828 $p=2.5\times10^{-7}$) and *NT5DC2* (rs2276824 $p=3.9\times10^{-10}$; rs12489828 $p=8\times10^{-17}$). Reciprocal conditional eQTL analyses for these two SNPs showed that rs12489828 was still strongly associated with *NT5DC2* expression (rs12489828 $p_{cond}=1.8\times10^{-8}$; rs2276824 $p_{cond}=0.17$), while rs2276824 remained marginally associated with *GNL3* expression (rs2276824 $p_{cond}=0.097$; rs12489828 $p_{cond}=0.78$), suggesting that the two signals at the *PBRM1* GWAS locus might each influence the expression of a different gene. In summary, at multi-signal GWAS loci for which target genes are unclear, the integration of GWAS and eQTL data can highlight plausible candidate genes and help implicate the multiple signals in target gene regulation.

331

Fine-mapping GWAS loci containing extensive allelic heterogeneity reveals complex patterns of association. *C.N. Spracklen[1], A.U. Jackson[2], H.M. Stringham[2], Y. Wu[1], M. Civelek[3], C. Fuchsberger[2], A.E. Locke[2], R. Welch[2], P.S. Chines[4], N. Narisu[4], A.J. Lusis[3], J.K. Kuusisto[5], F.S. Collins[4], M. Boehnke[2], M. Laakso[5], K.L. Mohlke[1].* 1) Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI; 3) Department of Medicine, University of California, Los Angeles, CA; 4) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 5) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.

Genome-wide association studies (GWAS) have identified many loci, but the underlying functional variants and target genes remain largely unknown. Interpreting GWAS signals can be complicated due to linkage disequilibrium (LD) between multiple functional variants. Using circulating adiponectin levels, a complex trait, we sought to fine-map GWAS loci containing multiple signals. We performed a GWAS using 9,262 nondiabetic individuals from the Metabolic Syndrome in Men (METSIM) study imputed to the GoT2D reference panel (>19M variants). At each locus, we performed stepwise conditional analyses to identify additional "distinguishable" signals (P_{cond}<1x10^{-5}) within 1 Mb; when pairwise LD is low (r^2~0), the signals are also independent. We continued stepwise analyses until the lead variant P_{cond}≥1x10^{-5} or the signal harbored no coding variant or lead expression quantitative trait loci (eQTL). We annotated signals using eQTLs in subcutaneous adipose tissue samples from 1,381 METSIM subjects and the Epigenomics Roadmap adipose data. We identified two loci (*CDH13, ADIPOQ*) associated with adiponectin levels containing 2 and 8 distinguishable signals, respectively, 2 and 4 of which are independent. Each signal at *CDH13* consisted of ≥8 noncoding variants (LD r^2>0.8). The 1st signal contains a lead adipose eQTL variant (rs12051272) and overlaps active enhancer marks. Within the previously unreported 2nd *CDH13* signal, 2 variants overlap active enhancer marks. At *ADIPOQ*, each of the 8 distinguishable signals consisted of 1-15 variants (LD r^2>0.8). Of 45 known nonsynonymous variants in *ADIPOQ*, only 2 were variable in METSIM and only rs62625753 (G90S; P_{init}=3x10^{-3}, P_{cond}=6x10^{-4}) was associated with adiponectin. The lead adipose eQTL for *ADIPOQ* was the 8th ranked signal (rs35469083; P_{init}=1x10^{-44}, P_{cond}=2x10^{-3}).The trait-raising alleles of signals 1 & 2 and of signals 1, 4 & 8 share haplotypes with frequency >0.37 (D'≥0.94; r^2=0.05-0.33), which may explain the strength of signal 1. All 8 signals contain ≥1 variant in a putative enhancer and may harbor distinct functional variants. Accounting for multiple signals resulted in a 1.6-fold increase in variance explained over the lead signals alone (5.9 vs 9.4%). Taken together, fine-mapping and annotation allowed us to identify novel distinguishable adiponectin association signals and potentially novel functional coding and regulatory variants at loci with complex allelic heterogeneity.

332

Connecting the regulatory dots at the GWAS discovery phase. *A. Madar[1], D. Chang[1], F. Gao[1], A. Sams[1], Y. Waldman[1], C. Cunninghame Graham[2], T. Vyse[2], A. Clark[1,3], A. Keinan[1].* 1) Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, 14853, USA; 2) Divisions of Genetics and Molecular Medicine and Immunology, King's College London, Guy's Hospital London SE1 9RT, UK; 3) Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, 14853, USA.

GWAS analysis typically begins in a discovery phase that is followed by additional analysis on regions that reach genome-wide significance, typically p<5x10^{-8}. Other regions are often discarded, despite clear evidence that GWAS signal extends far beyond p=5x10^{-8}. Indeed, in the absence of auxiliary information, this procedure limits the number of false positives as higher significance variants more often replicate in independent studies. However, as we show here, variants that are supported by auxiliary information relevant for the disease under study, achieve equal replication rate to all SNPs discovered at 5x10^{-8} at a lower significance, e.g. 1x10^{-3} for Crohn's disease (CD) and 1x10^{-6} for rheumatoid arthritis (RA). As a consequence, hundreds of additional variants can be discovered and significantly replicated. The auxiliary information here is based on stratifying SNPs by cell-type specific activity patterns of regulatory DNA to which a SNP localizes. These patterns are identified by an algorithm we developed to integrate ENCODE DNaseI-seq data from multiple cell types. Beyond improved discovery, we used the activity patterns of disease-associated regulatory regions to provide insight into disease etiology. For example, we find (i) a more prominent B cell-specific regulatory activity in ulcerative colitis (UC) and lupus when compared to related diseases of CD and RA; (ii) for schizophrenia, a significant T cell-specific regulatory role, but a B cell- and monocyte-specific role in Alzheimer's disease. We also determine (using meme-chip motif discovery algorithm) which TF binding sites (TFBS) underlie specific activity patterns, e.g. the combination of Ets1, Runx1 and Rorc (the master regulator of Th17 T cells) is unique to Th17-specific regulatory elements. Finally, we determined which of the associated regulatory regions had an immune-relevant TFBS harboring a common SNP or indel. For the six autoimmune disease GWAS we analyzed (CD, RA, UC, lupus, multiple sclerosis, and type 1 diabetes), 422 regulatory regions were identified with an FDR<0.001, out of which 168 (~40%) had a TFBS harboring a common SNP (84 expected by chance in such regulatory regions; p=8x10^{-13}, fisher exact test). These 168 disease-associated TFBS serve as excellent candidates for follow-up analysis. We conclude that integrating information at the discovery phase increases genetic discovery and interpretability, and provides mechanistic hypotheses for GWAS regulatory signals.

333

Identifying critical cell types in complex traits from purified and single-cell expression using a polygenic model. *D. Calderon[1], D. Golan[2,3], T. Raj[2,6,7], J. Pritchard[2,4,5].* 1) Biomedical Informatics, Stanford University, Stanford, CA; 2) Department of Genetics, Stanford University, Stanford, CA; 3) Department of Statistics, Stanford University, Stanford, CA; 4) Department of Biology, Stanford University, Stanford, CA; 5) Howard Hughes Medical Institute, Stanford University, Stanford, CA; 6) Departments of Neurology and Medicine, Brigham and Women's Hospital, Boston, MA; 7) The Broad Institute of MIT and Harvard, Cambridge, MA.

Genome-wide association studies (GWAS) have identified thousands of common genetic variants associated with hundreds of human traits. While significant progress has been made in identifying these variants, the relevant pathogenic cell types and cell networks affected are not well characterized. Here, we describe a new method to identify trait-relevant cell types based on a polygenic model of disease that can utilize any publicly available purified or single-cell expression data and association summary statistics. We assign a cell-specific expression influence score to each SNP based on available expression data, and then estimate the individual contributions of each cell type to the observed phenotypic variance using a polygenic regression model (polyTest). We applied this method to association statistics of over 20 human diseases and other complex traits, and expression data from purified immune cell populations from ImmGen and single-cell of the hippocampus and cortex from Zeisel et al. We validate previous findings of associating rheumatoid arthritis with CD4+ helper T-cells (β = 8.67, P < 5e-13), B cells (β = 4.26, P < 5e-13), and CD8+ cytotoxic T cells (β = 7.87, P < 5e-10). We discovered many newly linked pathogenic cell types, including a significant enrichment for myeloid cell types such as microglia (β = 2.21, P < 5e-7) for Alzheimer's disease, GABAergic interneurons (β = 9.91, P < 5e-5) for Parkinson's disease, and pyramidal cells in the hippocampus (β = 6.72, P < 5e-5) and cortex (β = 6.24, P < 5e-5) for Schizophrenia. Thus, our method not only validated known cell types, but also identified novel cell types that may be relevant to a given disease. This approach represents a powerful framework for understanding the effect of common variants to cell type networks contributing to disease, and may help to prioritize pathogenic cell types for functional studies.

334

Integrating genome-wide association and co-expression network data for novel gene discovery. *C.R. Farber[1,2], L.D. Mesner[1], J.P. Stains[3], S.M. Tommasini[4], M.C. Horowitz[4], C.J. Rosen[5].* 1) Center for Public Health Genomics, University of Virginia, VA; 2) Departments of Public Health Science and Biochemistry and Molecular Genetics, University of Virginia, VA; 3) Department of Orthopaedics, University of Maryland, Baltimore, MD; 4) 4Department of Orthopaedics and Rehabilitation, Yale School of Medicine, New Haven, CT; 5) Maine Medical Center Research Institute, 81 Research Drive, Scarborough, Maine 04074.

Genome-wide association studies (GWAS) have identified >70 associations for fracture, bone mineral density (BMD) and other fracture-related quantitative traits. However, it has proven difficult to identify causal genes at associations using genetic data alone. Here, we analyzed BMD GWAS data in the context of a bone co-expression network to identify putative causal genes. The approach was based on the idea that groups of causal genes will influence BMD through the same process (e.g. osteoblast-mediated bone formation) and such genes are often co-expressed. We overlaid the homologs of 127 genes implicated by 64 SNPs robustly associated with BMD onto a mouse bone co-expression network. We found that 25 of the 127 genes were members of two of 21 network modules (6 and 9) (enrichment P≤0.002). Modules 6 and 9 were significantly enriched for genes involved in the function of bone-forming osteoblasts. Of the 25 genes, 17 are well-known regulators of osteoblast activity and BMD, such as *Lrp5* and *Sp7*, demonstrating our ability to recover known osteoblast genes. Furthermore, GWAS SNPs representing associations containing module 6 and 9 genes were more likely than the other GWAS SNPs to overlap activating histone marks (such as H3K27ac) in osteoblasts, but not other cell/tissue types. Based on these results we hypothesized that GWAS genes with homologs in modules 6 and 9 are strong positional and functional candidates and these associations influence BMD by altering osteoblast activity. To provide support for this prediction we focused on an association on Chr. 14 containing the module 9 member, MAP/microtubule affinity-regulating kinase 3 (*MARK3*). Using GTEx data, we found that the lead BMD SNP (rs11623869) was associated with *MARK3* expression in multiple tissues. Furthermore, knockdown of *Mark3* decreased osteoblast commitment, but increased osteoblast differentiation and activity. *Mark3* knockout mice (*Mark3*$^{-/-}$) also had increased femoral BMD, trabecular bone volume fraction (BV/TV) and cortical thickness. Additionally, the transcript levels of module 9 genes were specifically increased in *Mark3*$^{-/-}$ femurs. Together, our results implicate osteoblasts as drivers of variation in human BMD and identify *Mark3* as a novel regulator of BMD and potential causal BMD GWAS gene. Our results highlight the power of using co-expression network data to inform GWAS.

335

DNA.Land: A community-wide platform to collect millions of genomes-phenomes. *Y. Erlich[1,2], A. Gordon[2], N. Pearson[2], K. Shee[2], J. Pickrell[1,2].* 1) Computer Science, Columbia University, New York, NY; 2) New York Genome Center.

Understanding the genetic architecture of complex traits is a key challenge for personalized medicine. A decade of common- and rare-variant association studies have shown that finding genomic contributions to such phenotypes typically entails comparing tens of thousands of individuals. Obtaining the genomes and phenomes of cohorts at such scale using traditional ascertainment methods are logistically challenging and cost-prohibitive. But approximately three million US individuals have already obtained genome-wide information on their DNA using Direct to Consumer (DTC) genomics companies such as 23andMe, AncestryDNA, or FamilyTreeDNA. These individuals are usually driven by self-interest in genealogy and ancestry research. In our previous studies, we built a 13-million member family tree by crowdsourcing information from the same vibrant citizen genealogy community. Building on such work, we developed a web-platform DNA.Land (https://DNA.Land) where anyone can contribute her or his own DTC-generated genome data for research. A critical concept of DNA.Land is reciprocation. To serve participants' curiosity in their genomes and family histories, our platform is built to efficiently offer analyses unavailable through DTC companies, including whole-genome imputation, refined ancestry inference, and kin-matching across company cohorts. We hope to work closely, trustworthily, and fruitfully with participants, to apply the platform for scientific benefit. We will discuss our efforts to collect family trees and phenotype streams using social media, while respecting individuals' preferences according to our data sharing guidelines. Our vision is that this platform will serve the human genetics-wide community to reach the massive scale of data needed to understand complex traits. This abstract presents the official launch of our platform.

336

The European Variation Archive: Integrating Open-Access Variation Datasets. *G.I. Saunders, J.D. Spalding, I. Medina Castello, C. Yenyxe Gonzalez, J. Kandasamy, F.J. Lopez, I. Lappalainen, J. Coll-Moragon, J.M. Mut Lopez, J. Paschall.* European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom.

In recent years a number of high-profile genetic variation datasets have been made publically available by consortia or research groups such as 1000 Genomes, ExAC, deCODE and UK10K. However, these datasets are often prepared using independent methodologies or processing pipelines and can be accessed using only custom websites or FTP servers. I describe the first full release of the European Variation Archive (EVA) that works to gather, normalize and annotate via standardized pipelines the variants from such datasets, and datasets that are submitted directly to us; EVA thus represents the largest collection of open-access genetic variation data available, worldwide.EVA currently contains data from more than 50 studies, describing ca. 125 million unique variants from more than 100,000 human samples. Access to these data is possible via the EVA API (http://www.ebi.ac.uk/eva/?API) or website (www.ebi.ac.uk/eva), where we offer the ability to download in VCF or CSV format at the database, study or custom query (e.g. 'all missense variants within the PAX6 gene with an AF > 0.2') level.Where consent allows, EVA stores aggregated variant data from the EMBL-EBI controlled-access resource, the European Genome-phenome Archive (EGA; www.ebi.ac.uk/ega), allowing users to easily identify key EGA datasets of intertest. Additionally, EVA works in collaboration with GA4GH to provide the GlobalAlliance API against all loaded datasets and we are part of the push towards a federated system of global genetic variation data sharing. We also work in collaboration with ClinVar at NCBI to provide a clinically based browser, specifically for the ca. 135 thousand human variants that have been associated with at least one phenotype and a clinical significance from the ACMG classification.

337

The Monarch Initiative: An open science integrated genotype-phenotype platform for disease and model organism discovery. *M.A. Haendel[1], N.L. Washington[2], N. Vasilevsky[1], J. Nguyen-Xuan[2], C. Condit[3], D. Smedley[4], M. Brush[1], S. Köhler[5], T. Groza[6], K. Shefchek[1], H. Hochheiser[7], S.E. Lewis[2], P.N. Robinson[5], C.J. Mungall[2], The Monarch Initiative.* 1) Dept. of Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, USA; 2) Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA; 3) San Diego Supercomputing Center, UC San Diego, La Jolla, California, USA; 4) Wellcome Trust Sanger Institute, Mouse Informatics group, Hinxton, UK; 5) Charité - Universitätsmedizin Berlin, Institute for Medical and Human Genetics, Berlin, Germany; 6) Garvan Institute, Kinghorn Centre for Clinical Genomics, Sydney, Australia; 7) Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

Attempts at correlating phenotypic aspects of disease with causal genetic and environmental influences are often confounded by the challenges of interpreting diverse data distributed across numerous resources. New approaches to data modeling, integration, tooling, and community practices are needed to make efficient use of these data. The Monarch Initiative http://monarchinitiative.org is an international consortium working on the development of shared data, tools, and standards to enable direct translation of integrated genotype, phenotype, and environmental data from human and model organisms to enhance our understanding of human disease. We utilize sophisticated semantic mapping techniques across a diverse set of standardized ontologies to deeply integrate data across species, sources, and modalities. Using phenotype similarity matching algorithms across these data enables disorder prediction, variant prioritization, and patient matching against known diseases and model organisms. These similarity algorithms form the core of several innovative tools. The Exomiser, which enables exome variant prioritization by combining pathogenicity, frequency, inheritance, protein interaction, and cross-species phenotype data. Our Phenotype Sufficiency tool provides clinicians the ability to compare patient phenotypic profiles using the Human Phenotype Ontology to determine uniqueness and specificity in support of variant prioritization. The PhenoGrid visualization widget illustrates phenotype similarity between patients, known diseases, and model organisms. Monarch develops models in collaboration with the community in support of the burgeoning genotype-phenotype disease research community. Our data, tools, and models are freely available at GitHub, and community contributions are welcome. Modular software development supports use of our tools in third-party applications, with active efforts involving widely-adopted patient phenotype capture and management systems, such as Appistry's CloudDX, NIH UDPICS, PhenoTips, and the KCCG Patient Archive. We have successfully used Exomiser to solve a number of undiagnosed patient cases in collaboration with the NIH Undiagnosed Disease Program. Ongoing interaction with the Global Alliance for Genetic Health (GA4GH) and other groups will catalyze the realization of our goal of a vital translational community focused on the collaborative application of integrated genotype, phenotype, and environmental data to human disease.

338

iCLiKVAL: an open-access tool for adding value to scientific literature one annotation at a time through the power of crowdsourcing. *T.D. Taylor, N. Kumar.* Laboratory for Integrated Bioinformatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan.

