ASHG 2024 Platform Abstracts

As of November 20, 2024

Instructions

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

Table of Contents

Session 09: A Heart to Heart on Cardiovascular Genetics in Health and Disease 24

Genome-wide Association Study for Resting Electrocardiogram in the Qatar Biobank Identifies 6 Novel Genes and Validates Novel Polygenic Risk Scores 24

Multi-ancestry GWAS meta-analysis of TG/HDLC ratio in the Million Veteran Program and UK Biobank 25

Machine learning enables discovery of rare coding variants in 17 genes for coronary artery disease 26

Deciphering rare non-coding LDL-C associations in over 246K individuals with whole genome sequencing 27

Proteome-wide Mendelian randomization identifies candidate causal proteins for cardiovascular diseases 29

Characterising the role of 46 candidate genes in early-stage atherosclerosis using CRISPR/Cas9 and live fluorescence imaging 30

Session 10: Advancements in Molecular and Cytogenetic Diagnostics 32

Somatic overgrowth and vascular malformations: Unveiling novel pathogenic variants and clinical utility through comprehensive genetic testing at Genetic Diagnostic Laboratory (GDL) 32

Alport syndrome: Genetic, clinical features and renal transplant outcomes ⭐ 33

Assessing the utility of long-read genome sequencing in undiagnosed rare developmental disorders 34

Inherited metabolic disorders in critically ill patients: Results of genome sequencing of 1,000 consecutive patients from a single centre 35

Advanced Chromosomal and Genomic Abnormality Detection in Hematological Malignancies: Leveraging Genomic Proximity Mapping as a Next-Generation Cytogenomics Tool 36

Living with your dynamic genome: T2T-CHM13 reference genome identifies Robertsonian translocation carriers in healthy newborn cohorts 37

Session 11: All the Single Cells 39

Rare and common genetic variants regulate single-cell expression of immune cells from 2,000 individuals ⭐ 39

Population-scale single-cell RNA-seq across five countries reveal Asian-specific genetic architecture of alternative splicing and complex disease 40

Interindividual cellular and transcriptional regulatory changes in human aging 41

Single-cell long-read sequencing analysis in endemic pemphigus foliaceus 42

Mapping the observable IBDverse: Identifying novel drivers of IBD susceptibility through population-scale, multi-tissue single-cell eQTL mapping..... 43

Single nucleus, multi-ancestry atlas of genetic regulation of gene expression in the human brain..... 44

Session 12: Beyond Genetic Discoveries: Novel Mechanisms of Neurodevelopmental Disorders 46

Monoallelic *de novo* variants in DDX17 cause a novel neurodevelopmental disorder 46

Maternally derived *de novo* variants in the non-coding spliceosomal snRNA *RNU4-2* are a frequent cause of syndromic neurodevelopmental disorders..... 47

Biallelic inactivating variants in *DMAP1* underlie a syndromic neurodevelopmental disorder 48

Loss-of-function of the Zinc Finger Homeobox 4 (*ZFHX4*) gene underlies a neurodevelopmental disorder 49

Biallelic *UGGT1* gene variants cause a congenital disorder of glycosylation 50

Unveiling the crucial neuronal role of the proteasomal ATPase subunit gene *PSMC5* in neurodevelopmental proteasomopathies..... 52

Session 13: Biobank Scale Genetic Data Resources for Studying Complex and Rare Human Diseases 55

100,000 Genomes of Europe: Unlocking genetic variability across Europe for science and health 55

The UAE Genome Program: Unique Genetic Insights from 43,608 Individuals..... 56

Structural variant discovery with GATK-SV in 97,940 short-read whole genomes from the *All of Us* Research Program 57

A complete telomere-to-telomere reference panel of 6404 human haplotypes improves imputation and phasing accuracy 58

Diversity in the NHGRI-EBI GWAS Catalog: addressing disparities while promoting accessibility and data sharing 59

Harmonizing the world's rare disease knowledge in Mondo..... 60

Session 14: Cancer Risk: Novel Genes and Mechanisms 62

An Atlas of Pan-cancer Susceptibility Genes Revealed by Intronic Polyadenylation Transcriptome-wide Association Study 62

Exploring gene-by-environment interactions in colorectal cancer risk using massively parallel reporter assays ⭐ 63

Genetic regulation of *TERT* splicing contributes to reduced or elevated cancer risk by altering cellular replicative potential ⭐ 64

Immune surveillance and cancer risk 65

Prevalence and effect of inherited chromosomally integrated human herpesvirus 6 in 735,434 human genomes 66

Session 15: Decoding Structural Variation at Scale 68

The contribution of linked structural variants to recent positive selection in humans 68

Haplotype-informed analysis of structural variation in 490,414 genomes and its effects on human health ⭐ 69

A phenome-wide association study of tandem repeat variation in 168,554 individuals from the UK Biobank 70

Contribution of Copy Number Variation in Disease Related Phenotypes Risk in 23andMe Research Cohort 71

Pangenome-derived copy number variation maps with global diversity and association analysis in BioBank scale data with ctyper 72

Novel short tandem repeats on the Telomere-to-Telomere reference genome are associated with Alzheimer's disease neuropathology 73

Session 16: I See Ghosts: Archaic DNA in Our Genomes 75

A refined analysis of Neanderthal-introgressed sequences in modern humans with a complete reference genome 75

Patterns of genomic and morphological variation in the mid-19th century burial remains of the Liberated Africans from St. Helena Island in the South Atlantic 76

Deciphering genetic contribution of ancient hunter-gatherer Jomon in Japanese Populations 77

Characterize the nature of ghost archaic introgression in African populations 78

Indirect inheritance of archaic ancestry in modern Peruvians 79

Archaic introgression in Samoans: population structure, genetic admixture, and health associations 79

Session 17: Creative Community Engagement: Gathering Data for Better Participatory Research 81

From Barbershop to Biopsy: Improving Access to Genetic Screening through the Cleveland African American Prostate Cancer Project 81

Co-Creating a story-based video collection to engage LGBTQIA+ community members with the *All of Us* Research Program: An engagement marketing and human entered design approach 82

Breaking Barriers: Project GIVE's Tele-Genetic Initiative for 100 Children with Rare Diseases at the Texas-Mexico Border 83

Assessing diverse communities' perspectives of precision research participation: the Precision rEsearCh pArticipatioN (PECAN) study 84

Genetic Services in Africa: Evidence-Based Recommendations for Policymakers and Healthcare Organizations ⭐ 85

Balancing Constitutional Protections in Genetic Research: Addressing Concerns of Minoritized Communities in Non-consented Tissue Reuse 86