There are nearly 26 million citations to various forms of scientific literature in PubMed. Searching this vast resource does not always give desirable or complete results for several reasons: missing abstracts, unavailability of full-article text, non-English articles, lack of keywords, etc. Ideally, every citation should include a complete set of keywords or terms that describe the original article in enough detail that searches, using natural language, return more relevant results; however, this would require countless hours of manual curation.Our goal is to make manual curation 'fun', social, and self-correcting, thus enriching resources like PubMed so that users are able to extract more valuable and relevant results, as well as improve discovery of their own work. We developed a web-based open-access tool for manual curation of PubMed articles, and other media types, using a crowdsourcing approach. We encourage the use of ontologies and support them as auto-suggest keyword terms, but we do not restrict users to these so as not to impose any limitations on the annotation types. Non-English annotation is also supported. Through this 'non-restrictive' approach, we hope to encourage more researchers, not just trained curators, to make use of this tool.We constructed a cross-browser and platform-independent application using the latest web technologies and a NoSQL database. Users perform searches to identify articles of interest, mark articles for review, load PDFs into the viewer, select annotations (values) within the text, and add appropriate keywords (keys) and relationship terms, if applicable. Article-specific comments can be made, key-value pairs can be edited and rated, live chats between users can be conducted, annotations can be added via Twitter, etc. Users can create both private and public communities and groups in which to work together on shared sets of articles. In addition, they can create project-specific controlled vocabularies with which to annotation the articles.As annotations accumulate in the database, our semantic search feature will improve and the more relevant the results, which can be precisely filtered. Where possible, we will pre-populate the database with other curated data from publicly available resources, and, via our REST API, make the annotations easily accessible to the entire research community. We are working to establish a single, easy-to-use resource for community-based curation of all online scientific literature.

339

A Practical Guide to Drug Discovery through Phenome-Wide Association Studies. *F. Sathirapongsasuti, D.A. Hinds, E. Karrer.* 23andMe Inc., Mountain View, CA.

Increased understanding of human genetics has the potential to inform drug discovery and development, a challenging and risky endeavor traditionally plagued by a high rate of failure at clinical stages. Phenome-wide association studies (PheWAS), which measure associations of defined genetic variants across a wide range of phenotypes, are made possible by routine collection and sequencing/genotyping of biospecimens, analyzed in conjunction with phenotypes from surveys or electronic medical records. Through the lens of the experiments of nature, PheWAS emerges as a promising method to identify new disease indications for existing drugs or novel targets. We propose a framework to identify such potential disease indications for targets deemed to be druggable. To illustrate the practical viability of our approach, we interrogated known drug targets: IL13 and IL4, both in a cytokine cluster in the 5q31 region, using 23andMe PheWAS results spanning 736 curated phenotypes. SNPs within IL13 and IL4 loci are highly associated with several immune and autoimmune conditions such as asthma, allergy, psoriasis, eczema, and rosacea. A conditional analysis with forward stepwise regression further deconvolutes the associations into several independent signals and highlights a missense mutation in IL13 as a possible causal variant. Clinical data supports our findings as IL13 antibodies lebrikizumab and tralokinumab are in Phase III clinical trials for the treatment of asthma while a dual IL4/IL13 receptor antibody dupilumab is in Phase III development for eczema. Psoriasis appears to be more strongly associated with IL13 variants, suggesting a repurposing opportunity for IL13-targeting drugs. This study helps demonstrate the potential application of PheWAS to identify disease indications for drug targets, with implications for improving the success rate and speed of drug discovery.

340

The Human Phenotype Ontology: Semantic unification of common and rare disease. *P. Robinson[1], S. Köhler[1], D. Moldenhauer[2], N. Vasilevsky[3], G. Baynam[4], T. Zemojtel[1], L. Schriml[5], W. Kibbe[6], P. Schofield[7], T. Beck[8], D. Vasant[9], A. Brookes[8], A. Zankl[10], N. Washington[11], C. Mungall[11], S. Lewis[11], M. Haendel[3], H. Parkinson[9], T. Groza[10].* 1) Institute for Medical Genetics, Charite-Universitaetsmedizin, Berlin, Germany; 2) University of Applied Sciences, 35390 Giessen, Germany; 3) Oregon Health & Science University, Portland, OR 97239-3098, USA; 4) School of Paediatrics and Child Health, University of Western Australia, Perth, 6840, Western Australia; 5) Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA; 6) Center for Biomedical Informatics and Information Technology, National Cancer Institute,Rockville, Maryland, 20850, USA; 7) University at Cambridge, Department of Physiology, Development and Neuroscience,Cambridge CB2 3EG, UK; 8) Department of Genetics, University of Leicester, Leicester LE1 7RH, UK; 9) European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD United Kingdom; 10) Garvan Institute of Medical Research, Darlinghurst, Sydney, NSW 2010, Australia; 11) Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

The Human Phenotype Ontology (HPO) provides a structured, comprehensive and well-defined set of over 11,000 classes (terms) that describe phenotypic abnormalities seen in human disease. The HPO project additionally provides a collection of over 116,000 disease-phenotype annotations to over 7,000 rare diseases; for instance, the disease Marfan syndrome [MIM:154700] is annotated to the HPO terms *Arachnodactyly* [HP:0001166], *Ectopia lentis* [HP:0001083], and 46 other HPO terms. The HPO has been used to develop algorithms and computational tools for differential diagnostics, prioritization of candidate disease-associated genes, as well as exome and clinical exome sequencing studies. A comparable resource has not been available for common diseases.Here we develop a concept recognition procedure that analyzes the frequencies of annotations of diseases to HPO terms as identified in over 5 million PubMed abstracts, employing an iterative procedure to optimize precision and recall of the identified terms. We derived disease models for 3145 common (complex) human diseases, comprising a total of 132,006 HPO annotations. We estimated the overall quality of the HPO annotations using a set of 41 randomly chosen common diseases that were subjected to comprehensive manual biocuration. The overall F-score (i.e., the harmonic mean of precision and recall) was 45.1% (mean Precision: ~60% and mean Recall: ~40%). We used the annotations to generate a network of phenotypic similarity for 1678 diseases belonging to 13 Disease Ontology categories such as *nervous system disease* or *respiratory system disease*, and demonstrated a highly significant pattern of interconnected phenotypic clusters. To understand the genetics of complex disease, it is important to consider the phenotypic and genetic overlap amongst diseases. For instance, ulcerative colitis and Crohn's disease share multiple susceptibility loci ("GWAS hits"). We investigated a total of 16,152 GWAS hits, of which 863 were associated with more than one disorder. We demonstrated pervasive and highly significant phenotypic sharing amongst complex diseases that share associations to the same SNP. A similar analysis showed phenotypic sharing between genetically linked rare and common diseases. The annotations presented here, as well as the HPO itself, are freely available.

341

Imputation in the Cloud: Lessons Learned and Future Directions. *C. Fuchsberger[1,2], L. Forer[3], S. Schönherr[3], D. Sayantan[1], F. Kronenberg[3], G. Abecasis[1].* 1) Ctr Statistical Genetics, Univ Michigan, Ann Arbor, MI; 2) Center for Biomedicine, European Academy of Bolzano/Bozen (EURAC), Bolzano, Italy; 3) Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria.

Genotype imputation is a key step in the analysis of genome-wide association studies (GWAS). The process is computationally demanding and typically requires access and familiarity to substantial computational resources as well as to a reference panel of sequenced genomes. Last year we have introduced a web-based service, called Imputation Server (https://imputationserver.sph.umich.edu), that facilitates access to new reference panels and simplifies user experiences. Here, we describe the lessons learned from providing a highly-accessed and privacy sensitive web-service to the genetic community. As of abstract submission, we have processed >700,000 genomes from >250 users. Moreover, we have extended our service to allow not only imputation using the 1000 Genomes Project and HapMap Consortium reference panels but also the Haplotype Reference Consortium European ancestry panel (64,976 haplotypes, 39,235,157 SNPs) which is expected to extend imputation to variants with frequencies of 0.1 – 0.5%. This panel is an agglomeration of disease study haplotypes and is otherwise cumbersome to access directly as a result of participant privacy protections as well as the large volumes of data involved. We show how the imputation server framework can be used as a model for other genetic analysis services, such as for the processing of mitochondrial next generation sequencing data (http://mtdna-server.uibk.ac.at) or the estimation of genetic ancestry (http://laser.sph.umich.edu). We discuss how these web-services can accelerate genetic research by greatly simplifying analysis steps for most users, allowing many researchers to devote their time to more interesting tasks.

342

LARVA: an integrative framework for Large-scale Analysis of Recurrent Variants in noncoding Annotations. *M. Gerstein[1,3,4], L. Lochovsky[1], J. Zhang[1], Y. Fu[1], E. Khurana[2].* 1) Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT; 2) Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY; 3) Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT; 4) Department of Computer Science, Yale University, New Haven, CT.

In cancer research, background models for mutation rates have been extensively calibrated in coding regions, leading to the identification of many driver genes, recurrently mutated more than expected. Noncoding regions are also associated with disease—for instance, recent studies have implicated the promoters of *TERT* (MIM: 187270), *PLEKHS1*, *WDR74*, and *SDHD* (MIM: 602690) in multiple cancer types. However, background models for noncoding regions have not been investigated in as much detail as coding regions. This is partially due to limited noncoding functional annotation. Also, great mutation heterogeneity and potential correlations between neighboring sites give rise to substantial overdispersion in mutation count, resulting in problematic background rate estimation. Here, we address these issues with a new computational framework called LARVA. It integrates variants with a comprehensive set of noncoding functional elements, modeling the mutation counts of the elements with a beta-binomial distribution to handle overdispersion. Moreover, LARVA uses regional genomic features such as replication timing to better estimate local mutation rates and mutational hotspots. We demonstrate LARVA's effectiveness on 760 whole-genome tumor sequences, showing that it identifies well-known noncoding drivers, such as the *TERT* promoter. Furthermore, LARVA highlights several novel highly mutated regulatory sites that could potentially be noncoding drivers. We make LARVA available as a software tool, and release our highly mutated annotations as an online resource (larva.gersteinlab.org). LARVA's broad applicability allows it to be used for the discovery of high mutation burden elements for any type of genetic disease. Furthermore, this framework may be used to find high germline mutation burdens, which may correspond to genetic disease risk alleles.

343

Leveraging variant prioritization information in *de novo* mutation analysis to identify novel autism candidate genes. *C.D. Huff[1], H. Hu[1], M. Li[2], H. Coon[3], M. Yandell[2].* 1) Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX; 2) Department of Human Genetics and USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, Utah; 3) Department of Psychiatry, University of Utah, Salt Lake City, Utah.

Evidence for a role of de novo mutations in human disease continues to accumulate. In Autism Spectrum Disorder (ASD), *de novo* mutations are estimated to account for approximately 30% of all sporadic cases. Existing tests for identifying *de novo* mutations contributing to disease risk are based on the total number of *de novo* mutations in each gene, with the significance level calculated based on gene length and sequence composition. The choice of which mutations to include in the test can have a considerable affect on power. Nonsense and splice junction mutations are nearly always evaluated, while missense mutations are often excluded. Functional variant predictors such as SIFT and PolyPhen-2 offer orthogonal information regarding potential disease causality. However, functional variant prediction information cannot be readily incorporated into existing *de novo* mutation tests except through highly suboptimal binary classification. To overcome this shortcoming, we propose a new statistical test, VARiant PRIoritization SuM (VARPRISM). VARPRISM assigns a functional weight to each *de novo* mutation and sums up the total weight from all *de novo* mutations in a gene to generate the test statistic. The functional weight is a Conservation-Controlled Amino Acid Substitution matrix (CASM) score introduced in VAAST 2. VARPRISM calculates the significance level of the test statistic through simulation, accounting for differential sequence composition and mutation rates. We demonstrate that VARPRISM substantially outperforms existing approaches in an analysis of reported *de novo* mutations in known ASD genes, with a 17% increase in relative power when 0.25% of subjects carried a de novo mutation in the gene of interest. We analyzed an existing dataset of 2,508 parent-offspring autism trios using this new approach, replicating 36 previously known autism susceptibility genes and identifying 14 novel candidate genes at FDR < 0.3. These candidate genes were significantly over-represented in schizophrenia ($p = 1.1x10^{-4}$) and intellectual disability ($p = 1.4x10^{-10}$) pathways as well as the learning and memory GO term ($p = 4.0x10^{-3}$). VARPRISM is implemented in the pVAAST software package; the underlying statistical framework in VARPRISM can also be used to jointly analyze *de novo* mutations and inherited variants in pedigrees in pVAAST.

344

SimDenovo: A simulation toolkit to understand the variability in *de novo* mutation burden in human disease. *V. Aggarwala[1], B.F. Voight[2,3].* 1) Genomics and Computational Biology Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA; 2) Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, PA; 3) Department of Genetics, Perelman School of Medicine, University of Pennsylvania, PA.

The rate of *de novo* mutation, a fundamental force in molecular evolution, varies substantially across the genome and strongly associates with complex diseases like Autism and Schizophrenia. Determining the factors underlying this variability holds the promise to discover novel disease-causing genes, detect signatures of natural selection, and identify functional, non-coding genomic elements with regulatory potential. Family based designs have been employed to estimate the rate of *de novo* mutation, though the sparsity of *de novo* events (~60 per family, and ~1 per exome) limits the resolution in which this variability can be characterized at fine scale.We hypothesized that population level data could instead be used to estimate variability in *de novo* rates across the genome. To test this hypothesis, we employ a novel sequence context based approach using millions of SNPs from the 1000 Genomes Project (Phase I). The key ideas of our approach are (a) determining the rate of polymorphism for windows of sequence contexts of different lengths, and (b) normalizing our context-based rates to the previously observed overall genome-wide *de novo* mutation rate of 1.2×10^{-8} per generation per site (Kong et al, Nature 2012).To validate our approach, we obtained *de novo* events from whole genome sequences of 78 trios (Kong et al, Nature 2012). We demonstrate that (i) a hepta-nucleotide sequence context window explains the distribution of observed *de novo* events better than the currently used tri or di-nucleotide context estimates (likelihood improvement of $>> 10^{100}$), and (ii) sequence motifs that predict a higher mutation rate (ApT dinucleotides, CAA<u>T</u>, or TA<u>C</u>G motifs) are enriched for *de novo* events ($P < 10^{-6}$).Based on our hepta-nucleotide sequence context *de novo* rate estimates, we report additional, undocumented variability in protein coding regions, resulting in a *de novo* rate of 1.72×10^{-8} per generation per site in exomes. We developed software that allows users to simulate *de novo* events in any number of individuals over any genomic region, based on our rates. Our tool - along with novel prioritization functions - can be used to develop new, nonparametric tests for evaluating burden of *de novo* mutations in complex disease.

345

Relationship Inference in Big Genetic Data with >100,000 samples. *W.-M. Chen, A. Manichaikul, S.S. Rich.* Center for Public Health Genomics, University of Virginia, Charlottesville, VA.

Motivation: Technological advances in high-throughput sequencing and custom genotyping arrays are making genetic studies larger than ever. An algorithm to accurately infer the degree of relationship between a pair of individuals using high-throughput SNP data without using allele frequency information, as implemented in our software package KING, has allowed rapid relationship inference in large datasets consisting of 10,000s of samples. However, relationship inference in even larger datasets (>100,000 samples) remains computationally challenging to existing software implementations. **Methods:** We present an improved algorithm for inference of close relationships (duplicates, 1st- or 2nd-degree relatives) using high-throughput genotype data. To achieve improved efficiency, we use a two-stage approach whereby we first select a subset of highly informative SNPs to obtain first-pass relationship estimates for all pairs of individuals, and follow-up using all SNPs on a small subset of pairs identified in the first stage analysis.**Results:** In a dataset consisting of ~30,000 individuals (recently published in Nature Genetics 47:381-6), out of ~450 million pairs of individuals to be examined, our new implementation correctly identified all pairs of MZ twins/duplicates in <1 minute on a single CPU, which projects to < 1 day among 1 million individuals. For the same dataset, it took 5 minutes to correctly identify all pairs of first-degree relatives. In another dataset that is from the 1000 Genomes Project (phase3v5), 2,504 individuals each with 81 million autosomal SNPs were analyzed. It took our new implementation 1 second, 2 seconds and 12 seconds respectively to correctly identify duplicates, and close relationships up to 1st- and 2nd-degree respectively, in contrast to 20 hours using the standard KING-robust algorithm.**Availability:** Our improved relationship inference algorithms are implemented in a freely available software package, KING, available for download at http://people.virginia.edu/~wc9c/KING.

346

Mixed Model Association with Family-Biased Case-Control Ascertainment. *T. Hayeck[1,2], N. Zaitlen[3], P. Loh[2,4], A. Gusev[2,4], N. Patterson[2], A. Price[1,2,4].* 1) Biostatistics, Harvard T.H. Channing School of Public Health, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 3) Lung Biology Center, School of Medicine, University of California, San Francisco, San Francisco, CA 94158,USA;; 4) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA.

Mixed Models have become the tool of choice for genetic association studies; however, existing mixed model methods may be poorly calibrated or underpowered in settings of family sampling bias and/or case control ascertainment. In our previous work (Hayeck et al. 2015 AJHG), we introduced a liability threshold based mixed model association statistic (LTMLM) that addresses the power loss of standard mixed model methods under case-control ascertainment. Here, we consider family-based case-control ascertainment, in which cases and controls are ascertained non-randomly with respect to family relatedness. Previous work has shown that this type of ascertainment can severely bias heritability estimates (Zaitlen et al. 2013 PLoS Genet), and we show here that it also impacts mixed model association statistics. We introduce a family based association statistic (LT-Fam) that is robust to this problem. Similar to LTMLM, LT-Fam is computed from posterior mean liabilities (PML) under a liability threshold model conditional on every individual's case-control status and the disease prevalence. To avoid biased heritability estimation, published narrow-sense heritability estimates are used to construct a properly calibrated LT-Fam statistic. In sib-pair simulations at varying levels of family ascertainment bias, the LT-Fam Statistic was correctly calibrated and achieved higher power than existing mixed model methods. At a prevalence of 1% for concordant sib pairs LT-Fam was properly calibrated (average chi-square = 0.999), while the Armitage Trend Test (ATT) was inflated (1.500) and standard mixed linear models (MLM) were deflated (0.742). Additionally, a 9% increase in test statistics was observed at causal markers after correcting other statistics for mis-calibration. The magnitude of the improvement depends on severity of family ascertainment bias, sample size, relatedness structure, and severity of case-control ascertainment. Looking at type 2 diabetes (T2D) case control data in the Jackson Heart Study (JHS), we down-sampled to increase relatedness among cases and again we again observed: ATT was inflated (1.337) and MLM was deflated (0.820), while LT-Fam was properly calibrated 0.980(0.002). In summary, we have demonstrated an increase in power with correct calibration for the LT-Fam association statistic under family-biased case-control ascertainment.

347

Dissecting a major linkage signal to identify potential causal variants for serum triglycerides in a founder population. *W.-C. Hsueh, A.K. Nair, S. Kobes, L.J. Baier, R.L. Hanson.* NIDDK, NIH, Phoenix, AZ.

Previously, we identified a major locus for serum triglycerides (TG) in 1,007 Pima Indians using a population-based genome-wide linkage analysis that uses the empirically estimated % of allele sharing IBD for all ~506,000 pairs of subjects. The locus on chr. 11q (LOD=9.3) explained 10.8% of variance (σ^2) in TG. Conditional measured genotype analyses identified 3 independently associated variants (rs147210663, rs2072560, and rs11357208) with nominal $p < 0.05$ that explained nearly the entire linkage signal. The SNP with the strongest association ($p=1.5*10^{-13}$), rs147210663, explained 6.9% σ^2. Its association was replicated using 4,002 additional samples ($p=3.1*10^{-47}$, % s2 explained: 4.7%). This SNP codes for an AlayThr substitution (A43T) with a known effect on APOC3 function, but did not entirely explain the linkage signal (reducing LOD to 2.2). Thus, we conducted further follow-up association studies of additional candidate variants in the region using a replication sample composed of 4,467 full- or partial-heritage Pima Indians.We conducted replication association analyses of rs2072560, rs11357208 and 7 additional SNPs, conditional on effects of rs147210663. The 7 additional SNPs were all in moderate to strong LD with rs2072560, including 4 SNPs reported by published genome-wide association studies (GWAS) of TG. Rs964184, in moderate LD with rs2072560 near the 3' UTR of *ZPR1*, had the strongest association with TG among all SNPs tested ($p=3.4*10^{-20}$, % σ^2 explained: 2.3%). This is the lead SNP from several previous GWAS, but has no known functional effects; thus, we analyzed additional variants with known functions related to lipid metabolism. Three SNPs in the adjacent *APOA5* locus (rs651821 at the 5'UTR, rs662799 in the promoter, both affecting *APOA5* expression; and rs3135506, a missense S19W SNP) were associated with TG ($p=2.7*10^{-13}$, $p=1.8*10^{-13}$ and $p=2.3*10^{-5}$, respectively). Including these 3 SNPs in a conditional analysis rendered the effect of rs964184 on TG non-significant (p=0.88, % σ^2 explained: ~0). We did not replicate an independent association between rs11357208 and TG, which may be partly due to the fact that this SNP is in moderate LD with rs964184 ($r^2=0.56$). These analyses suggest that the major locus for TG on chr. 11q in Pimas represents the effects, not only of the A43T SNP in *APOC3* (accounting for ~5% σ^2), but also of additional functional variants in/near a strong candidate gene (*APOA5*), which collectively account for >2% σ^2 in TG.

348

Systematic and large-scale investigation of twin and sibling concordance of 1723 traits in a nationally representative health claims cohort. *C.M. Lakhani[1], J. Yang[2], P.M. Visscher[2], C.J. Patel[1].* 1) Center for Biomedical Informatics, Harvard Medical School, Boston, MA; 2) Queensland Brain Institute, The University of Queensland, Brisbane, 4072 Queensland, Australia.

Family-based studies have been instrumental in understanding the relative contributions of differences genes and environments in variation in traits and disease. However, most documented twin studies have had low sample sizes and have not examined the heritability in different populations or strata systematically to document evidence of gene-environment interactions, leading to fragmented literature of twin-based heritability estimates. In this study, we re-purpose a US-based health claims data set in order to estimate concordance rates of 1723 traits in sibling (sib) and twin children (born on or after 1986). In our dataset (Total N= 34M), we have ascertained 750K female/female pairs (FF) sib pairs, 786K male/male sib pairs (MM), and 1.5M mixed gender sib pairs (MF). Further, we have ascertained 47K twin FF pairs, 45K twin MM pairs, and 47K twin MF pairs. For each of the 1723 traits (classified as combinations of ICD9 codes), we calculated sib and twin pro-bandwise concordance. The average concordance for 1723 traits was 0.0888 for twin MM (interquartile range [IQR]=(0, 0.22)), for twin FF 0.02 (IQR=(0,0.2)), for twin MF 0.03 (IQR=(0,0.13), for sib MM 0.0219 (IQR=(0, 0.0570)), for sib FF 0.02 (IQR=(0.004, 0.06)), and for sib MF 0.017 (IQR=(0.0006, 0.05)). For example, the twin MM concordance for asthma was 0.45 (n=3.8K, 95% confidence interval [CI]=(0.43, 0.47)), for twin FF 0.41 (n=3.1K, CI=(0.39, 0.43)), twin MF was 0.34 (n=4.5K, CI=(0.32, 0.36)). As expected, the sib concordances were lower, 0.27 (n=70K) 0.23 (n=55K), and 0.24 (N=126K) for MM, FF, and MF respectively. Similarly, for autism we observed concordance rates of 0.43 (n=745, CI=(0.39, 0.47)) for twin MM, 0.41 (n=249, CI=(0.35, 0.49)) for twin FF, 0.17 (n=714, CI=(0.13, 0.21)) for twin MF which were significantly higher than sib rates concordance of 0.13 for sib MM, 0.07 for sib FF, and 0.07 for sib MF respectively. In this report, we will estimate the heritability of traits using a modified variance components analysis and take advantage of demographic information (e.g., geography, sociodemographics, and age of parents) to compute 1723 concordances in different populations to search for evidence of gene-environment interactions. To our knowledge, our twin and sibling study is the largest of its kind in sample size and scope, providing a robust and comprehensive list of concordances in multiple populations in the US.

349

Heritability estimates for thirty-four traits in a large Ugandan cohort. *D. Heckerman[1], D. Gurdasani[2], C. Pomilla[2], R. Nsubuga[3], C. Kadie[4], C. Widmer[1], M. Sandhu[2].* 1) Microsoft Research, Los Angeles, CA; 2) Wellcome Trust Sanger Institute, Hinxton, UK, and the Department of Medicine, Cambridge; 3) MRC/ UVRI Uganda Research Unit on AIDS, Uganda; 4) Microsoft Research, Redmond, CA.

Heritability estimates for most traits are unknown in Africa. To address this shortcoming, we collected genome-wide SNP data for 4,778 individuals sampled throughout Uganda along with 34 traits including height, weight, results from a blood panel, and cholesterol measurements. In addition, we collected GPS locations as a proxy for environmental effects. To estimate narrow-sense heritability, we used a novel approach involving a linear mixed model with two random effects—one based on genetic markers and one based on GPS locations. Specifically, for the genetic random effect, we used identity-by-descent estimates from long-range phased genome-wide data (366K SNPs). For the environmental random effect, we constructed a radial basis function kernel, where the entry for a pair of individuals was the exponential of the negative scaled distance between the two individuals. The scaling parameter, as well as the weights of the two random effects, was determined by maximizing the restricted likelihood of the data. Without the environmental random effect, the estimate of heritability would be inflated whenever genetics and environment were correlated, because the influence of genetic markers would be counted twice—once through their direct effect on the trait and once through their effect on the trait via environment. Indeed, for all traits, heritability estimates were lower when GPS location was taken into account. Estimates varied from relatively modest (*e.g.,* 10% for the liver biomarker GGT) to substantial (*e.g.,* 71% for mean cell hemoglobin concentration). In addition to better estimates of heritability, this approach allowed us to estimate the amount of variance explained by location, which was significantly greater than zero for 22 of the traits. For each trait, we were also able to determine the geographical distance of the environmental effect (*i.e.,* the optimized value of the scaling parameter), which varied by more than four orders of magnitude across the traits. Finally, heritability estimates of the Ugandan cohort and the European cohort of Zaitlen *et al.* 2013 (neither corrected for location) showed some significant differences including those for height 50±4% vs. 69±2% and LDL 60±4% vs. 20±6%, possibly reflecting differences in allelic architecture of variants associated with these traits, and/or differences in interaction with environmental determinants between populations. Software is available at https://github.com/MicrosoftGenomics/FaST-LMM.

350

Leveraging whole genome sequencing in an internal study-specific imputation reference panel for family-based designs. *K. Iyer[1], L.R. Yanek[1,2], M.A. Taub[3], I. Ruczinski[3], D. Becker[1,2,3], L. Becker[1,2], R.A. Mathias[1,2,3].* 1) Medicine, Johns Hopkins University, Baltimore, MD; 2) GeneSTAR Research Program, Johns Hopkins University, Baltimore, MD; 3) Public Health, Johns Hopkins University, Baltimore, MD.

GWAS have identified common disease associated variants diseases either through direct genotyping or imputation relying on the Thousand Genomes Project (TGP). Importantly, rare variants relevant to disease are likely enriched in a study-specific reference panel ascertained on the basis of the disease in contrast to a population-based sample like the TGP. It is widely acknowledged that the use of an internal reference panel could facilitate better imputation of rare and disease-relevant variants. Here, we quantify the benefit of this strategy in a family-based sample comparing African Americans (AA) to European Americans (EA). Whole genome sequence data were available on N=129 AAs and N=157 EAs from the GeneSTAR Study comprising families ascertained on an early onset coronary artery disease proband. Genomes were phased in SHAPEIT following published TGP and protocol. Imputation was performed in the remaining GeneSTAR subjects (N=1203 and N=1990 AAs and EAs, respectively) using ~1M framework GWAS SNPs and IMPUTE2 for two different reference haplotype datasets (1) N=286 internal GeneSTAR genomes; and (2) N=1092 external population-based genomes from TGP. Comparisons between the two imputations was limited to high quality (imputation r2 metric >0.3) imputed SNPs. The GeneSTAR reference panel includes 23 and 14 million (M) total variants in AAs and EAS, respectively, of which 21.3M and 12.4M were successfully imputed. In contrast, more SNPs were imputed using the TGP reference panel (36.2M in AAs and 26M in EAs). However, three important metrics support the benefit of the internal reference panel: (1) we imputed 4.3M novel variants in the AAs and 2.4M novel variants in the EAs using the GeneSTAR reference panel; (2) <0.05% of successfully imputed variants were monomorphic when relying on an internal reference panel in contrast to 7% of those imputed from TGP being monomorphic/irrelevant in the study sample; and (3) the majority of imputed variants using the internal reference panel are rare with an MAF<5% (64% and 56% in AA and EA, respectively). These results offer strong metrics on the benefits in using an internal reference panel. Perhaps most important is that novel variants imputed represent a set of segregating variants *private* to the families in this disease-ascertained sample that are potentially important phenotype associations under our ascertainment scheme. Validation of this benefit through tests for association with key phenotypes is underway.

351

Combined analysis of over 60,000 exomes: genic constraint, widespread mutational recurrence, and impact on clinical variant interpretation. *D. MacArthur[1,2], Exome Aggregation Consortium.* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA.

The discovery of genetic variation has been empowered by the growing availability of DNA sequencing data from large studies of common and rare diseases, but these data are typically inconsistently processed and largely inaccessible to most genetics researchers. To increase the availability of such data, the Exome Aggregation Consortium (ExAC) has collected and performed joint variant calling across over 90,000 individuals sequenced in diverse population genetic and disease studies. Using extensive independent validation data we demonstrate that our joint variant calling approach improves accuracy, sensitivity and consistency of rare variant detection. At ASHG 2014 we publicly released variants and frequency data for a subset of over 60,000 exomes, a resource that has rapidly become the default reference data set for clinical genetics, and received over one million page views in its first six months of operation. Here we describe the scientific results that have emerged from analyses of this data set since its public release. We show that the unprecedented scale of the ExAC data set empowers the calculation of the depletion of loss-of-function variants in human genes, flagging over 1,000 genes with strong evidence for haploinsufficiency for which human phenotypes have not yet been characterized. We describe the surprising observation of widespread mutational recurrence (the same mutation arising multiple times independently), and demonstrate the strong impact of this phenomenon on many analyses (such as frequency spectra) in large collections of genetic variation. Finally, we show that many previously reported disease-causing mutations are also seen in ExAC individuals, and describe manual curation of the literature support for over 200 such variants, indicating the prevalence of interpretation errors as well as incomplete penetrance.

352

Population differentiation analysis of 54,734 European Americans reveals independent evolution of *ADH1B* gene in Europe and East Asia. *K.J. Galinsky[1,2], G. Bhatia[2,3], P. Loh[2,3], S. Georgiev[4], S. Mukherjee[5], N.J. Patterson[2], A.L. Price[1,2,3].* 1) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; 3) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA; 4) Google, Palo Alto, CA; 5) Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, NC.

Population differentiation is a widely used approach to detect the action of natural selection. Existing methods search for unusual differentiation in allele frequencies across discrete populations, e.g. using F_{ST}. Loci that are unusually differentiated with respect to the genome-wide F_{ST} or with respect to a null distribution of F_{ST} are reported as signals of selection. These approaches are particularly powerful for closely related populations with large sample sizes.However, population genetic data often is not naturally partitioned into discrete populations. We developed a test for selection that uses SNP loadings from principal components analysis (PCA). For a given PC reflecting geographic ancestry, under the null hypothesis of no selection, the square of the SNP loadings, rescaled by a scaling factor derived from the eigenvalue of the PC, follows a chi-square (1 d.o.f.) distribution. This statistic is able to infer selection with genome-wide significance, a key consideration in genome scans for selection. We confirmed via simulations that this statistic has correct null calibration under a wide range of demographies and is well-powered to detect selection at large sample sizes.We applied the method to a cohort of 54,734 European Americans genotyped on genome-wide arrays. PCs were inferred using our FastPCA software (running time: 57 minutes). The top 4 PCs corresponded to clines of Irish, Eastern European, Northern European, Southeast European and Ashkenazi Jewish ancestry, validated via PCA projection of samples of known ancestry. We detected genome-wide significant signals of selection at 4 known selected loci (*LCT*, *HLA*, *OCA2* and *IRF4*) and 3 novel loci: *ADH1B*, *IGFBP3* and *IGH*. 2 of the 3 novel loci could not be detected using discrete-population tests (or other existing tests). The *ADH1B* gene is associated with alcoholism (via the same coding SNP rs1229984 producing a signal in our selection scan) and has been shown to be under recent selection in East Asians (via a haplotype-based test for recent selection); we show here that it is a rare example of independent evolution on two continents. The *IGFBP3* gene and *IGH* locus have been implicated in breast cancer and multiple sclerosis, respectively. Our results show that application of our PC-based selection statistic to large data sets can infer novel, genome-wide significant signals of selection at loci linked to disease traits.

353

A direct estimate for the human mutation rate from autozygous sequences in thousands of parentally related pedigrees. *V. Narasimhan[1], R. Rahbari[1], A. Scally[2], Y. Xue[1], C. Tyler-Smith[1], R. Durbin[1].* 1) Dept of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK; 2) Dept of Genetics, University of Cambridge, Cambridge, UK.

There is still an ongoing debate over the estimate of the human mutation rate and examining heterozygous mutations within autozygous sequences offers a way to ascertain mutations that have occurred over a number of generations. In this study, we exome sequenced 4353 individuals with a range of parental relatedness to obtain a direct estimate using this approach. We employed a novel approach to discovering false negatives by alternating bases of reads covering a certain locus at appropriate ts/tv ratios at a set of random sites with a Bernoulli process and then recalling variants using these remapped reads through the same calling pipeline and accounted for false positives by resequencing biological replicates of 176 pairs. We mitigate the effect of gene conversion events by removing mutations already segregating in the population as well as examining clusters of mutations very close to one another. Using a supervised clustering approach we infer the number of generations to the TMRCA directly from the length distribution of the autozygous sections observed in each genome and verify this with pedigree information. To reduce errors in the calling of the autozygous segments as well as regions from higher order relationships that might appear directly adjacent and thereby confound a called region, we show that the mutation rate estimates obtained are insensitive to the number of bases we evaluate from each end of the autozygous segment. As we are ascertaining the mutation rate across multiple generations we greatly reduce the impact of under or over-counting due to mosaic mutations occurring during development of the germ-line and/or blood lineage, the rates of which have not been accurately assayed in humans. After carefully accounting for these factors, from autozygous sequences of >10Gb that span the entire genome, we observed 932 de novos mutations with an average number of separating generations of 6.63, which corresponds to a rate of 1.48*10^-8 per bp per generation (1.58-1.68 95% CI) that is in line with most exomic estimates. We believe that this is one of the largest sample sizes used in a direct estimate and the first to obtain a measurement in a non-European population.

354

Leveraging distant relatedness to quantify human mutation and gene conversion rates. *P. Palamara[1,2], L. Francioli[3], G. Genovese[2], P. Wilton[6], A. Gusev[1,2], H. Finucane[1,2], S. Sankararaman[2,4], S. Sunyaev[2,4], P. DeBakker[3], J. Wakeley[6], I. Pe'er[7], A. Price[1,2,5], The Genome of the Netherlands Consortium.* 1) Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 3) Department of Medical Genetics, University Medical Center Utrecht, Utrecht, Netherlands; 4) Department of Genetics, Harvard Medical School, Boston, MA, U.S.A; 5) Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, U.S.A; 6) Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, U.S.A; 7) Department of Computer Science, Columbia University, New York City, NY, U.S.A.

The rate at which human genomes mutate is a central biological parameter that has many implications for our ability to understand demographic and evolutionary phenomena. We present a method for inferring mutation and gene conversion rates using the number of sequence differences observed in identical-by-descent (IBD) segments together with a reconstructed model of recent population size history. This approach is robust to, and can quantify, the presence of substantial genotyping error, as validated in coalescent simulations. We applied the method to 498 trio-phased Dutch individuals from the Genome of the Netherlands (GoNL) project, sequenced at an average depth of 13×. We infer a point mutation rate of $1.66 \pm 0.04 \times 10^{-8}$ per base per generation, and a rate of $1.26 \pm 0.06 \times 10^{-9}$ for <20 bp indels. Our estimated average genome-wide mutation rate is higher than most pedigree-based estimates reported thus far, but lower than estimates obtained using substitution rates across primates. By quantifying how estimates vary as a function of allele frequency, we infer the probability that a site is involved in non-crossover gene conversion as $5.99 \pm 0.69 \times 10^{-6}$, consistent with recent reports. We find that recombination does not have observable mutagenic effects after gene conversion is accounted for, and that local gene conversion rates reflect recombination rates. We detect a strong enrichment for recent deleterious variation among mismatching variants found within IBD regions, and observe summary statistics of local IBD sharing to closely match previously proposed metrics of background selection, but find no significant effects of selection on our estimates of mutation rate. We detect no evidence for strong variation of mutation rates in a number of genomic annotations obtained from several recent studies.

355

Genetic diversity on the human X chromosome suggests there is no single pseudoautosomal boundary. *M. Wilson Sayres[1,2], D. Cotter[1], S. Brotman[1].* 1) School of Life Sciences, Arizona State University, Tempe, AZ; 2) Center for Evolution and Medicine, The Biodesign Institute, Arizona State University, Tempe, AZ.