Session 18: Machine Learning and AI Applications in Human Genetics 88

CellPhenoX: An eXplainable Cell-specific machine learning method to predict clinical Phenotypes using single-cell multi-omics 88

scPrediXcan: leveraging single-cell data for transcriptome-wide association studies at cell-type level through transfer learning 89

MILTON: Disease prediction with multi-omics and biomarkers empowers case-control genetic discoveries in UK Biobank 90

Human data, machine learning, and mouse models demonstrate that SPEN plays a critical role in cardiac development 91

Using machine learning to predict noncoding variant associations with sulcal patterns in congenital heart disease 92

Whole exome sequencing shows cystic fibrosis risk variants confer a protective effect against inflammatory bowel disease 93

Session 19: Mapping the Brain in Health and Disease 95

Common variants for migraine tested in 1,138,261 Europeans implicate biological processes with specific effects on head pain severity symptoms 95

Genetic regulation of the gene expression in fetal and adult brains explains GWAS signals from the East Asian population 96

Single-cell atlas of transcriptomic vulnerability across multiple neuropsychiatric and neurodegenerative diseases 97

A nucleotide-scale map of brain cell effects of neurodegenerative GWAS variants reveals distinct and shared causal disease mechanisms 98

Cell-Cell Communication Patterns in Alzheimer's Disease Dementia and Mild Cognitive Impairment Vary by Cortical Layers 99

Mechanisms of presenilin2 driven neuroinflammation: Impact of *PSEN2*-N141I variant on microglial response to Alzheimer's disease-relevant stimuli 100

Session 20: Moving Polygenic Risk Scores Closer to Clinical Implementation 102

Implementation of breast cancer polygenic risk scores in a personalized screening trial 102

All of Us diversity and scale improve polygenic prediction contextually with greatest improvements for under-represented populations 103

Polygenic risk score enriches for clinically significant prostate cancer in a screening program - the BARCODE 1 study results 104

Genetic and metabolomic determinants of disease in the UK Biobank 105

Performance of contemporary polygenic risk scores for atherosclerotic cardiovascular disease in the All of Us Workbench ⭐ 106

PGS Browser: a comprehensive analysis of 3,168 polygenic score models across 400,000 Finns ⭐ 107

Session 21: Multimodal Approaches to Interpreting the Non-Coding Genome: Evolution, Functional Genomics, and Machine Learning 109

Evolutionary conservation and functional analysis of neuronal regulatory elements in mammals 109

Uncovering gene regulatory differences between human and chimpanzee neural progenitors 110

Cross-species variant-to-function analyses implicate insomnia effector genes and reveal a highly conserved regulatory architecture at the *MEIS1* locus 111

Constructing cell type-specific enhancer-promoter regulatory interaction networks with massively parallel reporter assays 112

Unbiasedly partitioning the heritability of scRNA-seq data reveals that the vast majority of cell type-specific gene regulation lies in trans 113

Predicting and interpreting functional non-coding regulatory variants with base-resolution deep learning models of chromatin accessibility 114

Session 22: New Frontiers in Multi-ancestry Methods for Complex Traits 116

Multi-trait and multi-ancestry genetic analysis of comorbid lung diseases and traits improves genetic discovery and polygenic risk prediction 116

Dissecting ancestry-aware molecular causal effects for type 2 diabetes 117

Quantifying genetic effect heterogeneity across ancestral populations 118

AI-STAAR: An ancestry-informed association analysis framework for large-scale multi-ancestry whole genome sequencing studies 119

A multi ethnic meta analysis of genome wide association studies identified additional novel genomic loci associated with cervical cancer ⭐ 120

A multi-ethnic reference panel to impute classical and non-classical *HLA* class II alleles: Enhancing HLA Imputation Accuracy in Admixed Populations 121

Session 23: Not Only Genetics: Integrating Other Omics Approaches 123

Single nucleus multiome optimizations for postmortem human brain and large scale multiome profiling of Alzheimer's disease progression reveal novel gene regulatory mechanisms and effects of APOE 123

An Atlas of Protein Quantitative Trait Loci in Olink and Somascan Platforms Uncover Genetic Insights into Gastroenterological and Hepatological diseases 124

Computational Analysis of Microbiome Genetics in Head and Neck Squamous Cell Carcinoma 125

Prioritizing Alzheimer’s Disease genetic risk variants with massively parallel reporter assays and 3D chromatin structure 126

Multi-omic Profiling in a 61 day Pig Kidney to Human Decedent Xenotransplant Reveals a Concerted Acute Rejection Immune Response 127

Targeted CRISPRa/CRISPRi screen identifies functional variants and novel target genes at multiple renal cell carcinoma (RCC) susceptibility loci 128

Session 24: The Sex-Specific Landscape: Variation, Regulation, and Expression 130

CHD1-deficiency shows sexual dimorphism mediated by androgen exposure 130

Sex differences in brain cell-type specific chromatin accessibility in schizophrenia 131

Disentangling mechanisms underlying sex differences in gene regulation using population-scale multi-omics 132

Let's talk about sex: how biological sex affects functional variation across the genome to alter risk of human disease 132

Large-scale analyses of variants with sex-biased population allele frequencies, sex-biased association with phenotypes, and sex-biased allele penetrance in 43 human tissues 133

GenESIS: enhancing transferability of polygenic scores with gene-by-sex interactions 135

Session 25: Decoding Gene Expression Cis and Trans 137

Colocalization of >1,200 skeletal muscle genes with GWAS loci for musculoskeletal and cardiometabolic traits: a muscle eQTL study of 1,002 individuals 137

Identifying the molecular mechanisms of complex disease through a genome-wide *trans*-eQTL meta-analysis in 43,301 individuals 138

Polymorphic short tandem repeats shape single-cell gene expression across the immune landscape 139

Variation and regulatory mechanisms of the small RNA transcriptome across human tissues 141

Session 26: Genetic Approaches Informing Drug Targets and Mechanism 143

Replication of genetic associations across diverse ancestry groups is indicative of drug target success in clinical trials ⭐ 143

Prioritising New Antihypertensive Drug Targets and Unravelling Disease Modulation by Antihypertensive Drugs Using Mendelian Randomisation 144

Identification of plasma proteins as promising therapeutic targets to treat hypertension 145

Deep learning modeling of rare noncoding genetic variants in human motor neurons defines CCDC146 as a therapeutic target for ALS 146

Session 27: Interrogating Variant Function at Scale 148

High-throughput Deep Mutational Scanning to determine pathogenicity of Variants of Uncertain Significance in genes in the Sonic Hedgehog Pathway 148

PerturbVI: A scalable latent factor model to infer regulatory modules from large-scale CRISPR perturbation data 149

Changes in Kv11.1 (*hERG*/*KCNH2*) protein interactomes from hiPSC-derived cardiomyocytes of individuals with extreme QT interval polygenic scores and CRISPR edited rare variants 150

A shared autophagy pathway dysregulated in multiple neurodegenerative diseases revealed by phenotypic CRISPR screens of iPSC-derived neurons with familial mutations 151
Session 28: Liver, Laugh, Love: New Insights into Liver Disease 153
Genetic Determinants of Liver Function Markers in African Ancestry Populations 153
Investigating the Role of LYPLAL1 Loss-of-Function in Metabolic Dysfunction-Associated Steatotic Liver Disease 154
Characterizing 99 candidate genes for a role in MASLD and MASH using CRISPR/Cas9, *in vivo* imaging and deep learning in zebrafish larvae 155
Machine learning-based subtyping and validation with longitudinal patient data in metabolic dysfunction-associated steatotic liver disease 156
Session 29: Modeling Rare Neurodevelopmental Disorders in Human iPSCs and Mice ... 158
Rapid generation of mouse model mimicking VUS uncovers novel pleiotropy in neurodevelopmental disorders 158
Loss of SZT2 leads to an increase in outer radial glia by hyperactivation of mTORC1 in human brain organoids 159
Investigating NuRDopathies with GATAD2B-associated Neurodevelopmental Disorder (GAND): clinical evaluations and modeling with patient-derived IPSCs and mice 160
Variants in cohesin release factors define a novel class of cohesin balance disorders 161
Session 30: Novel Aspects of Modeling Genetic Architectures of Complex Traits 163
Selection, pleiotropy, and chance: why rare and common variant association studies often implicate different genes 163
Determining the driving factors shaping genetic architecture of complex traits in recently admixed populations 164
Genomic and ethnolinguistic diversity in >40,000 eastern and southern Africans highlights the ongoing impact of cultural affiliation shaping genetic variation 165
Detecting ongoing natural selection affecting allele frequencies across generations to uncover genetic variants contributing to disease susceptibilities 166
Session 31: Therapies for Genetic Disorders 168
A drug repurposing screen identifies NSAIDs and COX1/2 enzyme inhibition as potential therapies for MAN1B1-CDG, a rare congenital disorder of glycosylation. 168

NAGLU co-expressed with a modified phosphotransferase has increased mannose-6-phosphorylation and shows preclinical efficacy as a treatment for mucopolysaccharidosis IIIB (Sanfilipo B Syndrome) 169

Antisense oligonucleotide therapy in an individual with KIF1A-associated neurological disorder 170

Rescue of Proteus syndrome lethality in mice with prenatal miransertib treatment..... 171

Session 32: Unifying Multimodalities: Insights from Single Cell Analyses 173

Leveraging single-cell multi-omic profiling to investigate non-coding variants in Parkinson's disease ⭐ 173

Single-cell eQTL analysis in >2,000 individuals in conjunction with single-cell multiomics analysis in 271 individuals infers causal disease mechanisms 174

A Multiomics Single Cell Atlas Redefining the Human Maternal-Fetal Interface by Spatial Cellular Mapping 175

Identifying Noncoding Regulatory Variants by Multiome Single-Cell Sequencing in Prostate Cells..... 176

Session 43: All about Implementation 178

Implementation of hereditary cancer risk assessment in primary care settings: Strategies and proximal outcomes 178

The Million Veteran Program Return Of Actionable Results (MVP-ROAR) Study: Preliminary outcomes from participants receiving clinical genetic confirmation testing for familial hypercholesterolemia 179

Introducing an efficient framework to evaluate oncology and cardiology gene-disease validity leveraging clinicogenomic biobank data..... 180

Provider acceptance of patient-facing digital genetics service delivery tools: a qualitative study 181

Does universal testing under payer medical policy equate with genetic testing coverage for patients with ovarian, pancreatic, male breast, and early-onset colorectal cancer? 182

Costs and outcomes of opportunistic genomic screening: Findings from the Incidental Genomics randomized controlled trial 183

Session 44: Alzheimer's Disease from Gene Discovery to Multi-omics Integration 185

Discovering Genes Associated with Alzheimer’s Disease via multi-tissue and cell type Transcriptome-Wide Association Study 185

Integration of GWAS, 3D genomics, and CRISPRi screens in microglia implicates causal variants and genes at Alzheimer's disease loci, including at *TSPAN14* 186

Deciphering single-cell genomic landscape of brain somatic mutations in Alzheimer's disease 187

Large-scale proteomic and genomic analysis identify plasma proteins influencing human brain structure and Alzheimer's disease risk 188

Unraveling the Propagation of Functional Genetic Effects in Alzheimer's Disease on a Population Scale 189

Astrocytes from diverse ancestries reveal key differences in APOE expression and other AD risk genes across populations 190

Session 45: Disease Insights from Omic-Wide Approaches 192

Large-scale genome-wide association study meta-analysis across 1,962,069 individuals reveals insights into the genetic mechanisms of osteoarthritis 192

Multi-ancestry proteome-wide Mendelian randomization offers a comprehensive protein-disease atlas and potential therapeutic targets 193

Transcriptome-wide association study of early substance use reveals associations between tobacco use and predicted gene expression in adolescents 194

All by All of Us: common and rare variant association testing in 245,000 whole genomes across diverse ancestry groups 195

Genome-wide association study and predictors of lymphocyte-related blood cell traits in Hispanic/Latino newborns 196

Omic Risk Scores are Associated with Cross-Sectional and Longitudinal Chronic Obstructive Pulmonary Disease-Related Traits Across Three Cohorts 197

Session 46: Diverse Epigenetic Marks in Health, Diagnosis, and Disease 199

H3K36 methylation - a guardian of epigenome integrity 199

m^6A mediated epitranscriptomic dynamics in human brain development and disease 200

Most genetic effects on DNA methylation are shared across tissues 201

Multi-platform long read genomics identifies methylation outliers in rare disease 202

A novel single-cell sequencing method for *CHD2* variant classification in childhood epilepsies 203

In-utero rescue of neurological dysfunction in a mouse model of Wiedemann-Steiner syndrome 204

Session 47: From Variant to Function: Prediction and Understanding Variants Function . 206

Defining the function of disease variants with CRISPR editing and multimodal single cell sequencing 206

Functional genomics applied to mapping the gene regulatory mechanisms downstream of neuron-astrocyte interactions 207

CircRNA mediated polyadenylation alteration contribute to Alzheimer's disease pathogenesis 208

In Silico Module Perturbation Analysis unlocks a functional understanding of the dynamic gene networks in single-cell data 209

De Novo Precise Splice Site Predictor Using Deep Learning and Integration with Minimap2 for Enhanced Long-Read Sequence Alignment 210

Classification of rare nonsynonymous variants to identify individuals at low risk of disease: introducing variants of potential risk 211

Session 48: Novel Genetic, Genomic, and Epigenetic Resources in the Era of Big Data ... 213