Unlike the autosomes, recombination between the X chromosome and Y chromosome is thought be constrained to two small pseudoautosomal regions (PARs) at the tips of each sex chromosome. The PAR1 spans the first 2.6 Mb of the proximal arm of the human sex chromosomes and is conserved across most eutherian mammals. The PAR1 is separated from the nonPAR region on the Y chromosome by a Y-specific inversion that is hypothesized to suppress X-Y recombination. The much smaller PAR2, spanning the distal 320 kb of the long arm of each chromosome, was duplicated from the terminal end of the X to the terminal end of the Y in the common ancestor of humans. In addition to the PAR1 and PAR2, there is a human-specific X-transposed region (XTR) that is suspected to behave as a third pseudoautosomal region. Genetic diversity is expected to be higher in recombining regions than in nonrecombining regions. In this study we investigate patterns of genetic diversity across unrelated individuals and divergence between primates across all regions of the human X chromosome. We observe that genetic diversity in PAR1 is significantly greater than the non-PARs. However, we observe a gradual shift from higher to lower diversity across this region, not an abrupt shift at the proposed pseudoautosomal boundary, suggesting that non-homologous recombination is common on the human sex chromosomes and spans the pseudoautosomal boundary. In contrast, diversity in the PAR2 is not significantly elevated compared to the nonPAR, suggesting that recombination is not obligatory in the PAR2. Finally, diversity in the XTR is higher than both the surrounding nonPARs, and the PAR2, providing evidence that recombination may occur with some frequency between the X and Y in the XTR. Although they comprise only a small percentage of the genome, the PARs and the XTR house essential genes for both sexes and provide a unique opportunity to explore the dynamics of sex chromosome evolution.

356

Comparative Epigenomic Analysis of Regulatory Elements in Primate Stem Cells. *I. Narvaiza[1], C. Benner[1], M. Wang[1], M.C. Marchetto[1], M. Ku[1], T. Swigut[2], J. Wysocka[2], F.H. Gage[1,3].* 1) Laboratory of Genetics, The Salk Institute, La Jolla, CA; 2) School of Medicine, Stanford University, Stanford, CA; 3) Center for Academic Research and Training in Anthropogeny (CARTA), UC San Diego, La Jolla, CA.

We have previously shown that the comparison of gene expression in induced pluripotent stems (iPS) cells derived for human and non-human primates revealed differences in the control of mobile elements, which contribute to explain the higher levels of genome diversity in great apes. Here, we have further investigated the differences between humans and our closest living relatives by carrying out a comparative epigenomic study in human and chimpanzee iPS cells. For epigenomic profiling we analyzed genome-wide chromatin accessibility, and histone modifications associated to active and repressed regulatory elements by chromatin immunoprecipitation and sequencing (ChIP-seq). We found that gene-specific modulation of bivalency directly correlates with differences in gene expression between human and chimapanzee iPS cells. We also identified a number of divergent enhancers driven by differences in transcription factor binding motifs, and found a direct association between enhancer divergence and differences in target gene expression between human and chimpanzee iPS cells. Many regulatory elements overlap with transposable elements, however we observed that divergence in promoters is partially driven by different behavior of ancient transposons that were mobile before the human-chimpanzee split. These findings reveal a novel mechanism for epigenomic evolution in humans and chimpanzees, and demonstrate the value of primate iPS cells for comparative and evolution studies.

357

Transcriptome Diversity Associated with Ancestry and Diet in Ethnically Diverse East African Populations. *N.G. Crawford[1], Y. Ren[2], R.A. Rawlings-Goss[1,3], G.R. Grant[1], H. Hutton[1], M. Yeager[4,5], S. Chanock[4,5], A. Ranciaro[1], S. Thompson[1,6], J. Hirbo[1,7], W. Beggs[1], T. Nyambo[8], S. Omar[9], D. Meskel[10], G. Belay[10], C. Brown[1], H. Li[2], S.A. Tishkoff[1,11].* 1) Genetics, University of Pennsylvania, Philadelphia, PA; 2) Department of Biostatistics and Epidemiology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA; 3) National Science Foundation, Washington DC, USA; 4) Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), NIH, Bethesda, MD, USA; 5) Core Genotyping Facility, NCI-Frederick, Frederick, MD, USA; 6) Global Alliance for Chronic Diseases, London, UK; 7) Biological Sciences, Abbot Lab, Vanderbilt, Nashville, TN, USA; 8) Department of Biochemistry, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania; 9) Kenya Medical Research Institute, Center for Biotechnology Research and Development, 54840-00200 Nairobi, Kenya; 10) Department of Biology, Addis Ababa University, Addis Ababa, Ethiopia; 11) Department of Biology, University of Pennsylvania, Philadelphia, PA, USA.

African populations harbor the greatest levels of genetic diversity, have reduced linkage disequilibrium (LD), and contain the deepest divergence times among human populations. However, little is known about how this variation affects complex phenotypes such as gene expression. To investigate this question, we deep sequenced RNA obtained from the whole blood of 171 Africans from nine populations in East Africa. These populations represent a range of subsistence practices (hunter-gatherers, agriculturalists, pastoralists and agro-pastoralists) and linguistic diversity (Khoisan, Cushitic, Omotic, Nilo-Saharan, Semitic). Samples were sequenced to a mean of 50 million reads. In addition we also genotyped these samples on an Illumina 5.0 plus exome SNP array in order to characterize expression quantitative loci (eQTL) and allele specific expression (ASE). We applied surrogate variable adjustment to control for batch effects and other latent variables in the gene expression data. We find that hierarchical clustering and principle component analysis of differentially expressed genes recapitulate phylogenetic relationships inferred from genome-wide SNP array data. A pathway enrichment analysis of genes differentially expressed among populations clustered based on subsistence pattern revealed pathways associated with diet metabolism. We performed a meta analysis across populations to identify genetic variants associated with gene expression using joint eQTL and ASE models. The eQTL analysis both replicates many known associations identified in European whole blood studies and also identifies variants more common and/or private to African populations. Furthermore, we demonstrate that the lower LD in Africans improves the resolution of eQTLs previously identified in Europeans. This research represents the largest study to date of transcriptomic variation and eQTL analysis from whole blood among understudied and ethnically diverse Africans.

358

High-coverage RNA-sequencing Reveals Substantial Variation Associated with Geography, Environment and Endophenotypic Variation. *MJ. Fave[1], A.J. Hodgkinson[1,2], JP. Goulet[1], JC. Grenier[1], H. Gauvin[1], V. Bruat[1], T. de Maillard[1,3], E. Gbeha[1,5], E. Hip-Ki[1], Y. Idhagdour[4], P. Awadalla[1,3,5].* 1) Sainte-Justine University Hospital Research Centre, Department of Pediatrics, Faculty of Medicine, University of Montreal, Montréal, Québec, Canada; 2) Department of Medical and Molecular Genetics, Guy's Hospital, King's College London, U.K; 3) CARTaGENE, Sainte-Justine University Hospital, Montréal, Québec, Canada; 4) NYU Abu Dhabi, Abu Dhabi; 5) Ontario Institute of Cancer Research, University of Toronto, Toronto, Canada.

Phenotypic variation is the result of the combined effect of genetic variation with environmental influences. Gene-by-environment interactions are thought to be pervasive and may be responsible for a large fraction of the unexplained variance in heritability and disease risk. However, it has been particularly difficult to reliably identify robust gene-by-environment effects in humans. Studies mapping gene expression variation in humans have established that there is an abundant amount of inter-individual regulatory variation and that a significant fraction of it is heritable. Yet, a general understanding of the extent of variation of gene expression and how genetic regulatory variation is modulated by environmental factors is lacking. To systematically survey genetic, environmental and interaction effects on whole blood transcriptome, we combined RNASeq profiling with whole genome genotyping on 1,000 deeply endophenotyped individuals selected from over 40,000 participants in the CARTaGENE resource. Using haplotype-based methods on genome-wide genotyping, we detected fine-scale genetic structure within the province, and were able to identify within-province migrant individuals. We document substantial geographical variation in whole blood gene expression in this founder population that follows a south-north cline in the province of Quebec. In addition to the strong signature of geographic regional effects on gene expression, we reveal a substantial impact of environmental factors on global gene expression profiles overpowering that of the genotype. Expression profiles of migrants are more similar to those of individuals presently living in the same region than to those of individuals with the same ancestry but living in a different region. Genes involved in oxygen transport and inflammation are enriched among the differentially expressed genes between regions, suggesting an impact the highly urbanized environments on expression profiles. We also report several instances of genome-wide significant transcriptional gene-environment interactions (environmental eQTLs) that may have a clinical impact for individuals carrying specific genotypes in a given environment. These findings suggest that environmental variation can significantly alter disease genetic risk in both direct and indirect fashion and call for placing regulatory variants in the context of their geographical distribution and associated environmental exposures.

359

The importance of assaying the matched normal when sequencing cancer genomes. *E. Helman, M. Clark, R. Alla, D. Church, S. Boyle, A. Patwardhan, S. Luo, J. Harris, N. Leng, C. Haudenschild, R. Chen, J. West.* Personalis, Inc, Menlo Park, CA.

Targeted sequencing assays are increasingly used to identify tumor mutations that guide therapeutic decisions. Interpretation of a cancer variant's origin and therapeutic impact poses a set of new challenges. Recent studies have indicated that jointly analyzing a tumor with its matched normal can accurately discriminate between tumor-specific (somatic) and inherited (germline) mutations. However, procurement of a matched sample is often logistically impractical. In the absence of a matched normal, large databases and analytical techniques are currently used to identify cancer variants in tumor sequencing data. The necessity of the matched normal for accurate detection of cancer-relevant mutations remains an open question.To compare tumor-only and tumor/normal analysis of cancer samples, we collected a set of >100 formalin-fixed (FFPE) and fresh frozen cancer samples of various tumor types, where matched normal blood or adjacent tissue was available. We performed augmented target enrichment sequencing (exome and large cancer gene panel) of both DNA and RNA. The data was analyzed using cancer bioinformatics pipelines that detect base substitutions, small insertions/deletions, copy number alterations, and gene fusions in both tumor-only and tumor/normal modes. Variants were annotated using a described clinical actionability filtering strategy.We find that 47% of mutations detected in tumor-only mode are reclassified as germline variants when analyzed together with the matched normal sample. These include mutations in hereditary cancer predisposition genes, such as *BRCA1, RB1*, *MSH2*, and other genes with American College of Medical Genetics guidelines. Importantly, by applying our clinical filtering strategy, we discover a number of clinically actionable mutations that are present in the matched normal. Copy-number alterations are also refined in the tumor/normal analysis.The effects of administering targeted therapies to patients with germline mutations in the relevant gene are largely unknown. We suggest that mutations determined to be germline through matched normal sequencing represent variants that may be important for hereditary cancer knowledge and tumor treatment, and should be reported as such. For NGS-based cancer interpretation to guide clinical decisions in a practical and cost-effective manner, both tumor-only and tumor/normal analysis must be available until more is known about the effects of treatment based on somatic and germline variants.

360

Insights into somatic mutation-driven cancer genome evolution: A study of 3,000 cancer genomes across 9 cancer types. *Z. Zhao[1,2], F. Cheng[1].* 1) Department of Biomedical Informatics, Vanderbilt Univ, Nashville, TN; 2) Department of Cancer Biology, Vanderbilt Univ, Nashville, TN.

Genomic instability has been recognized as a hallmark of cancer for several decades. So far, there are few general mathematical models to quantitatively examine how perturbations of a single gene drive subsequent genetic changes. Massive amounts of genomic alteration data have been recently generated, but they present researchers with a dilemma: does this genomic instability contribute to cancer, or is it simply a byproduct of cellular processes gone awry? Thus, quantifying whether the perturbation of any single gene in a cancer genome is sufficient to shape adaptive cancer genome evolution would help us better understand the fitness of somatic cells through genomic alterations. In this study, we developed the gene gravity model derived from Newton's law of gravitation to study the evolution of cancer genomes by incorporating the transcriptional and somatic mutation profiles of ~3,000 tumors across 9 cancer types from The Cancer Genome Atlas into a broad gene network. This model postulates that if two genes had high mutation rates and strong gene co-expression in a given cancer type, they would have a higher gravitation score (*G*) and create a higher risk of inducing mutations to other genes. We found that mutations of a cancer driver gene tended to uniquely cause cancer genome instability and shape adaptive cancer genome evolution by inducing mutations in other genes. Importantly, this functional consequence is often generated by the combined effect of genetic and epigenetic (e.g., chromatin regulation) alterations. In addition to the above findings at the cellular level, we identified six new cancer genes (*AHNAK, COL11A1, DDX3X, FAT4, STAG2,* and *SYNE1*), each of which significantly increased cancer genome mutation rates. Finally, we provided statistical evidence that aneuploidy is a common genetic mark of cancer, due to a higher risk of genomic instability uniquely induced by cancer driver genes on the X chromosome in comparison to autosomal chromosomes. In summary, our results provided novel insights into the genomic instability that propels adaptive cancer genome evolution. Moreover, we first time provided statistical evidence using somatic mutations that aneuploidy is a common genetic factor of cancer. This work may help elucidate the functional consequences and evolutionary characteristics of somatic mutations during tumorigenesis towards driving adaptive cancer genome evolution.

361

Systematic analysis of mutation distribution in three dimensional protein structures identifies cancer driver genes. *A. Fujimoto[1], Y. Okada[1,2], K. Boroevich[1], T. Tsunoda[1], H. Taniguchi[1], H. Nakagawa[1].* 1) RIKEN Center for Integrative Medical Sciences; 2) Tokyo Medical and Dental University.

Protein tertiary structure determines molecular function, interaction, and stability of the protein, therefore distribution of mutation in the tertiary structure should provide us with biological insight into molecular function of the protein and can facilitate to identify new driver genes in cancer. To analyze mutation distribution in protein tertiary structures, we applied a novel three dimensional permutation test (3D permutation test) to the mutation positions. We analyzed somatic mutation datasets of 21 types of cancers obtained from exome sequencing conducted by the TCGA project. Of the 3,608 genes that had ≥3 mutations in the regions with tertiary structure data, 98 genes showed significant skew in mutation distribution after adjustment for multiple testing (FDR $q\text{-}value \leq 0.1$). Known tumor suppressors and oncogenes were significantly enriched in these identified cancer gene sets with significant 3D skew. Physical distances between mutations in known oncogenes were significantly smaller than those of tumor suppressors, indicating that mutations in oncogenes were more tightly clustered. Twenty-three genes, including *TP53, PIK3CA*, *PTEN*, *CDKN2A*, *KMT2C, DHX9* and *PARG,* were detected in multiple cancers. Candidate genes with significant 3D skew of the mutation positions included tumor suppressor genes, oncogenes, kinases, apoptosis related genes, a RNA splicing factor, a miRNA processing factor, an E3 ubiquitin ligase and transcription factors. Our study suggests that systematic analysis of mutation distribution in the tertiary protein structure can help identify cancer driver genes, and contribute to the functional interpretation of the role of the mutations.

362

Insights, mechansims and fundamental significance of copy-neutral loss of heterozygosity detected in oncology samples. *S. Schwartz[1], B. Williford[1], R. Burnside[1], I. Gadi[1], V. Jaswaney[1], A. Penton[1], K. Phillips[1], H. Risheg[2], J. Schleede[1], J. Tepperberg[1], P. Papenhausen[1].* 1) Cytogenetic Department, Laboratory Corporation of America, Research Triangle, NC; 2) Dynacare/Laboratory Corporation of America, Seattle, WA.

While chromosome microarray analysis (CMA) has been utilized mainly for the detection of gain or loss of genetic material in constitutional specimens, SNP-CMA can detect copy-neutral loss of heterozygosity (CN-LOH). Although unusual in constitutional studies, it is a more common phenomenon in neoplastic cells leading to duplication of a pathogenic mutation. To understand this alteration, we have studied over 5,000 patients with the Affymetrix® Cytoscan® HD microarray that permit detection of CN-LOH. The patients diagnoses include a variety of neoplastic disorders such as; ALL, AML, CLL, CML, Lymphoma, Myeloma, MPN and MDS. Overall, the presence of CN-LOH varied from 10.4% of patients with multiple myeloma, to as high as 42.3% in patients with MPN disorders. The studies have provided important insight, as well as, a deeper understanding with regard to CN-LOH: 1) Every chromosome arm was involved in CN-LOH with the exception of three (4p, 10p, 20p), suggesting the high prevalence of tumor suppressor genes; 2) Different chromosome arms were preferentially involved in different disorders (ALL – 9p, AML – 13; CLL – 13, 14, 20q, Myeloma - 16q, MPN -9p, MDS – 4q, 7q, 11q, 14); 3) However, 17p (resulting in a TP53 homozygous mutation) is involved in all of the disorders; 4) The occurrence of 2 or more CN-LOH regions varied from 0% with MPN to 44.4% with Lymphoma; 5) In the majority of the disorders the CN-LOH usually occurred with copy number gains and losses, but was seen as an isolated finding with 40% of AML cases, 57.5% of MDS and 72.7% of MPN; 6) A rare expansion of homozygosity that visibly demonstrated distinctly different frequencies in the same chromosome arm (e.g. CLL – 8.6% of patients), suggesting clonal evolution; 7) There appears to be an underlying mechanism leading to CN-LOH as for any chromosome arm, with the initiation site for the CN-LOH clustering at one location (even if the putative gene was at a more distal location); 8) The presence of CN-LOH ranged from ~9% to 100% of the sample; 9) The percent CN-LOH could be used to track the progression of clonal involvement and in some MDS could be detected before being definitively diagnosed pathologically; 10) The CN-LOH studies have provided additional mechanistic insight demonstrating a distinct correlation with the evolution to deletion homozygosity, denotation of transplant failure in HLA-identical sibling transplants and pseudo interstitial CN-LOH associated with complex rearrangements.

363

Genomic analysis reveals novel secondary drivers and progression pathways in skin basal cell carcinoma. *X. Bonilla[1], L. Parmentier[2], B. King[3], G. Kaya[4], H.J. Sharpe[5], T. McKee[6], V. Zoete[7], P.G. Ribaux[1], F.A Santoni[1], K. Popadin[1], M. Guipponi[1,8], M. Garieri[1], C. Verdan[6], K. Grosdemange[4], O. Sumara[9], M. Eilers[9], F.J. de Sauvage[5], I. Aifantis[3], O. Michielin[7], S.E. Antonarakis[1,8,10], S.I. Nikolaev[1,8].* 1) Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; 2) Department of Dermatology, Hôpital de Sierre, Sierre, Switzerland; 3) Department of Pathology, NYU School of Medicine, NYC, USA; 4) 4 Department of Dermatopathology, University Hospitals of Geneva, Switzerland; 5) Department of Molecular Oncology, Genentech Inc., CA, USA; 6) Department of Clinical Pathology, University Hospitals of Geneva, Switzerland; 7) Swiss Institute of Bioinformatics, Switzerland; 8) Department of Genetic Medicine and Laboratories, University Hospitals of Geneva, Switzerland; 9) Department of Biochemistry and Molecular Biology,University of Würzburg, Germany; 10) iGE3, Institute of Genetics and Genomics of Geneva, Switzerland.

Skin basal cell carcinoma (BCC) is the most common malignant neoplasm in humans. Most BCCs are caused by aberrant activation of the hedgehog pathway (SHH) by mutations in *PTCH1, SMO* or *SUFU*. BCC's varying morphology and aggressiveness, as well as the incomplete response to pharmacological treatment in 50% of cases, may be associated with the acquisition of secondary driver mutations. We sought to fully characterize the genomic landscape of BCCs in order to identify additional genes and pathways contributing to BCC development and progression.We performed exome-sequencing of fresh or frozen tumors from 119 cases/matching germline samples and targeted sequencing of a panel of cancer genes in a validation cohort of 141 FFPE samples. BCC is the tumor with most somatic mutations per Mb (higher than melanoma). Primary driver mutations or SCNAs in *PTCH1*, *SMO, SUFU* or *TP53* were identified in 96% of samples. We observed secondary driver mutations in 55% of cases. MutSigCV analysis detects *PTCH1* and *TP53* (q-value=0), *SMO* (0.037) and *PTPN14* (0.00095) as significantly mutated genes. Remarkably, we identified highly recurrent mutations in *MYCN* (26% of cases), *PPP6C* (15%), *FBXW7* (6.5%) and *STK19* (12%) as well as smaller fractions of samples harboring well known oncogenic mutations in *KRAS* and *NRAS* (2%), *ERBB2* (4%), and *RAC1* (2%). Validation experiments determined that *MYCN* mutations promote N-Myc stabilization by impairing its ubiquitination by FBXW7. Furthermore, we show SHH-independent activation of the HIPPO pathway in 30% of cases via truncating mutations in *PTPN14*, a tyrosine phosphatase involved in cell proliferation control through cytoplasmic sequestration of YAP1, a key transcription factor. Immunohistochemistry confirms the enrichment of nucleus-localized YAP1 in *PTPN14*-mutated tumors. Finally, in a subset of tumors we have identified somatic mutations in genes of the MAPK pathway.This study is the largest and most comprehensive analysis of BCC to date. We report *MYCN* oncogenic point mutations for the first time and we provide insight into signaling mechanisms that may act in combination with *PTCH1-SMO* to promote tumor progression via secondary oncogenic mutations downstream of *GLI* (*MYCN*, *FBXW7*) or known oncogenic mutations not previously described in BCC (*PPP6C, MAPK, RAC1*), as well as through mutations in SHH-independent tumor suppressor genes (*PTPN14*). These results reveal novel tumorigenic pathways and may offer additional therapeutic opportunities in BCC.

364

Recurrent Somatic Mutation in the MYC Associated Factor X in Brain Tumors. *H. Nikbakht[1,2], M. Montagne[3], N. Jabado[4], P. Lavigne[3], J. Majewski[1,2].* 1) Department of Human Genetics, McGill University, Montreal, Quebec, Canada; 2) Genome Quebec Innovation Center, Montreal, Quebec, Canada; 3) Department of Biochemistry, Sherbrooke University, Sherbrooke, Quebec, Canada; 4) Department of Pediatrics, McGill University, Montreal, Quebec, Canada.

The MYC proto-oncogene is a known key factor in the development of diverse cancers. MYC Associated Factor X (MAX) plays a central role in the MYC-MAX-MAD gene regulatory network; however, its direct involvement in cancer has not yet been reported.In this study we report discovery of a novel recurrent somatic mutation in MAX gene in brain tumors. Within a set of 172 high-grade astrocytoma (HGA) that were exome sequenced by our group, we identified 6 tumors (3.5%) with R51Q mutation. We further investigated the presence of MAX mutations in other tumors in public databases and identified 14 additional R51Q mutants. Our findings show that this mutation always appears later in tumor development in a sub-clonal fashion, and is always accompanied by at least one driver such as Histone 3 K27M or IDH1 R132H.Using biophysical assays we demonstrate that MAX R51Q mutation has a very specific effect on the binding efficacy between MAX and other proteins in its regulatory network, as well as to DNA. Our results show that Max R51Q binds less efficiently to DNA in a homodimer form compared to Max wild type. On the other hand, this mutation has an opposite effect when MAX heterodimerizes with Myc. These effects are more pronounced in non-specific DNA as compared to E-Box containing genes.To further understand the effects of this mutation on the progression and development of brain tumors, we explored a unique opportunity to study a patient with bilateral thalamic pediatric astrocytoma in which the primary tumors (left and right thalamus) had nearly identical mutation profiles except for the presence of the MAX R51Q exclusively in the right tumor. The MYC-MAX-MAD regulatory complex is known to affect acetylation of Histone 3 Lysine 27 (K27ac) and thus affect the transcription levels of the downstream genes. To study these effects, we used ChipSeq and RNASeq and investigated the redistribution of the K27ac as well as the differentially expressed genes between the MAX mutated and wild type tumors.We identify MAX as a new cancer gene, particularly relevant to brain cancer. Our results show the possible effects of MAX mutation in promoting tumor progression and development. It also suggests the effect of this mutation on the spread of the tumor. These findings shed new light on the mechanisms underlying cancer progression, spread, and the involvement of MYC signaling in the development of brain tumors which, in turn, can point us towards new targets for therapeutic approaches.

365

The driver landscape of parathyroid carcinoma. *C. Pandya[1], A.V. Uzilov[1], J. Bellizzi[2], S.D. Li[1], W. Yu[3,4,5], M. Stevenson[6], B. Cavaco[7], B.T. Teh[3,4,8], R.V. Thakker[6], H. Morreau[9], A. Arnold[2], R. Chen[1].* 1) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Center for Molecular Medicine, University of Connecticut School of Medicine, 263 Farmington Avenue, Farmington, CT; 3) Laboratory of Cancer Epigenome, Division of Medical Sciences, National Cancer Centre Singapore, Singapore; 4) Division of Cancer and Stem Cell Biology, Duke-National University of Singapore Graduate Medical School, Singapore; 5) National University of Singapore Graduate School for Integrative Sciences and Engineering, Singapore; 6) Academic Endocrine Unit, Nuffield Department of Clinical Medicine, Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford UK; 7) Molecular Endocrinology Group, Molecular Pathobiology Research Centre Unit of the Portuguese Institute of Oncology from Lisbon Francisco Gentil and Chronic Diseases Research Centre, Faculty of Medical Sciences, New University of Lisbon, Lisbon, Portugal; 8) Cancer Science Institute of Singapore, National University of Singapore, Singapore; 9) Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands.

Parathyroid carcinoma (PC) is a rare malignancy leading to severe hyperparathyroidism with intractable hypercalcemia and metabolic complications. Effective therapeutic intervention for disease that persists after surgical excision still eludes the clinic as the molecular basis of oncogenesis is poorly understood. We analyzed whole-exome sequencing (WES) data from 18 primary/metastatic parathyroid carcinomas and matched normals as WES provides an unbiased view of the driver landscape. Primary data from 10 novel cases fulfilling stringent selection criteria (clinically sporadic presentations and demonstrating local invasion of tumor into surrounding tissues and/or distant metastasis) was pooled with 8 other cases (Yu et al., J Clin Endocrinol Metab 2015, 100:E360). Somatic mutations, germline variations and copy number alterations were identified. Noting that biallelic inactivation is required for oncogenesis driven by a classic tumor suppressor gene, we show that 8/18 patients exhibit inactivation of CDC73 (only established tumor suppressor in PC), with the first inactivation in the germline in 3 cases. Notably, we have identified the first observed recurrent somatic mutation in PC - ADCK1 p.I482M. ADCK1, a kinase on which no experimental studies have been previously carried out, is expressed in parathyroid tissue and several head & neck cancers. Furthermore, we identified driver genes novel to PC: FAT3 (N=3) and TNRC6A (N=2). No additional mutations were discovered in PRUNE2 beyond the ones described earlier. Additionally, somatic mutations in PIK3CA, TGFBR2, MTOR and other cancer associated genes were observed. Based on our analyses, we can partition the cohort by cancer pathway dysregulation as: Wnt pathway, Hippo pathway, Notch/TGF-beta pathway and PI3K/AKT pathway. Observed copy number variations are similar to results described earlier for e.g. chr13 loss (N=8/18). Typically, G:C>A:T and A:T>G:C somatic transitions are observed in other cancers. However, we observe G:C>C:G (N=6) and G:C>T:A (N=3) transversions at elevated rates, indicating presence of mutator phenotypes causing uncommon mutation types. 5/18 tumors with high rates of somatic G:C>C:G transversion had biallelic somatic inactivation of CDC73. Intriguingly, we observe the absence of somatic mutations in most established cancer genes. Thus, our data suggest the driver landscape of parathyroid carcinoma is distinct from commonly sequenced cancer types and requires further investigation.

366

Epigenomic profiling of prostate cancer identifies differentially methylated genes in *TMPRSS2:ERG* fusion positive versus negative tumors. *M.S. Geybels[1,2], J.J. Alumkal[3], I.M. Shui[1], M. Bibikova[4], B. Klotzle[4], M. Rinckleb[5], A. Luedeke[5], C. Maier[5], E.A. Ostrander[6], J. Fan[4], Z. Feng[7], J.L. Stanford[1,8].* 1) Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA; 2) Department of Epidemiology, GROW School for Oncology and Developmental Biology, Maastricht University, Maastricht, the Netherlands; 3) Division of Hematology and Medical Oncology, Knight Cancer Institute, Oregon Health and Science University, Portland, OR; 4) Illumina, Inc., San Diego, CA; 5) Institute of Human Genetics and Department of Urology, Faculty of Medicine, University of Ulm, Ulm, Germany; 6) Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD; 7) MD Anderson Cancer Center, Houston, TX; 8) Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA.

The *TMPRSS2:ERG (T2E)* gene fusion is a common somatic alteration in prostate cancer and preliminary evidence suggests that *T2E* fusion positive tumors represent a biologically distinct subtype of prostate cancer at the epigenetic level. We investigated epigenome-wide DNA methylation profiles in relation to tumor *T2E*-status in a population-based cohort of 496 prostate cancer patients. Fluorescence in situ hybridization (FISH) 'break-apart' assays were used to determine *T2E* fusion status and 266 patients (53.6%) had fusion positive PCa. Differentially methylated CpGs were identified (False Discovery Rate Q-value <0.00001; n = 27,946), and the 25 top-ranked CpGs (Q-values ≤1.53E-29) had a mean methylation difference of at least 25% between patient groups. The 25 CpGs mapped to nine genes, eight of which also showed changes in mRNA expression in the same patients' tumor tissue: *CACNA1D*, *C3orf14*, *SEPT9*, *NT5C*, *GREM1*, *PDE4D*, *TRIB2*, and *KLK10*. The calcium-channel gene *CACNA1D* is a known *ERG*-target that showed gene body hypomethylation and mRNA overexpression in fusion positive prostate cancer. Some of the top-ranked CpGs were differentially methylated in tumors with higher versus lower Gleason scores in fusion negative prostate cancer, in particular a CpG island in the gene body of *PDE4D*. Analysis of The Cancer Genome Atlas (TCGA) dataset provided confirmatory evidence for our top findings. Results from this large epigenome-wide study suggest that *T2E* fusion positive and negative prostate cancer represent epigenetically distinct subgroups and differentially methylated genes between these subgroups may highlight specific therapeutic targets.

367

Large-scale Inference of Activating and Repressive Nucleotides in Human Cell Types Using Tiling Reporter Assays. *J. Ernst[1], T.S. Mikkelsen[2], M. Kellis[2,3].* 1) UCLA, Los Angeles, CA; 2) Broad Institute, Cambridge, MA; 3) MIT, Cambridge, MA.

Massively parallel reporter assays have enabled genome-scale validation experiments towards gaining a systems-level view of gene regulation. A series of studies have demonstrated their use for testing thousands of predicted enhancers, dissecting regulatory motifs within them, and testing synthetically-designed sequences. However, even with tens of thousands of sequences tested in a single assay, it has been impractical to dissect large numbers of regions at nucleotide level resolution, without an a priori knowledge of predicted regulatory motifs, limiting their large scale use to validation, but not discovery.Here, we overcome this limitation, and present a new Bayesian tiling inference approach, which combines experimental tiling of regulatory regions using 31 sequences of length 145bp at 5bp intervals covering 295bp in total with computational inference of the resulting signal to infer a nucleotide-level view of regulatory activity across thousands of regulatory regions. By exploiting the multiple overlapping sequences in a probabilistic framework, our method is also robust to noisy or missing measurements, and enables high resolution inferences with a very small number of tested sequences per target region. This enables the *de novo* discovery of individual binding sites, and inference of their activating or repressive action in a single experiment across thousands of candidate regions. In contrast, activating and repressive sites are generally not distinguishable in current DNase hypersensitivity footprinting assays, as they both show footprints.We apply this method to more than 15,000 regions in the human genome, in two ENCODE cell types, selected based on the presence of DNase hypersensitivity and chromatin marks covering a diverse range of regulatory regions, including enhancers, promoters, and insulator regions. Our method resulted in a regulatory activity score for more than 4.5 million nucleotides, which we used to predict bases of activation and repression. These nucleotides showed strong enrichments for motifs associated with activation or repression in the cell type.Our method enables an unbiased, *de novo*, and high-resolution view of regulatory bases, which complements current motif scanning and DNase hypersensitivity footprinting approaches, and provides the first nucleotide-level view of activating and repressive sites across a sizeable fraction of the regulatory human genome.

368

Full-length mRNA sequencing uncovers a widespread coupling between transcription and mRNA processing. *S.Y. Anvar[1,2], E. de Klerk[1], M. Vermaat[1,2], J.T. den Dunnen[1,2,3], S.W. Turner[4], P.A.C. 't Hoen[1].* 1) Human Genetics, Leiden University Medical Center, Leiden, Netherlands; 2) Leiden Genome Technology Center, Leiden University Medical Center, Leiden, Netherlands; 3) Clinical Genetic, Leiden University Medical Center, Leiden, Netherlands; 4) Pacific Biosciences, Menlo Park, CA 94025, USA.

High-throughput RNA sequencing helps deciphering the global landscape of RNA expression. However, a comprehensive survey of transcriptional and posttranscriptional events in the same mRNA molecule is compromised by the short read length of second-generation sequencing platforms. Here, we analyzed Pacific Biosciences long sequencing reads capturing full-length mRNA molecules in MCF-7 human breast cancer cells. We obtained 7.4 million single-molecule long sequencing reads representing full-length mRNA molecules. From the 14,385 genes with detectable expression, 49% produced multiple transcripts. A total of 93 candidate fusion genes were identified based on the inter-chromosomal or distant intra-chromosomal split-alignment of transcripts to the human reference genome. In addition, 42% of identified transcripts in MCF-7 are potentially novel in comparison with the GENCODE annotation.Our long read-based survey and quantification of transcripts demonstrates a striking degree of coordination between transcription initiation, splicing and polyadenylation. In nearly half of the genes the selections of alternative transcription start sites, alternative exons or alternative polyadenylation sites are interdependent. Notably these couplings can occur over large distances, and a particular selection of a transcription start site at the 5'-end of a transcript can influence the choice of the polyadenylation at the 3'-end. Interestingly, alternative polyadenylation sites that are coupled with alternative splicing events are depleted for known polyadenylation signals and enriched for binding motifs for RNA binding proteins from the muscle blind (MBNL) family. Our data suggest a coordinating role for MBNL proteins in the regulation of splicing and polyadenylation. Our findings demonstrate that our understanding of transcriptome complexity is far from complete. Full-length transcript sequencing provides excellent opportunities to study largely unresolved mechanisms that coordinate transcription and mRNA processing and the effect of genetic variants on transcript structure.

369

The Role of RNA Polymerase II Pausing in the Mediation of Human Gene Expression. *J. Boden[1], V.G Cheung[2,3,4].* 1) Cancer Biology Graduate Program, University of Michigan, Ann Arbor, MI; 2) Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI; 3) Life Sciences Institute, University of Michigan, Ann Arbor, MI; 4) Departments of Pediatrics and Human Genetics, University of Michigan, Ann Arbor, MI.

Pausing of RNA polymerase II (RNAPII) near gene promoters is a key regulatory step that was originally identified in *Drosophila* (Rougvie and Lis 1988); here we investigated promoter proximal pausing in human cells. Pausing occurs when RNAPII is held in a transcriptionally active state 20-60 nucleotides downstream of the transcription start site through its interaction with two pausing factors, negative elongation factor (NELF) and DRB sensitivity-inducing factor (DSIF). RNAPII pausing has also been shown to regulate the expression of *MYC*, *FOS* and *JUNB* (Krumm, A. et al. 1992, Plet, A. et al. 1995, Aida, M. et al. 2006). The goal of this study is to broaden our understanding of the role of RNAPII pausing on human gene expression genome-wide. To identify a set of paused genes, we performed precision run-on sequencing (PROSeq) (Kwak et al., 2013). From the 10,000 genes in this data set, we identified over 4,000 genes that are enriched for RNAPII in the promoter regions relative to the gene body as indicated by the "pausing index." Among these genes, the average pausing index is 54, the median pausing index is 19; with and the a range maximum of pausing indices from is 3,000 to 1. The genes in this dataset include *GSK3B, WNT2B* and *DVL2* which are involved in WNT signaling, *TGFB1*, *TGFB2* and *SMAD5* which are components of the TGFb signaling pathway and *EIF5*, *RPL9* and *EEF2* which are involved in translation. To confirm that these genes are regulated by pausing, we knocked down the pausing factor, NELF, and performed RNA sequencing. We looked for changes in gene expression presumably due to loss of RNAPII pausing. Even at one time-point following NELF knockdown, approximately 20% of the genes with a pausing index greater than 1 showed a 50% or more change in expression levels. Both the PROseq and gene expression data shown here were obtained in colorectal cancer cells. In this presentation, I will describe our findings in human cancer and normal cells by presenting data from PROSeq, NELF knockdown as well as RNAPII, DSIF and NELF chromatin immunoprecipitation to illustrate human gene regulation by proximal promoter pausing.

370

Reappraising the protein-coding potential of GENCODE using high stringency mass spectrometry. *J.M. Mudge[1], J. Wright[2], J. Choudhary[2], J. Harrow[1].* 1) Computational Genomics, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom; 2) Proteomics, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom.