The developmental Genotype-Tissue Expression projects 213

The Clinical Genome Resource (ClinGen): Advancing Genomic Knowledge through Global Curation 213

The New York Genome Center ALS Consortium combines postmortem tissue transcriptomics with whole genome sequencing to empower biological discovery 214

FILER 2.0: Unified access to >100,000 omics datasets across >1,000 cell types and tissues 216

Whole exome sequencing of 44,028 British South Asians in Genes and Health uncovers 2,917 genes with putative human knockouts for systematic characterization 217

Enhanced Genetic Insights from Brain Region-Specific GWAS Using Deep Unsupervised Learning Derived Endophenotypes on UK Biobank T1-Weighted MRI Data 218

Session 49: Polygenic Risk Scores: Novel Methods for Modeling Risk 220

JointPRS: A Comprehensive Framework for Genetic Prediction Across Populations Incorporating Genetic Correlation and Combining Meta-Analysis and Tuning Strategies 220

A Novel Polygenic Risk Scoring Framework Integrating Common and Rare Variants for Enhanced Genetic Prediction Across Ancestries 221

Modeling diagnostic code dropout of schizophrenia in electronic health records improves phenotypic data quality and transferability of polygenic risk scores for a diverse Veteran cohort 222

A Deep Ensemble Encoder Network Method for Improved Polygenic Risk Score Prediction 223

Integrative polygenic score modeling with tissue-specific annotation improves polygenic scores transferability 224

Functional gene embeddings improve rare variant polygenic risk scores 225

Session 50: The Context of All in Which We Live: Gene by Environment Interactions 227

The Genetic Basis of Environmental Exposures in the Personalized Environment and Genes Study (PEGS) 227

Nature versus nurture of glucose homeostasis trajectories in children 228

Decomposing sex-different phenotypic correlations in the UK Biobank into genetic and environmental components 229

Neanderthal introgression modifies the response to environmental stimuli in modern humans 230

Assessing cellular contexts of type 2 diabetes-associated variants at scale 231

Alternative polygenic score approaches aid in detecting genetic modification of the relationship between adiposity and cardiometabolic risk 232

Session 51: 3D Chromatin and Epigenomics 234

Dissecting the genetic underpinnings of chromatin loops and their relationship to transcriptional regulation 234

Comprehensive Single-Nucleus Analysis of Genetic Regulation on Gene Expression and Chromatin Accessibility in Human Kidneys to understand of genetic basis of chronic kidney disease 235

Genetic and Epigenetic Insights into the Aging of the Human Retina 236

Single-cell genomics, QTLs, and regulatory networks for 388 human brains 237

Session 52: Computational Methods for Causal Variant Prioritization 239

Footprint quantitative trait loci (fpQTLs) reveal non-coding causal variants associated with transcription factor binding for liver traits 239

A new variant-to-disease score prioritizing causal variants in GWAS 240

Robust fine-mapping in the presence of LD mismatch 241

Do deep genome language models help pinpoint causal variants in statistically fine-mapped loci? 242
Session 53: Dysfunction at the Powerhouse: Molecules, Models, and Organisms 244
Investigating the role of seryl-tRNA synthetase (*SARS2*) in mitochondrial biology and human recessive disease 244
COXFA4 Dysfunction Leads to ODC Dysregulation: A Link to Mitochondrial Disease Mechanism 245
A multiomic approach to elucidate muscle-specific pathogenesis of SUCLA2-deficient mitochondrial myopathy 246
Identification and targeting of ABHD18 as a strategy to alleviate TAZ mutant phenotypes 247
Session 54: Expanding the Table: Considerations for Inclusion in Genetics and Genomics 249
Equity-focused implementation illuminates diverse perspectives in rare disease research 249
Use of exclusion criteria to select critically ill newborns for rapid genome sequencing captures precise genetic diagnoses missed by use of conventional inclusion criteria . 250
Reprogenomics, Ethics and Inclusivity: Perspectives from Sex and Gender Diverse Communities 251
The NIH INCLUDE Project: Over five years of transformational research for people with Down syndrome 252
Session 55: Insights into Somatic Mosaicism and Human Diseases..... 254
A personalized multi-platform assessment of somatic mosaicism in the human frontal cortex 254
Reconstructing Cell Lineage in Human Brain Using Somatic Mutations in Microsatellites 255
Somatic genomic changes in single ischemic human heart cardiomyocytes 255
Genotype-Informed Single-Cell RNA-Seq Reveals Somatic Loss of Heterozygosity in Hemimegalencephaly with *PIK3CA* Mutations 256
Session 56: Neurogenomic Approaches Translating Risk Variants to Disease 258
Complex Structural Genome Variation in the Genetic Architecture of Neuropsychiatric disorders: Insights from Human Population Analysis and from Postmortem Brains of Individuals with Psychiatric Disorders 258

Sex differences of the spatiotemporally dynamic FMRP-RNA interactome in the human brain 259

ASXL1 mutations drive mitochondrial dysfunction, resulting in disrupted mTOR signaling and cellular proliferation in Bohring Opitz Syndrome 260

Translating *IGHMBP2* variants with a patient-specific neuromuscular junction system: Personalized medicine rescue 261

Session 57: Population Genetics Methods Matter 263

Characterizing features affecting local ancestry inference performance in diverse admixed populations 263

A genealogy-based approach for revealing ancestry-specific structures in admixed populations 264

Deep learning-augmented models of gnomAD v4 enable estimation of LoF mutational constraint for all human genes 265

Genotype Representation Graphs: Enabling Efficient Analysis of Biobank-Scale Data 266

Session 58: Scaling Structural Birth Defects 268

Whole genome analysis of 137 trios with CHARGE-phenotype overlap 268

Functional validation of a novel gene associated with orofacial clefts 269

Analysis of rare *de novo* variants in 5707 congenital heart disease (CHD) trios identifies three novel CHD genes 270

Unraveling the Diverse Genetic Architecture of Structural Birth Defects 271

Session 69: Complex Traits and Other Omics 273

Genetically predicted leukocyte telomere length from 800,000 individuals identifies novel phenotypic associations 273

Multiomics approach identifies novel genes for Skeletal Class III malocclusion 274

Complex interactions of copy number variants on rare and common disorders 275

Uncovering the nuclear genetic basis of mitochondrial DNA heteroplasmy 276

A phenome-wide association study of the structural variants in 467,152 UK Biobank genomes identifies non-coding structural variants associated with human diseases .. 277

Detecting large complex structural variants from human genome assemblies 278

Session 70: Exploring the Genetic Spectrum of Obesity 280

An abdominal obesity missense variant in the transcription factor and thermogenesis gene *TBX15* shows signals of adaptation to cold in Finns and affects downstream adipocyte expression in *trans* 280