While modern RNA sequencing methodologies allow us to appreciate the size of the human transcriptome, there is a limit to what they can tell us about its *functionality*. It is particularly important to establish which transcripts are translated. In 2014, Kim *et al*. and Wilhelm *et al*. published 'draft maps' of the human proteome based on mass spectrometry data, with both claiming the existence of hundreds of novel protein-coding regions. While these datasets will likely transform our understanding of the transcriptome, questions have been raised about their true false discovery rates. Here, we describe the intersection of this work with the ongoing GENCODE annotation project, which seeks to describe the full extent of transcriptional complexity (www.gencodegenes.org). We present a complete reanalysis of the human tissue data from these experiments, using multiple search algorithms combined with strict significance filtering. We use a radically different search space, including all GENCODE annotations alongside the UniProt database and thousands of RNAseq models and *ab initio* gene predictions. Starting with over 52 million spectra, we find 600 potential novel protein-coding regions, 71 of which are supported by multiple peptides. Each locus has been subjected to manual analysis by the HAVANA annotation group, allowing for a wide range of orthogonal data sources - including ribosome profiling - to be examined for additional supporting evidence. While we find a small number of truly novel protein-coding genes, we find no convincing support for the translation of lineage-specific lncRNAs based on current biological paradigms. Furthermore, over 90% of the loci identified are pseudogenes. As will be discussed, such results prompt a variety of conflicting biological and experimental interpretations. In conclusion, we will propose a set of comprehensive guidelines for the integrated analysis of proteomics and transcriptomics datasets in annotation projects, suitable for use in the wider community.

371

Post-translational mechanisms buffer protein abundance against transcriptional variation. *S.C. Munger[1], J.M. Chick[2], P. Simecek[1], E.L. Huttlin[2], K.B. Choi[1], D.M. Gatti[1], N. Raghupathy[1], K.L. Svenson[1], S.P. Gygi[2], G.A. Churchill[1].* 1) The Jackson Laboratory, Bar Harbor, ME; 2) Harvard Medical School, Cambridge, MA.

The Central Dogma provides a framework for understanding the regulation of protein expression. Transcriptional regulation of protein expression is driven by the availability of mRNA, which in turn is determined by rates of translation and degradation as well as sequestration and processing of mRNA. Under transcriptional regulation, protein abundance should be correlated with the abundance of its cognate mRNA. Post-translational mechanisms that modulate the rates of translation, protein stability or degradation can effectively uncouple protein from mRNA abundance. Recent studies have reported low correlation between protein and mRNA abundance, which has led to the prediction of widespread buffering of protein expression against variation in mRNA levels. However, the relative importance of transcriptional and post-transcriptional modes of protein regulation remains poorly understood. Our study builds on earlier observations, integrating new techniques in mass spectrometry to expand the breadth of protein quantification, new methods to accurately quantitate transcript abundance in RNA-seq data, and new mouse models with extensive genetic variation to perturb mRNA and protein expression. We identified 1728 proteins that are regulated by local genetic variation; for 80% of these proteins, the causal variant acts proximally on transcript abundance, and consequently there is high correlation between protein and transcript abundance. Further, we identified 1362 proteins that are regulated by distant genetic variation. In stark contrast to local associations, nearly all distant loci appear to act on target protein abundance independent of their cognate mRNA via unknown post-transcriptional mechanism(s) – aka protein buffering. We applied a novel mediation approach to identify causal regulatory proteins and transcripts underlying these distant loci. Our analysis revealed an extensive network of protein-protein interactions that act to achieve stoichiometric balance of functionally related enzymes and subunits of multimeric complexes, and moreover provides new insights into the specific mechanisms of protein buffering.

372

Convergence of genes and pathways influencing neurodevelopment following suppression of ASD-associated chromatin modifiers and transcriptional regulators in human neural progenitor cells. *S. Erdin[1,2], A. Sugathan[1,2,3], P. Manavalan[1], K.M. Hennig[1,3], S.D. Sheridan[1,3], C.M. Seabra[1,2], A. Stortchevoi[1], A. Ragavendran[1,2], M. Biagioli[1,3], S.J. Haggarty[1,3], J.F. Gusella[1,2,3,4], M.E. Talkowski[1,2,3].* 1) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; 3) Department of Neurology, Harvard Medical School, Boston, MA; 4) Department of Genetics, Harvard Medical School, Boston, MA.

A surprisingly high proportion of the genes that have been identified as highly penetrant risk factors for autism spectrum disorders (ASD) are regulators of transcription and chromatin remodeling, which can influence a broad spectrum of biological pathways. Here, we mimicked loss-of-function (LoF) mutations using RNAi and genome-editing methods to suppress a series of genes known to confer substantial risk of ASD when disrupted by *de novo* LoF mutations (*CHD8, TCF4, SATB2, AUTS2*) in an isogenic line of iPSC-derived neural progenitor cells (NPCs). Our initial analyses revealed that *CHD8* suppression resulted in alterations in gene expression or chromatin binding sites of a large number of genes and pathways associated with transcriptional regulation, neurodevelopmental functions, and ASD, and these altered genes were strongly correlated with co-expression modules that influenced developmental timing in the brain. When we integrated these data with the results from suppression of *TCF4, SATB2,* and *AUTS2* in the same NPC line, we find a strong degree in overlap between genes regulated by *CHD8* and genes regulated by *TCF4* in both up and down directions*,* and the networks of proteins in which they function. Statistically significant overlap was also detected to a lesser degree between *TCF4, CHD8,* and *SATB2*, but not *AUTS2* above chance expectations based on permutation. The transcriptional consequences of reduced expression of these regulatory genes converged most strongly on two subnetworks associated with cell adhesion and Wnt/Notch signaling networks, both of which were significantly enriched for ASD-associated genes defined by LoF mutations from previous exome sequencing studies and were associated with distinct expression profiles during developmental timing from BrainSpan data. These results indicate that highly penetrant LoF mutations associated with severe forms of ASD and other neurodevelopmental disorders such as Pitt-Hopkins syndrome converge on a smaller number of shared pathways that have substantial impact on genes that are expressed early in neurodevelopment. Independent replication of these findings, and CRISPR/Cas9 mutagenesis of additional genes in these subnetworks is ongoing.

373

High-throughput analysis of gene-environment interactions across 250 cellular conditions. *F. Luca[1], G. Moyerbrailean[1], O. Davis[1], C. Harvey[1], A. Alazizi[1], D. Watza[1], X. Wen[2], R. Pique-Regi[1].* 1) Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI; 2) Department of Biostatistics, University of Michigan, Ann Arbor, MI.

Adaptations to local environments have played major roles in shaping allele frequency distributions in human populations. Yet, a mismatch between genotype and environment may be responsible for higher disease risk. Recent studies have shown that GxE interactions can be detected when studying molecular phenotypes that are relevant for complex traits (e.g. infection response eQTLs in immune cells). Despite these relevant examples, the extent to which the environment can modulate genetic effects on quantitative phenotypes is still to be defined. Here we have developed a high-throughput approach to achieve a comprehensive characterization of GxE interactions in humans. To this end, we have investigated the transcriptional response to 50 treatments in 5 different cell types (for a total of 250 cellular environments and 3 individuals per cell type). Across 56 cellular environments (cell type/treatment with large changes in gene expression) we discovered 9548 instances of ASE (FDR<10%), corresponding to 8923 unique ASE genes. We found that in an individual sample, on average, 0.5% of genes with heterozygous SNPs are ASE genes. For a given gene, the probability of ASE is negatively correlated with average expression, with a 4.2 fold decrease per 10x increase in FPKMs. On the other hand, we find a 1.3 fold increase in probability of ASE per 2x change in expression in response to treatments. These results suggest that phenotypic variation in expression levels for genes highly expressed (housekeeping genes) is constrained to be small. Treatment conditions, instead, may unveil the functional role of regulatory variants advantageous in specific environmental conditions. Of genes regulated through GxE interactions (conditional-ASE), we observe that the majority of ASE is consistent across conditions, consistent with previous condition-specific eQTL analyses. Overall, we find 248 loci with evidence for GxE interaction, 120 with control-only ASE and 128 with treatment-only ASE genes. We observe also a trend for increasing evidence of conditional-ASE for genes with larger differential expression. Our characterization of ASE across cell types and environmental exposures will contribute to the understanding of how GxE interactions have shaped human phenotypes in different environments, and how adaptation to these environments has contributed to variation in complex traits.

374

Functional dissection of *BCL11A* enhancer by Cas9-mediated *in situ* saturation mutagenesis. *M.C. Canver[1], E.C. Smith[1], F. Sher[1], L. Pinello[2], N.E. Sanjana[3], O. Shalem[3], D.D. Chen[1], P.G. Schupp[1], D.S. Vinjamur[1], S. Garcia[2], S. Luc[1], Y. Fujiwara[4], T. Maeda[5], G.C. Yuan[2], F. Zhang[3], S.H. Orkin[1,4], G. Lettre[6], D.E. Bauer[1].* 1) Division of Hematology/Oncology, Boston Children's Hospital, Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Stem Cell Institute, Department of Pediatrics, Harvard Medical School, Boston, MA; 2) Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA; 3) Broad Institute of MIT and Harvard, McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences and Department of Biological Engineering, MIT, Cambridge, MA; 4) Howard Hughes Medical Institute, Boston, MA; 5) Division of Hematology, Department of Medicine, Brigham and Women's Hospital, Boston, MA; 6) Universite de Montreal, Montreal Heart Institute, Montreal, Quebec H1T 1C8, Canada.

Enhancers, critical determinants of cellular identity, are commonly identified by correlative chromatin marks and gain-of-function potential, though only loss-of-function studies can demonstrate their requirement for gene expression in the native genomic context. Previously we identified an erythroid enhancer of *BCL11A*, subject to common genetic variation associated with fetal hemoglobin (HbF) level, whose mouse ortholog is necessary for erythroid BCL11A expression. Here we develop pooled CRISPR-Cas9 guide RNA libraries to perform *in situ* saturation mutagenesis of the human and mouse enhancers. This approach reveals critical minimal features and discrete vulnerabilities of these enhancers. Despite conserved function of the composite enhancers, their architecture diverges. Only the human sequences score as a "super-enhancer". The human enhancer is composed of three DNase I hypersensitive sites (DHSs). The common SNPs associated with HbF level localize to two of the DHSs, each of which has a modest requirement for gene expression. The essential human sequences, which appear primate-specific, are found in the third DHS and are not subject to common genetic variation. We sequenced this enhancer in 1,366 African Americans with sickle cell disease and identified 7 rare variants (minor allele frequency 0.03-0.5%). Burden analysis demonstrates that rare variants at this DHS are associated with HbF level (P=0.002, +2.8% HbF per rare allele), independent of the known effects of the common haplotypes. We show that a single cleavage mediated by CRISPR-Cas9 followed by indel production at critical enhancer sequences is sufficient to cause robust disruption of BCL11A and de-repression of HbF in primary human erythroid precursors. These results highlight an example in which common trait-associated genetic variation provides merely a minimal estimate of the biologic significance of the underlying non-coding element. Moreover these studies validate the *BCL11A* erythroid enhancer as a target for therapeutic genome editing for the β-hemoglobin disorders.

375

***NGLY1* disease causes mitochondrial respiratory chain dysfunction and induction of oxidative stress.** *J. Kong[1,2], M. Peng[2], E. Nakamaru-Ogiso[3], M. He[1], J. Ostrovsky[2], Y-.J. Kwon[2], E. McCormick[2], T. Suzuki[4], Y. Argon[1], M.J. Falk[2].* 1) Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA; 2) Division of Human Genetics, Department of Pediatrics, The Children's Hospital of Philadelphia and University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 3) Department of Biochemistry and Biophysics, University of Pennslyvania Perelman School of Medicine, Philadelphia, PA; 4) The Institute of Physical and Chemical Research, Riken, Saitama, Japan.

BACKGROUND: Mitochondrial respiratory chain (RC) diseases and some Congenital Disorders of Glycosylation (CDG), such as *NGLY1* disease, are multi-system disorders with overlapping clinical features. Interestingly, *NGLY1* is among the most dysregulated genes in human mitochondrial diseases (Zhang Z, Falk MJ, *IJBCB*, 2014). Conversely, diagnostic tissue biopsy studies in two unrelated *NGLY1* patients evaluated in the CHOP Mitochondrial-Genetics Diagnostic Clinic revealed dysregulated mitochondrial mass and/or mitochondrial DNA content. While the *NGLY1*-encoded cytosolic N-glycanase has known functions in the endoplasmic reticulum (ER)-associated degradation (ERAD) or folding (ERAF) pathways as well as in the autophagic/lysosomal proteolytic pathways, it has not previously been linked to mitochondrial function. **METHODS:** In the present study, we systemically characterized the effects of *NGLY1* deficiency on mitochondrial function in *C. elegans*, mouse embryonic fibroblasts (MEFs), and patient fibroblast cell lines. **RESULTS:** An *NGLY1*-/- (knock-out) *C. elegans* worm strain (RB1452) was short-lived compared to wild-type (WT, N2 Bristol) at 20°C (median lifespan 11 versus 15 days, $p < 0.0001$). Fluorescence-based imaging of mutant worm physiology showed 25% reduced mitochondrial mass, 23% increased matrix superoxide burden, and 24% reduced mitochondrial membrane potential compared to WT animals. Similarly, FACS analysis of *NGLY1*-/- MEFs showed 40% increase in matrix oxidant burden when normalized to mitochondrial mass, and 48% relative reduction mitochondrial membrane potential. Oxygen consumption was reduced by 20% in both *NGLY1*-/- MEFs and NGLY1 patient fibroblasts relative to controls. When *NGLY1* fibroblasts were grown in galactose (stressing OXPHOS), there was a 9% decrease in mitochondrial membrane potential, 65% increase in matrix oxidant burden, and 40% decrease in cell viability. Transcriptional dysregulation of components in both mitochondrial and ER unfolded protein response (UPR), and in autophagy/lysosomal degradation pathways, was evident in both *NGLY1* patient cells and *NGLY1*-/- MEFs.**CONCLUSIONS:** *NGLY1* deficiency significantly reduces mitochondrial mass and respiratory function while increasing oxidant stress, lysosomal stress, and both the ER and mitochondrial stress responses. These data across three species show that *NGLY1* dysfunction directly impairs mitochondrial mass and function, while significantly inducing both oxidative and ER stress.

376

Mutations in pyruvate dehydrogenase phosphatase regulatory subunit (*PDPR*) are a novel cause of fatal neonatal cardio-encephalopathy with corneal clouding and lactic acidosis. *J. Christodoulou[1,2,3], M. Nafisinia[1,2], J. Crawford[4], R.M. Brown[5], G.K. Brown[5], M.H. Menezes[1,2], L.G. Riley[1,2], W.A. Gold[1,2], R.J. Taft[4,6,7], C. Simons[4].* 1) Genetic Metabolic Disorders Research Unit, Western Sydney Genetics Program, the Children's Hospital at Westmead, Sydney, NSW, Australia; 2) Discipline of Paediatrics & Child Health, Sydney Medical School, University of Sydney, Sydney, NSW, Australia; 3) Discipline of Genetic Medicine, Sydney Medical School, University of Sydney, Sydney, NSW, Australia; 4) Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia; 5) Oxford Medical Genetics Laboratories, The Churchill Hospital, Oxford, OX3 7LE, UK; 6) Illumina, Inc., San Diego, CA 92122, USA; 7) Departments of Integrated Systems Biology and of Pediatrics, George Washington University, Washington, DC 20052, USA.

Background: Pyruvate dehydrogenase (PDH) phosphatase regulatory subunit (PDPR) is a FAD-containing subunit of the PDH phosphatase (PDP), regulating the oxidative decarboxylation of pyruvate to form acetyl-CoA.**Aim:** To identify the genetic basis of a novel clinical entity with a unique clinical profile and a fatal course in the neonatal period.**Patients & Methods:** Family 1: identical twins born to non-consanguineous Australian parents had a very similar clinical course, with neonatal onset lactic acidosis (LA), epileptic encephalopathy, bilateral corneal clouding, and hypertrophic cardiomyopathy, which proved fatal in the first week of life. Family 2: the first child of non-consanguineous parents of Indian and Indian/Fijian background had neonatal onset seizures, LA, cardiomyopathy and bilateral corneal clouding. Whole exome sequencing (WES) for one of the probands in family 1 and Sanger sequencing of all 19 *PDPR* exons for the proband from family 2 were performed. Western blot analysis of PDPR, PDH, PDH E1-alpha subunit (phospho-S293) levels, and measurement of PDH enzyme activity in patient fibroblasts following lentiviral transduction of wild-type *PDPR*, was performed to confirm pathogenicity.**Results:** WES uncovered compound heterozygous mutations in *PDPR* (c.1343G>A, p. Arg448His and c.2401A>G, p.Ser801Gly) in the affected twins from family one. We identified a homozygous mutation (c.1986T>A; p. Asn662Lys) (heterozygous in each parent) and a heterozygous mutation (c.1279C>T; p.Arg427Cys) (heterozygous in mother) in the proband from family 2. Protein levels of PDPR, PDH and PDH E1-alpha subunit were significantly reduced by 45%, 70% and 90% respectively in patient fibroblasts from family 1 and 85%, 75% and 40% respectively in patient fibroblasts from family 2. Exogenous wildtype expression of PDPR in patient fibroblasts rescued protein levels of PDH by 90%, and 50% and PDH E1-alpha subunit by 80% and 95% in the patients from the family 1 and 2 respectively. Similarly, PDH enzyme activity levels were restored by exogenous expression of PDPR in patient fibroblasts. More recently two additional unrelated cases have been identified that are currently undergoing further functional characterization.**Conclusion:** Our discovery of this novel Mendelian disorder with a very specific phenotype leads us to recommend that patients with neonatal encephalopathy, cardiomyopathy, corneal clouding and lactic acidosis should be screened for mutations in *PDPR.*.