Functional characterization of 14 obesity-associated genes using CRISPR in human white adipose tissue implicates *SLTM* as a novel lipid accumulation gene 281

Prioritization of effector genes within body mass index loci yields molecular insight into the biology of body weight regulation 282

The phenotypic variability, dose-response, and temporal effects in polygenic prediction of adiposity traits 284

Weight loss with semaglutide is influenced by traditional metabolic risk factors and BMI-associated genetic variants 285

Session 71: Long-Read Transcriptomes in Health and Disease 287

POISEN: A Bioinformatics Pipeline to Identify Poison Exons in Long-Read Transcriptomes 287

The Spatial Atlas of Human Anatomy (SAHA) project: Unveiling cellular landscapes of health and diseases and orchestrating a new paradigm in precision medicine 288

Genome-wide profiling of highly similar paralogous genes using HiFi sequencing..... 289

Combining spatial transcriptomic with snRNA-seq data enhances differential gene expression analyses 290

Benchmarking detection of technically challenging pathogenic variants with long-read sequencing and a head-to-head comparison with short-read sequencing in a clinical diagnostic laboratory 291

A variety of molecular mechanisms cause copy number gains at 17p11.2 locus causing Potocki-Lupski syndrome: understanding patients with CNVs that do not include *RAI1 gene* 292

Session 72: Pharmacogenomics: DNA and Drugs 294

Incorporation of Local Ancestry (LA) in a GWAS of warfarin dose requirement in African Americans (AAs) identifies a novel CYP2C19 Splice QTL ⭐ 294

Genome-wide association study on ACE-inhibitor switching identifies missense variants in *NTSR1* and *CACNA1H* 295

Epigenetic patient stratification via contrastive machine learning refines hallmark biomarkers in minoritized children with asthma 296

Understanding the impact of drug perturbations on disease-specific protein networks 297

Prioritization of icosapent ethyl for the potential reversal of metabolic dysfunction associated fatty liver disease using a genetically informed drug repurposing pipeline ⭐ 298

Combining genetics with real-world patient data enables ancestry-specific target identification and drug discovery 300

Session 73: Stats Just Wanna Have Fun: New Methods in Statistical Genetics 302

Integrative Statistical Framework for Detecting Divergent Selection and Linking to Disease 302

Genome-Wide Assessment of Pleiotropy Across >1000 Traits Among >1.5 Million Participants of Diverse Biobanks ⭐ 303

Pleiotropic heritability quantifies the shared genetic variance of common diseases ... 304

ENCODE cCRE-based WGS analysis of 100 traits in UK Biobank identifies 1,987 associations driven by rare-variants 305

Enhancing regulatory variant prioritization via long-range DNA sequences and multi-task learning 306

Trans-modeling of large-scale proteomics data uncovers enriched protein-protein interactions and drug targets 307

Session 74: The Non-coding Genome: From Nucleotide to Protein 309

Interpreting Regulatory Differences between Species in Terms of Potential Cis- and Trans- Mechanisms 309

Nanopore sequencing of chromatin accessibility 310

BRAIN-MAGNET: A novel functional genomics atlas coupled with convolutional neural networks facilitates clinical interpretation of disease relevant variants in non-coding regulatory elements 311

Variable number tandem repeats (VNTRs) regulate epigenome and transcriptome in human prefrontal cortex 312

CRISPRi perturbation screens and eQTLs capture different target genes for non-coding GWAS variants 313

Connecting rare variation to extremes of plasma protein levels 314

Session 75: Tick-Tock: The Aging Genome 316

Cell-type-specific effects of aging on the human prefrontal cortex transcriptome across the lifespan 316

Characterization of de novo retrotransposition events in the aging germline 317

Linking Rare Non-Coding Variants Associated with Human Longevity to Cellular Senescence via Integrative Functional Genomic Approaches 318

Longitudinal Proteomic Aging Index Construction using Functional Principal Component Analysis 319

Epigenetic age acceleration across chronological age groups and its modifiable lifestyle risk factors in middle-aged and elderly adults 320

Session 76: Translating Genetics into Screening Programs 321

Identification of actionable genetic variants in 4,198 volunteers from the Viking Genes research cohort and implementation of return of results 321

Early Check Genome Sequencing of Newborns to Detect Genetic Risk of Type 1 Diabetes 322

Combining gene genealogies and pedigrees to inform genetic screening programs 323

Rate and profile of secondary findings in 381 participants in the DDD-Africa from the DR Congo ⭐ 324

An association study without genotype sharing for uncovering germline susceptibilities in pediatric cancers..... 325

Biobank-scale genotype-to-phenotype analyses reveal the challenges in using exome sequencing for population screening 326

Session 77: Exploring Omics: From Genomes to Microbiomes 328

Institution-wide access to a scalable, clinical grade genomic sequencing platform advanced rare disease research and improved clinical outcomes in a pediatric setting 328

Project Baby Lion - Introducing ultra-rapid genome sequencing in German neonatal and pediatric ICUs 329

Assessing HiFi genome sequencing as first-tier test in rare disease genetics 330

Expanding the human gut microbiome atlas of Africa 331

Session 78: Genetics of Human Brain: Regulation, Disease Risk, and Assortative Mating 332

Establishing the Molecular Foundation of Brain Anatomy in Living Individuals 332

Mapping the regulatory effects of rare non-coding variants across cellular and developmental contexts in the brain 333

The largest to-date exome study of autism spectrum disorder triples the number of autism-associated genes 334

Assortative Mating Across Nine Psychiatric Disorders: Consistency and Persistence Across Cultures and Generations 335

Session 79: Lessons from Height 337

Leveraging whole-genome sequencing data from 750,000 diverse-ancestry individuals across biobanks to understand the genetic architecture of common anthropometric traits 337

Impact of rare coding variants on height prediction in a diverse set of >1 million individuals 338

Machine learning reveals 3D regulatory mechanisms for height-associated haplotypes 339

GWAS of infant and early childhood height in up to 70 000 children: Genetic influences on the early phases of childhood growth 340

Session 80: Linking Non-coding Variation to Function via Diverse Epigenetic Mechanisms 342

Single cell multi-omics and 3D genome architecture reveals novel pathways and targets of metabolic dysfunction-associated steatohepatitis 342

Machine learning identifies chromatin features that predict the sensitivity of regulatory sequences to inhibition of BAF chromatin remodeling activity 343

Response sQTLs in primary human chondrocytes identify novel putative osteoarthritis risk genes 344

Massively parallel reporter assay highlights the importance of B cell activation in uncovering latent QTLs, especially for eQTLs 345

Session 81: Rare Variants and Admixture Modeling in Diverse Population 347

Large-scale admixture mapping in the *All of Us Research Program* improves the characterization of cross-population phenotypic differences 347