377

Signal transducer and activator of transcription 2 (STAT2) deficiency is a novel disorder of mitochondrial fission. *R. Shahni[1], C.M. Cale[2], G. Anderson[3], L.D. Osellame[4], S. Hambleton[5], T.S. Jacques[3,6], Y. Wedatilake[1], J.W. Taaman[7], E. Chan[2], W. Qasim[2], V. Plagnol[8], A. Chalasani[9], M.R. Duchen[10], K.C. Gilmour[2], S. Rahman[1,11].* 1) Genetics and Genomic Medicine, UCL Institute of child health, London, UK; 2) Molecular Immunology Unit, Great Ormond Street Hospital, London, UK; 3) Histopathology Unit, Great Ormond Street Hospital, London, UK; 4) Department of Biochemistry and Molecular Biology, Monash University, Melbourne 3800, Australia; 5) Primary Immunodeficiency Group, Institute of Cellular Medicine, Newcastle University, UK; 6) Developmental Neurosciences, UCL Institute of Child Health, London, UK; 7) Department of Clinical Neurosciences, UCL Institute of Neurology, Rowland Hill Street, London, UK; 8) UCL Genetics Institute, London, UK; 9) Neurometabolic Unit, National Hospital for Neurology and Neurosurgery, London, UK; 10) Cell and Developmental Biology, University College London, UK; 11) Metabolic Unit, Great Ormond Street Hospital, London, UK.

Background: Disorders related to mitochondrial dysfunction are highly heterogeneous owing to involvement of nuclear DNA as well as mitochondrial DNA. Mitochondria are dynamic organelles and constantly join and divide by the processes of fusion and fission respectively, forming tubular networks in order to maintain their function. Proteins involved in mitochondrial fusion and fission include mitofusins, optic atrophy 1 (OPA1), dynamin-related protein 1 (DRP1), mitochondrial fission protein 1 and mitochondrial fission factor (MFF). Unregulated mitochondrial fission may occur in primary mitochondrial diseases (*DRP1* and *MFF* mutations) and neurodegenerative disorders (Alzheimer, Huntington and Parkinson diseases), but the exact cellular mechanisms remain unknown.**Methods:** Whole exome sequencing and an in-house analysis bioinformatics pipeline was used to identify the underlying defect in two siblings presenting severe neurological deterioration and long mitochondria following viral infection. Following mutation identification, transcript and protein levels of fission and fusion factors were analysed, and lentiviral transduction of the candidate gene was performed in patient and control cells, as well as candidate gene knockdown in SHSY-5Y cells, in order to unravel the cellular effects of the pathogenic mutation.**Results:** Both patients shared a novel homozygous stop mutation (c.1836C>A p.Cys612Ter) in *STAT2* (a gene involved in innate immunity), which segregated within the family. Both siblings, and a third STAT2 deficient patient, shared a cellular phenotype characterised by DRP1 inactivation leading to impaired mitochondrial fission. *STAT2* knockdown in SHSY5Y cells using siRNA recapitulated the fission defect, which was rescued in all three patient fibroblasts by lentiviral transduction with wild-type *STAT2*.**Conclusion:** Our findings reveal important interactions between innate immunity and mitochondrial dynamics and demonstrate that STAT2 is a novel regulator of mitochondrial fission. We propose that modulation of JAK-STAT activity may represent a novel therapeutic avenue for mitochondrial diseases, the vast majority of which remain incurable.

378

Smith-Lemli-Opitz Syndrome iPS cells demonstrate abnormal neuronal differentiation due to 7-dehydrocholesterol impairment of Wnt/β-Catenin signaling. *F.D. Porter[1], A.N. Ton[1], C.A. Wassif[1], Y. Xin[2], P.E. O'Halloran[3], N. Malik[1], C. Cluzeau[1], I.M. Williams[1], N.S. Trivedi[3], W.J. Pavan[3], W. Cho[2], H. Westphal[1], K.R. Francis[1].* 1) NICHD, NIH, Bethesda, MD; 2) Department of Chemistry, University of Illinois at Chicago, Chicago, Il; 3) NHGRI, NIH, Bethesda, MD.

Smith-Lemli-Opitz syndrome (SLO) is an inborn error of cholesterol synthesis due to mutation of *DHCR7*. DHCR7 reduces 7-dehydrocholesterol (7DHC) to cholesterol leading to increased 7DHC and decreased cholesterol levels. The SLO phenotype includes congenital malformation and cognitive impairments. The pathological mechanisms underlying the SLO phenotype are not well delineated. In particular, the toxic effects of 7DHC versus decreased cholesterol in the pathology of SLO have not been fully characterized. To gain insight into mechanisms underlying SLO pathology we derived induced pluripotent stem cells (iPSC) from 5 SLO subjects, and, as controls from unaffected and lathosterolosis subjects. Specific mutations, pluripotency, and sterol biochemistry were confirmed in all cell lines. When cultured in cholesterol containing medium, SLO iPSC were biochemically and phenotypically comparable to control iPSC. However, when cultured in cholesterol depleted medium, SLO iPSC accumulated 7DHC and underwent accelerated neuronal differentiation. The accelerated neuronal differentiation was confirmed by treatment of control iPSC with AY9944, an inhibitor of DHCR7, and by CRISPR/Cas9 induced mutation of *DHCR7* in a control iPSC line. The abnormal neuronal differentiation phenotype was not observed in lathosterolosis iPSC which accumulate lathosterol rather than 7DHC, nor in control iPSC treated with U18666A, an inhibitor of DHCR24 which results in accumulation of desmosterol. These data demonstrate that a specific toxic effect of 7DHC rather than decreased cholesterol underlies the abnormal neuronal differentiation phenotype. Whole genome transcript analysis was performed to gain insight into the underlying mechanism, and we identified altered gene expression indicative of a defect in Wnt/β-catenin signaling. A defect in Wnt/β-catenin signaling was confirmed by demonstrating that restoration of Wnt/β-catenin signaling ameliorated the abnormal neuronal differentiation phenotype. Furthermore, surface plasmon resonance analysis combined with single molecule imaging demonstrated disruption of the Wnt receptor complex by 7DHC. Consistent with the *in vitro* iPSC studies, *Dhcr7*-/- mice demonstrate a defect in cortical cellular proliferation and decreased β-catenin activity. These data clearly implicate 7DHC as a toxic metabolite in SLO and suggest that pharmacological modulation of Wnt/β-catenin signaling could be explored as a novel therapeutic approach.

379

A mouse model of *cblC* deficiency displays reduced survival, growth retardation and combined methylmalonic acidemia and hyperhomocysteinemia. *M. Arnold[1], J. Sloan[1], N. Achilly[1], G. Elliot[2], J. Fraser[1,3], B. Brooks[4], C. Venditti[1].* 1) Organic Acid Research Section, Genetics and Molecular Biology Branch, NHGRI, NIH, Bethesda, MD; 2) Mouse and ES Core Facility, NHGRI, NIH, Bethesda, MD; 3) Medical Genetics Training Program, NHGRI, NIH, Bethesda, MD; 4) Unit on Pediatric, Developmental and Genetic Ophthalmology, NEI, NIH, Bethesda, MD.

Cobalamin C deficiency (*cblC*) is stated to be the most common inborn error of intracellular cobalamin metabolism and is caused by mutations in *MMACHC*, a gene responsible for the processing and trafficking of intracellular cobalamin. Mutations in *MMACHC* impair the activity of two cobalamin-dependent enzymes: methylmalonyl-CoA mutase (MUT) and methionine synthase (MTR). Patients display methylmalonic acidemia, hyperhomocysteinemia, hypomethionemia and variably manifest intrauterine growth retardation, anemia, heart defects, failure to thrive, white matter disease, neuropathy, neurocognitive impairment, and a progressive maculopathy and pigmentary retinopathy that causes blindness. Efforts to develop a viable animal model have proven unsuccessful to date: a knockout mouse generated from a *Mmachc* gene trap resulted in embryonic arrest by day E3.5 (Moreno-Garcia 2014). To create a viable animal model of *cblC* deficiency, we used TALENs to edit exon 2 of *Mmachc*, near the location of the common mutation seen in humans - c.271dupA p.R91KfsX14. 11 founder mice harboring 10 different alleles were generated. Two mutations were further investigated: an early frameshift null allele [c.165_166delAC p.P56CfsX4 (Δ2)] and another [c.162_164delCAC p.S54_T55delinsR (Δ3)] that results in a deletion-insertion predicted to produce an intact but mutant enzyme. At birth, an expected 1:2:1 Mendelian segregation was observed for the *Mmachc*$^{\Delta 3}$ allele (n=30 litters, 218 mice, χ^2 p>0.1) but not for *Mmachc*$^{\Delta 2}$ (n=19 litters, 134 mice, χ^2 p<0.001), in which the proportion of *Mmachc*$^{\Delta 2/\Delta 2}$ mice was decreased, suggesting partial embryonic lethality caused by this mutation. *Mmachc*$^{\Delta 2/\Delta 2}$ and *Mmachc*$^{\Delta 3/\Delta 3}$ mice were growth retarded, hypopigmented, and displayed decreased survival with 100% lethality by 32 days. Compared with wild type controls (n=7), *Mmachc*$^{\Delta 2/\Delta 2}$ (n=4) and *Mmachc*$^{\Delta 3/\Delta 3}$ (n=6) mutants displayed significantly elevated plasma methylmalonic acid, homocysteine, cystathionine and decreased methionine (p<0.05-0.001 for all metabolites). These new mouse models represent the first viable mammalian models of *cblC* deficiency and recapitulate the phenotypic and biochemical features of the disorder. Furthermore, the pigmentation abnormalities, seen in both mouse and zebrafish *cblC* models, suggest that neural crest and/or melanosomal defects may possibly contribute to *cblC*-disease related pathophysiology and correlations in humans should be explored.

380

Defects in SLC33A1 impair copper ATPase trafficking and contribute to the clinical and biochemical phenotypes of Huppke-Brendel syndrome. *L. Yi[1], W.H. Tan[2], P. Huppke[3], S.G. Kaler[1].* 1) Section on Translational Neuroscience, Molecular Medicine Program, Eunice Kennedy Shriver National Institute of Child Health and Human Development; 2) Division of Genetics and Genomics Manton Center for Orphan Disease Research, Boston Children's Hospital; 3) Department of Pediatrics and Pediatric Neurology, Georg August University, Göttingen. Germany.

Mutations in the acetyl-coA transporter SLC33A1 cause a complex autosomal recessive phenotype known as Huppke-Brendel syndrome that features congenital cataracts, hearing loss, profound neurodevelopmental delay, and death during infancy or early childhood (Huppke et al., *Am J Hum Genet* 2012; 90:61-8. The presence of low serum copper and ceruloplasmin and cerebellar atrophy similar to Menkes disease in affected patients implied possible effects on ATP7A, the copper-transporting ATPase implicated in that disorder. SLC33A1 normally mediates N-terminal acetylation, a reversible post-translational modification of numerous proteins, including at least six other ATPases (Choudhary et al. *Science* 2009; 325(5942):834-40). We employed tandem mass spectroscopy to document acetylation of ATP7A. We then used CRISPR to knock out *SCL33A1* in HEK293T cells and studied ATP7A trafficking in response to copper stimulation after overexpression of a Venus-tagged ATP7A construct. In contrast to normal HEK293T cells, ATP7A failed to traffick from the *trans*-Golgi compartment to the plasma membrane in *SLC33A1* knockout cells upon high copper loading. Fibroblasts available from two affected patients were then studied directly and both showed a partial defect in endogenous ATP7A trafficking in response to copper. These findings 1) help to explain the clinical and biochemical phenotype of this recently identified condition, 2) reveal new information about the mechanism of ATP7A trafficking, and 3) suggest potential therapeutic approaches to this disorder. .

381

Inhibition of CTR1 by antisense oligonucleotides in mouse model of Wilson's disease reduces copper accumulation and improves liver pathology. *T.R Grossman, R.B Johnson, G. Hung, B.P Monia, M. McCaleb.* Antisense Drug Discoverey, ISIS PHARMACEUTICALS, Carlsbad, CA.

Wilson's disease (WD) is an autosomal recessive genetic disorder resulting from a loss of function mutation to the ATP7B gene, causing excessive copper accumulation in the liver, brain, kidney, and cornea. Clinical manifestations include hepatic disease ranging from mild hepatitis to acute liver failure or cirrhosis and neurological symptoms. WD is fatal if untreated and early initiation of therapy with copper chelators and / or zinc salts is essential for favorable outcome. However, current treatments may cause severe side effects such as neurological deterioration, hypersensitivity syndrome and bone marrow depression.Here, we developed a specific and highly efficacious 2nd-generation antisense oligonucleotides (ASO) targeting mouse high affinity copper transporter 1 (CTR1). CTR1 plays a critical role in systemic copper homeostasis, participating in hepatocyte copper uptake from blood, release of copper from enterocytes into the blood after absorption from the diet, and copper re-uptake in the kidney. Therefore, ASO mediated systemic inhibition of CTR1 could potentially provide great therapeutic benefit in WD. Treatment of toxic milk mice , a mouse model for WD, with CTR1 ASO for 12 weeks resulted in a significant reduction in CTR1 mRNA levels in the liver, kidney, and small intestine (83%, 50% and 80% knockdown compared to saline controls, respectively). While toxic milk mice exhibited greatly increased hepatic copper (400-600 PPM), treatment with CTR1 ASO began to show significant reductions in liver copper content by week 4 of treatment, and reached a maximal reduction in liver copper at week 6 (~86 PPM). Moreover, animals treated with CTR1 ASO showed an improved liver pathology and reduced copper accumulation in the brain. A study to evaluate the levels of copper in urine and feces of toxic milk mice revealed that copper levels are significantly increased (30-fold) in feces of mice treated with CTR1 ASO for 6 weeks. Our results suggest that the reduction in CTR1 expression in the intestine and liver leads to reduced copper absorption from the diet and reduced uptake of copper into the liver, resulting in improved liver pathology and reduced accumulation of copper in the brain. We propose that CTR1 ASO could be considered as a novel therapeutic strategy for Wilson's disease patients.

382

B4GALNT1 deficiency as a Cause of Hereditary Complex Movement Disorder with Parkinsonism Features: a New Inborn Error of Metabolism affecting Glycosphingolipid Biosynthesis. *C. Lourenco, M. Almeida, C. Leprevost, Y. Anikster, W. Marques jr.* Neurology, Univ Sao Paulo, Ribeirao Preto, SAO PAULO, Brazil.

Introduction: Hereditary spastic paraplegias (HSPs) comprise a complex and heterogeneous group of neurological disorders. Although majority of cases of HSPs are due to genes involved in axonal growth or vesicular trafficking, there is an overlooked group of HSPs that can be caused by inborn errors of metabolism (IEMs) . Adrenomyeloneuropathy, late-onset biotinidase deficiency, cerebrotendineous xanthomatosis are among relatively known metabolic causes of HSPs. **Objective:** To report a new hereditary metabolic cause of HSP in a Brazilian family caused by enzyme deficiency of beta-1,4-N-acetyl-galactosaminyl transferase 1 (B4GALNT1), involved in ganglioside biosynthesis. **Methodology:** After excluding the traditional IEMs associated with HSPs and molecular analysis of SPG11 and SPG15 genes, whole exome sequencing (WES) effort was performed . All sequencing results were imported and analyzed by the GEnomes Management Application (GEM.app). **Results:** Mutations in the B4GALNT1 gene (in the SPG26 locus) were identified in all affected patients in the family; parents were heterozygous carriers of the mutation. Patients affected by this disease have early onset spastic paresis, mild intellectual disability, cerebellar ataxia, strabismus and some can develop psychiatric disturbance. Male hypogonadism was also noticed. Brain MRI showed non especif white matter changes in older patients. **Conclusions:** Although there are many IEMs involved in ganglioside catabolism presenting as neurodegenerative disorders, this enzyme deficiency is the second human disorder identified in the pathway of ganglioside biosynthesis, suggesting that other human diseases can be caused by metabolic errors in this biochemical pathway.

383

74 SNPs associated with education provide insights into brain function and disorders. *J.J. Lee[1], T. Esko[2,3,4,5], Social Science Genetic Association Consortium.* 1) Department of Psychology, University of Minnesota Twin Cities, Minneapolis, MN, USA; 2) Department of Genetics, Harvard Medical School, Boston, MA, USA; 3) Broad Institute of MIT and Harvard University, Cambridge, MA, USA; 4) Divisions of Endocrinology and Genetics and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA; 5) Estonian Genome Center, University of Tartu, Tartu, Estonia.

Educational attainment (years of schooling) is commonly used as a proxy for cognitive function. We conducted a genome-wide association meta-analysis of educational attainment, using a combined sample of nearly 300,000 individuals. We identified 74 approximately independent genome-wide significant SNPs, and a polygenic score constructed from all genotyped polymorphisms accounts for 3.2% of the variance. Several distinct analyses, including comparisons of population-based and within-family estimates of genetic effects, suggest that very little of our signal is due to population stratification. We found significant genetic correlations between educational attainment and many other traits, including cognitive performance (0.75), intracranial volume (0.34), bipolar disorder (0.28), Alzheimer's disease (–0.31), body mass index (–0.26). The genes in our identified loci are preferentially expressed in neural tissue, particularly during prenatal development, and implicated in shaping features of the brain such as its overall size, morphology, and architecture of axon-dendrite connections. The genetic sites disproportionately reside in regions regulating gene expression in the fetal brain. Specific pathways implicated by our analyses include several that are well known to developmental neurobiologists (signaling by Robo receptor) and also some whose importance to neural development has become clear more recently (signaling by EGFR). Some of the specific prioritized genes in our loci have already been implicated in axon genesis/guidance (*TBR1*, *NFIB*, *DCC*), language and cognition (*FOXP2*), intellectual disability (*MEF2C*, *SMARCA2*, *GRIN2A*), autism spectrum disorder (*BCL11A*, *QRICH1*), and neurodegenerative disorders (*HTT*, *MAPT*). Our results can shed light on how biological and social influences interact to produce variation in educational achievement.

384

Biallelic Mutations in Human Accelerated Regions (HARs) are Associated with Abnormal Social and Cognitive Behavior. *R.N. Doan[1], B.I. Bae[1], M. Nieto[2], B. Cubelos[3], S. Al-Saad[4], N.M. Mukaddes[5], C.A. Walsh[1,6,7], The Homozygosity Mapping Consortium for Autism.* 1) Genetics and Genomics, Boston Children's Hospital, Boston, MA; 2) Departamento de Biologia Molecular, Centro de Biologia Molecular 'Severo Ochoa', Universidad Autonoma de Madrid, UAM-CSIC, Nicolas Cabrera, 1, Madrid 28049, Spain; 3) Department of Molecular and Cellular Biology, Centro Nacional de Biotecnologia, CNB-CSIC, Darwin 3, Campus de Cantoblanco, Madrid, 28049 Spain; 4) Kuwait Center for Autism, Kuwait City 73455, Kuwait; 5) Istanbul Faculty of Medicine, Department of Child Psychiatry, Istanbul University, Istanbul 34452, Turkey; 6) Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA; 7) Departments of Pediatrics and Neurology, Harvard Medical School, Boston, MA.

Comparative genomics have identified regions in the genome that are potentially involved in human evolution, but such comparisons provide no direct insight into how changes in DNA sequence may affect function. Human accelerated regions (HARs) represent 2,737 genomic loci conserved among non-human species, while showing elevated divergence in humans compared to other species. At least some HARs are thought to contribute to neurodevelopmental functions underlying the unique social and behavioral traits of humans. We hypothesized that a subset of essential HARs underlie human-specific behaviors, and hence could be subject to mutation in cognitive and social disorders. Strikingly, we find that biallelic HAR point mutations in whole genome sequence (WGS) and array capture sequence of the "HAR-ome" of 215 families enriched for consanguinity revealed a significant excess in affected individuals, suggesting a contribution to 7% of these consanguineous ASD cases. Even more, we demonstrate functional implications of HAR mutations on enhancer activity impacting regulation of genes such as *CUX1*, which regulates spine density in the most recently-evolved, upper-layer cerebral cortical neurons and *MEF2C*, a transcription factor implicated in social and cognitive disorders. Our data provide the first genetic evidence of HARs regulating neurodevelopment underlying for human social and cognitive behavior.

385

B56δ-related protein phosphatase 2A dysfunction identified in patients with intellectual disability. *G. Houge[1], D. Haesen[2], L.E.L.M. Vissers[3], S. Mehta[4], M.J. Parker[5], M. Wright[6], J. Vogt[7], S. McKee[8], N. Cordeiro[9], T. Kleefstra[3], M.H. Willemsen[3], M.R.F. Reijnders[3], S. Berland[1], E. Hayman[10], E. Lahat[10], E.H. Brilstra[11], C.L.I. van Gassen[11], E. Zonneveld-Huijssoon[11], C.I. de Bie[11], A. Hoischen[3], E.E. Eichler[12], R. Holdhus[13], V.M. Steen[13], S.O. Døskeland[14], D.E. FitzPatrick[15], M.E. Hurles[16], V. Janssens[2], DDD project, Welcome Trust Sanger Institute, Cambridge CB10 1SA, UK.* 1) Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, N-5021 Bergen, Norway; 2) Laboratory of Protein Phosphorylation and Proteomics, Department of Cellular and Molecular Medicine, KU Leuven - University of Leuven, B-3000 Leuven, Belgium; 3) Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Centre for Neuroscience, Radboud University Medical Center, NL-6500 HB Nijmegen, the Netherlands; 4) East Anglian Medical Genetics Service, Addenbrookes Hospital, Cambridge CB2 0QQ, UK; 5) Sheffield Clinical Genetics Service, Sheffield Children's Hospital, Sheffield S10 2TH, UK; 6) Northern Genetics Service, Newcastle upon Tyne Hospitals, Newcastle upon Tyne NE1 3BZ, UK; 7) West Midlands Regional Genetics Service, Birmingham Women's Hospital, Birmingham B15 2TG, UK; 8) Northern Ireland Regional Genetics Centre, Belfast City Hospital, Belfast BT9 7AB, UK; 9) Children's Services - NHS Ayrshire & Arran, Rainbow House, Ayrshire Central Hospital, KA12 8SS, UK; 10) Pediatric Neurology Department, Asaf Harofeh Medical Center, Zriffin , 70300 Israel; 11) Department of Medical Genetics, UMC Utrecht, NL-3508 AB Utrecht, the Netherlands; 12) Department of Genome Sciences, University of Washington, Seattle WA 98195- 5065, USA; 13) Department of Clinical Science, University of Bergen, N-5021 Bergen, Norway; 14) Department of Biomedicine, University of Bergen, N-5021 Bergen, Norway; 15) MRC Human Genetics Unit, MRC Institute of Medical Genetic and Molecular Medicine, Edinburgh EH4 2XU, UK; 16) Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

Here we report inherited dysregulation of protein phosphatase activity as a novel cause of intellectual disability (ID). De novo missense mutations in two different subunits of serine/threonine protein phosphatase 2A (PP2A) were identified in 16 individuals with mild to severe ID, long-lasting hypotonia, epileptic susceptibility, frontal bossing, mild hypertelorism and downslanting palpebral fissures. PP2A holoenzymes comprise catalytic (C), scaffolding (A) and regulatory (B) subunits that determine subcellular anchoring, substrate specificity and physiological function. Ten patients had mutations (E198K, E200K, P201R or W207R) that introduced a basic charge in a highly conserved acidic loop of the PPP2R5D-encoded regulatory B56**δ** subunit, including six individuals who had the same E198K mutation. Five others had de novo mutations (P179L, R182W and R258H, all also cancer-associated) in the *PPP2R1A*-encoded scaffolding Aα subunit. Large ventricles causing macrocephaly and suspicion of hydrocephalus were features in some cases, and all Aα cases had partial or complete corpus callosum agenesis. Functional studies showed that the mutant A and B subunits were stable and uncoupled from phosphatase activity: mutant B56**δ** had deficient A and C binding, while mutant Aα still bound B56**δ** but either had deficient C binding, or bound a C subunit with significantly diminished specific activity. This suggested a dominant-negative effect where mutant B56**δ**, mutant Aα-B56**δ**, or catalytically impaired C-(mutant Aα)-B56**δ** complexes could hinder access of PP2A activity to B56**δ**-anchored PP2A substrates, supported by finding hyperphosphorylation of GSK3β, a B56**δ**-regulated substrate, upon mutant subunit overexpression. Such an effect was also in line with clinical observation indicating a correlation between the degrees of ID and biochemical disturbance.

386

Clinical Indexing Genes Affected by Copy Number Variation in Neurodevelopmental Disorders. *M. Uddin[1,2], G. Pellecchia[1,2], D. Merico[1,2], B. Thiruvahindrapuram[1,2], M. Zarrei[1,2], T. Nalpathamkalam[1,2], K. Tammimies[3], M. Gazzellone[1,2], R.KC. Yuen[1,2], S. Walker[1,2], A. Chan[1,2], L. D'Abate[1,2], A. Noor[4], M.T. Carter[5], G. Yoon[5], P. Kannu[5], C.R. Marshall[1,6], M. Speevak[7], D.J. Stavropoulos[6], S.W. Scherer[1,2,8,9].* 1) The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada; 2) Program in Genetics and Genome Biology (GGB), The Hospital for Sick Children, Toronto, Ontario, Canada; 3) Center of Neurodevelopmental Disorders (KIND), Neuropsychiatric Unit, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden; 4) Department of Pathology and Laboratory Medicine, Division of Diagnostic Medical Genetics, Mount Sinai Hospital, Toronto, Ontario, Canada; 5) Division of Clinical and Metabolic Genetics, Department of Pediatrics, The Hospital for Sick Children, University of Toronto, Toronto, Ontario M5G 2L3, Canada; 6) Genome Diagnostics, Pediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada; 7) Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada; 8) McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada; 9) Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

A significant challenge in clinical genomics testing is determining whether copy number variation (CNV) affecting a gene or multiple genes, will manifest as disease. The increasing recognition of the importance of gene dosage effects in neurodevelopmental disorders(NDDs), prompted us to develop a novel computational approach to estimate the potential etiologic effects of CNV in these conditions. Initially,we developed an *'exon transcriptome-mutation contingency index'* built upon the concept that an inverse correlation tends to exist between exon expression level in brain genes and the burden of rare missense mutations they carry in population controls. Following this approach, previously, we found that specific *critical exons* were significantly enriched in individuals with autism relative to their unaffected siblings(Uddin et al., *Nature Genetics* 2014), and our updated analyses with additional transcriptional and control data added specificity (ranges $P < 1.64 \times 10^{-15}$ to 1.31×10^{-299}) yielding 2,989 candidate genes for having a role in NDDs. Separately, using weighted correlation network analysis (*WGCNA*),we constructed an unbiased protein module involved in neurogenesis, cell signaling, and synaptogenesis from genome-wide mass spectrometer protein expression data from multiple prenatal and adult tissues.We show that the protein module is significantly enriched for *critical exons* in prenatal($P < 1.15 \times 10^{-50}$, OR = 2.11) and adult($P < 6.03 \times 10^{-18}$, OR = 1.55)) developmental periods, and this approach yielded 1,206 proteins for which the corresponding gene is prioritized to have a role in NDDs.We then compared this gene list obtained from our *critical exon* and *WGCNA* strategies, respectively, and found each independently to be associated with CNVs annotated to be previously characterized as clinically significant(CS) (487 genes) and variants of uncertain significance (VOUS)(834 genes). The CS- and VOUS- CNV datasets were based on hand-annotation of 12,440 cases and 9,692 controls from our unpublished clinical research laboratory data. With this combined analysis,and by assessing other large NDD-CNV data, a subset of genes previously considered to be VOUS were identified as new candidate genes for neurodevelopmental disorders(examples being *GIT1*,*PPP1R9A* and *MVB12B*).In summary, we have developed a quantifiable approach that begins to enable the indexing of genes affected by CNVs detected by microarrays(or any other genomic test) for their potential role in NDDs.

387

Functional characterization of novel *DEAF1* mutations in clinical whole-exome sequencing of intellectual disability patients and its regulation of the *RAI1* gene. *L. Chen[1,2], P. Jensik[3], M. Walkiewicz[1], J.T. Alaimo[1], S.V. Mullegama[1], S.H. Elsea[1].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, houston, TX; 2) Department of Cellular and Genetic Medicine, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China; 3) Department of Physiology, Southern Illinois University, Carbondale, IL USA.

Clinical whole-exome sequencing (WES) has exponentially expanded the number of candidate genes for genetically diverse phenotypes such as intellectual disability (ID). Through WES analysis of patients with ID, we identified three novel pathogenic variants in the functional domains of *Deformed Epidermal Autoregulatory Factor-1* (*DEAF1*) gene. *DEAF1* encodes a transcription factor acting both as an activator and a repressor and is involved in early central nervous system development. Electrophoretic mobility shift assays (EMSA) and subcellular localization assays demonstrate that variants R226W and R246T in the SAND domain affect DEAF1 binding affinity, while K305del in NLS domain retains DEAF1 in the cytoplasm, preventing transport to the nucleus, suggesting that the transcriptional function is compromised. We performed a genome-wide scan for the putative DEAF1 binding consensus sequence in gene regulatory regions, and interestingly, one binding site localized to intron 2 of the dosage-sensitive gene *retinoic acid induced 1 (RAI1)*, the causative gene for Smith-Magenis syndrome (SMS) when haploinsufficient, which suggests that DEAF1 may directly regulate expression of *RAI1*. Patients carrying pathogenic *DEAF1* variants display ID, speech delay, sleeping disturbance, neurobehavioral problems, and a high pain threshold, implying a partial phenotypic overlap with SMS. Therefore, we tested DEAF1 interactions with intron 2 of *RAI1* using EMSA and found a direct *in vitro* DNA-protein interaction. Using luciferase reporter assays, we tested the role of DEAF1 in the transcriptional regulation of *RAI1* and observed a significant increase in luciferase activity driven by the putative DEAF1 binding sequence in intron 2 of *RAI1*, suggesting *DEAF1* acts as an activator of *RAI1* gene expression. Our results demonstrate that loss of function mutations within the SAND and NLS domains of *DEAF1* result in severely compromised transcriptional activities and that *DEAF1* acts as an upstream activator of the dosage sensitive gene, *RAI1*. These data suggest that *DEAF1* and *RAI1* may function in the same pathway and that DEAF1 modifies *RAI1* expression, impacting phenotypes commonly observed in both SMS patients and patients carrying pathogenic *DEAF1* variants. These findings provide evidence for common molecular pathways across genetic syndromes with ID.

388

Investigating the transcriptome wide impact of expanded polyalanine tract mutations in *ARX* contributing to intellectual disability and seizures. *C. Shoubridge[1,2], K. Lee[1,2], J. Gecz[1,2], T. Mattiske[1,2].* 1) University of Adelaide, Adelaide, South Australia, Australia; 2) Robinson Research Institute, University of Adelaide, South Australia, Australia.

Aristaless related homeobox *(ARX)* gene encodes a paired-type homeodomain transcription factor with critical roles in embryonic development. Mutations in *ARX* give rise to intellectual disability (ID), epilepsy and brain malformation syndromes. Over half of all mutations in *ARX* lead to expansion of two (of four) polyalanine tracts in the ARX protein. Our recent investigations in mice modelling the two most frequent polyalanine expansions (PA^E) seen in human patients demonstrated that aggregation of mutant Arx protein does not occur in the embryonic brain (1). Instead, we identified a marked reduction in mutant Arx protein expression in the developing forebrain (1). While both mice models display similar reductions in mutant Arx protein, each displays a distinct phenotype that recapitulates the phenotypes seen in patients; $Arx^{(GCG)7}$ (PA1) mice have seizures in addition to learning and memory problems, with seizures not reported in the $Arx^{432\text{-}455dup24}$ (PA2) mice. To investigate the mechanism underpinning these differences we chose a transcriptome wide approach of RNA-seq to analyse gene expression in the developing (12.5dpc) mouse forebrain. Here we report a total of 283 genes significantly deregulated (Log2FC >+/-1.1, P-value <0.05) when both mutations are compared to wild-type (WT) animals. When each mutation is considered separately, a greater number of genes were deregulated in the severe PA1 mice (825) than is noted in the PA2 animals (78) when compared to WT. The majority of genes significantly deregulated in PA1 mice are also perturbed in the milder PA2 mice, but fail to reach significance at this time point. A recent study has demonstrated estradiol administration during early postnatal development prevented spasms in infancy and seizures in adult Arx PA1 mutants (2). Our analysis shows that estrogen response elements are located upstream of promoters in a third of all degregulated genes (229/858, 27%). Our RNA-seq analysis at E12.5 dpc of embryonic development identified a 'core' pathway including two key effectors; *Twist1* and *HDAC4*. This pathway contains down stream targets enriched with estrogen repose elements. Hence, we predict that key pathways disrupted in the mutant mice contribute to the phenotypes and that transcriptional changes involved in the estrogen response pathway may be potential targets for treatment of infantile epileptic seizures.1.Lee et al., Hum Mol Genet 2014 23(4):1084-942. Noebels et al., Sci Transl. Med 6, 220ra12 2014.

389

WD-repeat 47 is essential for the normal brain development through interaction with SCG10 in tubulin-associated processes. *M. Kannan[1,7], M. Roos[2,7], L. McGillewie[3], C. Wagner[1], C. Chevalier[1], U.K. Sanger Mouse Genetics Project[4], G. Grenningloh[5], C. Kinnear[3], Y. Herault[1], B. Loos[2], B. Yalcin[1,6,8].* 1) Institute of Genetics and Molecular and Cellular Biology, Illkirch, 67404, France; 2) Department of Physiological Sciences, Stellenbosch University, South Africa; 3) SAMRC Centre for Tuberculosis Research, Department of Biomedical Sciences, Stellenbosch University, South Africa; 4) Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, United Kingdom; 5) École polytechnique fédérale de Lausanne, Switzerland; 6) Center for Integrative Genomics, University of Lausanne, Switzerland; 7) Joint first authors; 8) Corresponding author.

WD-repeat (WDR) proteins are one of largest eukaryotic family and are involved in a variety of cellular functions ranging from signal transduction to cytoskeletal assembly and apoptosis. More than 246 WDR proteins have been identified, however little is known about their role in neurodevelopment. In this study, we set out to investigate the role of 21 of these WDR proteins using knockout (KO) mouse models with a particular focus on WDR47. *Wdr47* is a novel WD-repeat gene of unknown function, with the highest level of expression in the brain. To gain insight into *Wdr47* function, we developed and characterized a hypomorph as well as a full KO model. First, we established that *Wdr47* is an essential gene for the normal development of the brain. *In vivo* inactivation of *Wdr47* resulted in severe lethality, neuroanatomical defects as well as major behavioral deficiencies. We then investigated cellular morphology *ex vivo*, and found a striking malformation of the growth cone of cortical and hippocampal neurons. Knowing that *SCG10* is highly enriched in growth cones, we checked its co-localization with *Wdr47*. We also carried out yeast two-hybrid screen of a human foetal brain cDNA library and, validated an interaction between WDR47 and SCG10. Furthermore, using *in vitro* knockdown assays of *Wdr47*, we were able to recapitulate *in vivo* findings but also very interestingly found that WDR47 played a role in the migration of neurons. Knockdown of WDR47 significantly reduced neuronal migration distance and velocity, decreased filopodia-like extensions, led to an expanded nuclear envelope and increased autophagic flux. Super-resolution structured illumination microscopy (SR-SIM) revealed a highly convoluted perinuclear tubulin network. These data suggest that WDR47 regulates tubulin dependent processes such as neuronal migration, filopodia formation, axonal outgrowth and autophagy. Our study allows us for the first time to investigate a large number of novel WDR genes involved in brain function. Together KO mice with knockdown assays help unravel a key role of Wdr47 for the normal brain development.

390

Single-cell RNA-Seq of Human Cajal-Retzius Neurons in developing brain. *J. Kim, M. Lin, R. Dominguez, T. Souaiaia, C. Walker, A. Camarena, J. Nguyen, J. Herstein, M. Francois, W. Mack, J. Rossen, C. Liu, O. Evagrafov, R. Chow, J.A. Knowles.* Zilka Neurogenetic Institute, USC, Los Angeles, CA.

The human nervous system is comprised of numerous heterogeneous populations of cells. We have pioneered the use of next-generation RNA sequencing for single-cell expression profiling (Qiu et al., 2012). More recently, our group has performed whole-cell patch clamp, followed by cytoplasm extraction and single-cell RNA sequencing of neurons from intact brain slices or organotypic brain slice cultures. We have sequenced nearly 1,000 cells from adult cerebellum and the temporal, parietal and frontal lobes, as well as fetal brain and spinal cord from gestational weeks 10-20. For these samples, RNA was amplified using the aRNA method (Vn gelder et al., 1990) followed by library construction using the Illumina TruSeq Stranded mRNA preparation kit. On average, ~6,000 genes per cell were detected. Principle component analysis (PCA) revealed heterogeneity in gene expression between cells from different types of tissue. One particular cell type, Cajal-Retzius neurons, is of special interest, as they are morphologically distinct and they play critical role in defining the cortical structure of the human brain. Electrophysiological recording revealed for the first time, spontaneous synaptic activity and action potential firing in human Cajal-Retzius neurons. We have found that the majority of the Cajal-Retzius cells (n>50) [JAK1] cluster separately from subplate neurons using PCA. These clusters have then been used to identify sets of genes that are significantly over- and under-expressed in Cajal-Retzius neurons, as compared to subplate neurons. Our data demonstrate the synergistic potential of combining extensive functional phenotyping and transcriptome analysis at the single-cell level.