Multi-ancestry GWAS for hypermobile Ehlers-Danlos Syndrome 348

Rare variant associations and fine-scale population structure in the Genes & Health Study of >44,000 British South Asians 349

Network comparison of ancestry-specific genetically correlated diseases in a meta-analysis of phenome-wide association studies from 1 million individuals 350

Session 82: Read All about It: Transcriptomic Insights from New Sequencing Technologies 352

Identifying pathogenic variants that cause Mendelian conditions using long-read transcript sequencing 352

Single-Cell Omics for Transcriptome CHaracterization (SCOTCH): isoform-level characterization of gene expression through long-read single-cell RNA sequencing ... 353

Applications of long-read RNA sequencing improves the design and interpretability of RNA-based therapeutics 354

Combined long- and short-read RNA sequencing of pathogen stimulated primary immune cells identifies the expression of uncharacterized genes and transcripts 355

Session 83: Splice Splice Baby: Isoform Expression in Health and Disease..... 357

An atlas of expressed transcripts in the prenatal and postnatal human cortex ⭐ 357

Uncovering the brain-specific genetic regulation of splicing by mapping splicing quantitative trait loci in 10,887 post-mortem brain RNA-seq samples..... 358

A high throughput splicing assay for characterization of rare variants of unknown significance 359

Defining the landscape of poison exon splicing events in the human brain: implications for neurodevelopmental and neurodegenerative disorders 360

Session 84: Strategies to Interpret Germline Variants in Cancer Predisposition Genes ... 362

Applying scalable machine-learning approaches to generate evidence to impact variants of uncertain significance in Lynch syndrome genes 362

When clinical meets molecular: why, when and how do *CTNNA1* germline variants cause hereditary diffuse gastric cancer development 363

Comparing scalable and automated vs. ACMG/AMP variant interpretation strategies for BRCA1 and BRCA2 in a large clinicogenomic cohort from six US-based health systems 365

Expanding the reach of paired DNA and RNA sequencing: Results from 450,000 consecutive individuals from a hereditary cancer cohort 366

Session 87: Framing Heritability for Complex Traits..... 368

Fine-mapped insertions and deletions disproportionately impact 78 diseases and complex traits 368

Heritability and effect-size distribution of rare and de novo protein-coding variation ... 369

Uncovering the contribution of rare variants to the heritability of complex traits: Insights from the UK Biobank whole genome sequencing data ... 370

Partitioning genetic and non-genetic contributions to epigenetic-defined endotypes of allergic phenotypes in children ... 371

Session 88: Keeping It REnAL! Genetic Studies of Kidney Disease ... 373

GWAS of multiple renal function biomarkers and kidney multi-omics prioritizes new chronic kidney disease genes ... 373

Characterization of a novel *ASAH2* variant associated with diabetes and kidney failure in Tongan and Samoan patients ... 374

KidneyGenAfrica: Putative novel genetic loci and improved polygenic prediction for kidney function derived from aggregating 10 continental African genome-wide association studies ⭐ ... 375

SLC6A19 loss of function is associated with improved kidney function and metabolic reprogramming of kidney cells ... 376

Session 89: Long-Read Sequencing Offering New Insights into Neurological Disease ... 378

Long-read sequence and assembly of autism reference genomes ... 378

Long-read sequencing to diagnose Autosomal recessive Parkinson's disease in diverse populations ... 379

Mapping parent-of-origin methylation pattern during development by long-read 5-base HiFi sequencing reveals novel imprinting motifs and insight into human disease ... 380

Identification of *FXN* protomutation alleles explains the unequal population distribution of Friedreich ataxia ... 381

Session 90: Tumor Genome Landscape Studies ... 382

A common missense polymorphism in the *PARP1* gene is associated with distinct tumor transcriptomic, immune and clinical profiles in high grade serous ovarian cancers ... 382

Characterization of the immunosuppressive microenvironment driven by HBV-infected tumor cells in hepatocellular carcinoma through single-cell sequencing ⭐ ... 383

Single-cell RNAseq revealed multiple resistance mechanisms in patient-derived xenograft model of rectal cancer during treatment ... 384

Session 91: Unraveling the Complexity of Polygenic Inheritance ... 386

Beyond known genes and relationships for craniofacial abnormalities ... 386

Non additive interactions between rare variants and lifestyle factors contribute to obesity 387

Polygenic risk of rheumatoid arthritis regulates the abundance of circulating regulatory T cells 388

An atlas of associations between plasma proteins biomarkers and polygenic risk scores for cancer and other complex human diseases 389

Session 92: Genetic Information in Breast Cancer Risk Assessment and Screening 391

Comprehensive Genetic Risk Assessment for Breast Cancer in a Diverse Cohort: Preliminary Findings from the eMERGE Study 391

Investigating genotype-estrogen interactions in breast cancer through a combined molecular and epidemiological approach 392

Finding the pathogenic variant in the haystack: using breast-cancer-related family history from electronic health records to identify patients who should be prioritized for genetic testing 393

Polygenic risk score (PRS) significantly improves breast cancer (BC) risk assessment for diverse ancestries 394

Session 93: Modeling Ataxia and Neuropathy 396

Sumo1 mutation modifies behavioral performances in Fragile X associated tremor/ataxia mouse model 396

SNX13 and SNX14 influence neuronal lipid homeostasis and associate with spinocerebellar ataxia and intellectual disability syndromes 397

Investigating R-Loop Formation as a Potential Pathomechanism in Spinocerebellar Ataxia 27B Using iPSC-Derived GABAergic Neurons 398

Predictive modeling to define the locus heterogeneity of tRNA synthetase-related peripheral neuropathy 398

Session 94: More than One Way to Break a Gene - Variant Effects on RNA 400

Unveiling the hidden role of RNA stability as a link between genetic variation and disease 400

Splice Switch: An investigation on the effect of a sQTL on PAPR2 isoforms and subsequent influenza A virus susceptibility 401

Modulation of the impact of genetic mutations on human health by transcriptional adaptation 402

Systematic analysis of nonsense variants uncovers peptide release rate as a novel modifier of nonsense-mediated mRNA decay efficiency 403

Session 95: Phenomenal PheWAS 405

Genome-wide association studies in a large Korean cohort identify novel quantitative trait loci for 36 traits and illuminate their genetic architectures 405

The phenomic landscape of gain- and loss-of-function genetic variants across diverse human populations 406

Phenome-wide study reveals multiple diseases and biomarkers causally associated with alcohol consumption 407

Phenome-Wide Association of *APOE* Alleles in the *All of Us* Research Program 408

Session 96: Technology for Translation 410

A randomized study of a digital genetic health portal (MyCancerGene) for patients who have received germline cancer genetic test results as compared to usual care 410

GenAI-powered approaches in advancing genetic testing education and communication: an exploratory study in Pharmacogenomics 411

Machine learning predictions to shorten diagnostic odysseys in Level IV NICUs 412

The All of Us Research Program data release 2024 (CDR v8): Powering genomic research through All of Us 413

Session 09: A Heart to Heart on Cardiovascular Genetics in Health and Disease

Location: Room 401

Session Time: Wednesday, November 6, 2024, 8:00 am - 9:30 am

Genome-wide Association Study for Resting Electrocardiogram in the Qatar Biobank Identifies 6 Novel Genes and Validates Novel Polygenic Risk Scores

Authors: M. Saad[1], N. Khan[1], A. Shaar[1], **K. Kunji**[2], A. Khan[3], M. Elshrif[1], B. Mohammed[4], M. Thamer Ali[5], A. Al Haj zen[6], K. Kiryluk[3], G. Nemer[7], A. Fahed[8]; [1]Qatar Computing Res. Inst., Doha, Qatar, [2]Qatar Res. Computing Ctr. (HBKU), Doha, Qatar, [3]Columbia Univ., New York, NY, [4]Hamad Med. Corp., Doha, Qatar, [5]Heart Hosp., Doha, Qatar, [6]Coll. of Hlth.and Life Sci., Hamad Bin Khalifa Univ., Doha, Qatar, [7]Hamad Bin Khalifa Univ., Doha, Qatar, [8]Broad Inst. of MIT and Harvard, Boston, MA

Abstract:

Background: Electrocardiography (ECG) is one of the most valuable non-invasive diagnostic tools in determining the presence of many cardiovascular diseases. Genetic factors are important in determining ECG abnormalities and their link to cardiovascular diseases. Genome-wide association studies (GWAS) and polygenic risk scores (PRS) have been conducted for various ECG traits such as QT interval and QRS duration. However, these studies mainly focused on cohorts of European descent.
Methods: In this study, GWASs for 6 ECG traits (RR, PR, QTc, QRS, JT, and PW) were conducted in a Middle Eastern cohort from the Qatar Biobank, comprising 13,827 subjects with whole genome sequence (WGS) data. Middle Eastern PRSs were developed using clumping and thresholding, and their performance was compared to 26 published PRSs.
Results: Seventy-four independent loci were obtained with genome-wide significance across the 6 traits ($P < 5\times10^{-8}$). Out of the 74 loci, 67 (90.5%) were previously reported and 7 loci (9.5%) were novel and contained 6 genes: *STAC* and *CSMD1* for PR, *ANK1* and *NCOA2* for QRS, *LSP1* for QTc, and *MKLN1* for PW. All 26 published PRSs showed good performance in our cohort. PGS002276 showed the best performance for QTc ($R^2 = 0.059$, $P = 4.83\times10^{-185}$), PGS002166 showed the best performance for QRS ($R^2 = 0.024$, $P = 1.53\times10^{-75}$), and PGS000905 for PR ($R^2 = 0.053$, $P = 2.57\times10^{-165}$). Some of these PRSs were associated with cardiovascular diseases. For example, PGS003500, a QTc PRS, was significantly associated with cardiomyopathy (Odds Ratio per 1SD=1.58, 95% CI [1.23 - 2.01], $P = 2.42\times10^{-4}$). Middle Eastern PRSs substantially outperformed published PRSs and did not perform well in the UK Biobank data. Ten pathogenic variants, including 3 that

are specific to Qataris, were observed in 17 long QT syndrome genes and were carried by 19 individuals. The QTc average was larger for mutation carriers (415.6 ± 23.5 vs 402.3 ± 18.5 in non-carriers). Five-year follow data did not show a significant change in ECG patterns, regardless of mutation status and PRS values. Four individuals out of 2302 had prolonged QTc intervals over the two timepoints.
Conclusion: In this first GWAS for ECG traits in the Middle East using WGS data, 7 novel loci (6 genes) were identified. Published PRSs performed well, but newly developed Middle Eastern-specific PRSs performed the best. Novel variants in long QT syndrome genes were observed for the first time in Qataris. Follow-up data did not show significant changes in ECG patterns.

Multi-ancestry GWAS meta-analysis of TG/HDLC ratio in the Million Veteran Program and UK Biobank

Authors: P. Kho[1,2,3], S. Koyama[4,5,6], R. Guarischi Sousa[2], S. Clarke[1,2,7], P. Tsao[8,3,2], T. Assimes[1,2,3,9]; [1]Dept. of Med., Div. of Cardiovascular Med., Stanford Univ. Sch. of Med., Stanford, CA, [2]VA Palo Alto Hlth.care System, Palo Alto, CA, [3]Cardiovascular Inst., Stanford Univ. Sch. of Med., Stanford, CA, [4]Broad Inst., Cambridge, MA, [5]Cardiovascular Res. Ctr. and Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [6]Dept. of Med., Harvard Med. Sch., Boston, MA, [7]Dept. of Med., Stanford Prevention Res. Ctr., Stanford Univ. Sch. of Med., Stanford, CA, [8]Dept. of Med., Stanford Univ. Sch. of Med., Stanford, CA, [9]Dept. of Epidemiology and Population Hlth., Stanford Univ. Sch. of Med., Stanford, CA

Abstract:

The triglyceride-to-high-density lipoprotein cholesterol (TG/HDLC) ratio is a marker of cardiovascular health, correlated with insulin resistance and type 2 diabetes (T2D). We present a multi-ancestry genome-wide association study (GWAS) meta-analysis of the TG/HDLC ratio, using data from the Million Veteran Program (MVP) and UK Biobank (UKB). Our analyses involved up to 985,855 participants from diverse ancestries, including 799,556 European-, 117,201 African-, 54,626 Admixed-American-, 7,314 East Asian-, and 8,158 South Asian-participants.
Our multi-ancestry GWAS meta-analysis uncovered 1,283 independent genome-wide significant SNPs associated with the TG/HDLC ratio, of which 906 were novel. Notably, 204 of these SNPs were unique to TG/HDLC and had not been previously associated with either TG or HDLC. Further, 32 SNPs were specific to the African-ancestry GWAS, underscoring the potential for novel discoveries in understudied populations. Phenome-wide genetic correlation analyses revealed strong associations between the TG/HDLC ratio and traits

related to lipid metabolism, T2D, and anthropometric measures. Traits such as reticulocyte count, alcohol consumption, and sex hormone-binding globulin (SHBG) were also correlated with TG/HDLC ratio. Elevated reticulocyte count may be linked to hyperinsulinemia in insulin-resistant states, while alcohol consumption and SHBG levels potentially influence glycemic and lipid profiles.
Tissue enrichment analysis pinpointed key tissues relevant to the TG/HDLC ratio: subcutaneous adipose, visceral omentum adipose, adrenal glands, liver, and pancreas. We conducted transcriptome-wide association studies using eQTL data from these relevant tissues obtained from GTEx. These analyses, along with fastBAT gene-based analysis, consistently prioritized 44 genes across different methods, highlighting their relevance to the TG/HDLC ratio. Moreover, we assessed the causality of plasma proteome from UKB in MVP data and revealed 29 likely causal proteins for the TG/HDLC ratio.
This ongoing study will incorporate additional biobank data and conduct functional follow-up studies on the prioritized genes. In summary, our multi-ancestry GWAS meta-analysis of the TG/HDLC ratio identifies several novel loci and highlights three important contributing factors: substantial scale, with inclusion of nearly 1 million subjects; inclusion of participants from diverse ancestries; and focus on TG/HDLC as a distinct lipid phenotype. These findings may pave the way for targeted therapeutic strategies and personalized medicine approaches to enhance cardiovascular health across diverse populations.

Machine learning enables discovery of rare coding variants in 17 genes for coronary artery disease

Authors: B. Petrazzini[1], I. Forrest[1], G. Rocheleau[1], H. Vy[1], C. Marquez-Luna[1], R. Chen[1], J. Park[1], K. Gibson[1], S. Goonewardena[2], W. Malick[1], R. Rosenson[1], D. Jordan[3], R. Do[1]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Univ. of Michigan, Michigan, MI, [3]Icahn Sch. of Med. at Mt Sinai, New York, NY

Abstract:

Coronary artery disease (CAD) is a heterogenous disease caused by multiple risk factors. Over the past decade, low prevalence of CAD in the general population and suboptimal phenotyping have hindered rare variant discovery efforts. Importantly, progression of atherosclerosis leads to diverse clinical manifestations of CAD which can result in phenotypic differences within cases and subclinical CAD in controls. Traditional phenotypes of cases versus controls lack granularity to capture these meaningful differences within classes. A recent study demonstrated that an in-silico score for CAD (ISCAD) built using machine learning and clinical features in electronic health records can

capture disease progression, severity and underdiagnosis on a spectrum (PMID 36563696). Here, we evaluated whether ISCAD can strengthen genetic association analyses to discover new rare coding variants for CAD.
Using whole exome sequences of 604,914 individuals from the UK Biobank, All of Us Research Program and Bio*Me* Biobank we tested ISCAD for association with 2,738,849 rare (minor allele frequency<=0.01) missense or protein-truncating variants and aggregates of deleterious ultrarare variants (minor allele count<5) in 17,883 genes and 50 biological processes. Finally, we evaluated prior genetic, biological and clinical evidence supporting the role of associated variants and genes in CAD.
The study identified 17 genes associated with ISCAD. These are known CAD genes including *LDLR* (β=0.37, P=1.18x10^{-15}), *APOB* (β=0.32, P=5.53x10^{-12}) and *APOC3* (β=-0.13, P=1.81x10^{-11}), as well as genes not previously associated with CAD using rare variants including *HECTD4* (β=0.68, P=4.79x10^{-17}), *SH2B3* (β=0.17, P=1.67x10^{-13}), *GCK* (β=0.45, P=5.58x10^{-11}), *MCL1* (β=-0.043, P=1.22x10^{-8}), *EML3* (β=-2.36, P=4.79x10^{-8}), *OSBPL3* (β=-0.81, P=9.47x10^{-8}), *SOS2* (β=-0.049, P=2.05x10^{-7}), *LIPG* (β=-0.042, P=3.36x10^{-7}) and *PLCB3* (β=0.077, P=3.85x10^{-7}). Examining multiple lines of evidence, we demonstrated that 14 (82.35%) genes have at least moderate genetic, biological or clinical support for their role in CAD. Moreover, we discovered biological processes including pancreatic β cell function and progression through the cell division cycle in gene set analyses. Finally, an excess of deleterious ultrarare coding variants in 321 known CAD genes pooled together (β=0.026, P=7.20x10^{-9}) demonstrates that there are additional associated rare coding variants in these genes.
These results reveal new genetic mechanisms for CAD. Importantly, this work demonstrates that digital markers of complex disease can accelerate discovery of rare coding variation.

Deciphering rare non-coding LDL-C associations in over 246K individuals with whole genome sequencing

Authors: M. Selvaraj[1], X. Li[2], Z. Li[3], E. Buren[4], X. Lin[5], G. Peloso[6], P. Natarajan[7], TOPMed Lipids Working Group; [1]MGH, Boston, MA, [2]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [3]Northeast Normal Univ., Changchun, China, [4]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA, [5]Harvard T.H. Chan Sch Pub Hlth, Boston, MA, [6]Boston Univ., Boston, MA, [7]Massachusetts Gen. Hosp., Boston, MA

Abstract:

Background: Low-density lipoprotein cholesterol (LDL-C) is a heritable risk factor for cardiovascular disease. Recent genome-wide association studies (GWAS) have identified numerous loci related to blood lipid levels, but the role of rare non-coding variants is less well-understood. Whole-genome sequencing (WGS) allows exploration of these variants; however, large sample sizes are required to provide novel insights. Our study used a computationally efficient meta-analysis approach for functionally informed aggregation-based tests that combines score statistics and their covariances with the MetaSTAAR framework using WGS data from two large independent datasets (UK Biobank, n=173,982; TOPMed, n=72,175), yielding the largest WGS analysis for LDL-C. **Methods:** We ascertained deep-coverage WGS and LDL-C measures from UK Biobank and NHLBI TOPMed freeze 10 (23 cohorts). We harmonized and normalized LDL-C from individual cohorts and adjusted for age, sex, ethnicity-by-cohort, PCs and accounted for lipid-lowering medications. To enable efficient WGS meta-analysis for aggregation-based tests across UK Biobank and TOPMed freeze 10, we implemented the MetaSTAAR workflow, which stores variant summary statistics efficiently. In addition to single variant analyses, we performed gene-centric coding and non-coding set-based, and region-based sliding window meta-analyses of rare variants (MAF < 1%) for LDL-C. Finally, we replicated our findings in *All of Us* WGS data (AOU, n=86,540). **Results:** We generated variant summary statistics and covariance matrices for UK Biobank and TOPMed, independently. We processed 571M and 660M variants from TOPMed and UKBiobank, respectively, in which 92M variants had a minor allele count >20. We used 5 gene-centric coding and 7 non-coding variant masks and filtered genome significant aggregates based on Bonferroni-correction [0.05/(20K*masks)]. Before conditional analysis we obtained 70 and 111 aggregates significantly associated with LDL-C for coding and non-coding regions, respectively. After adjusting for known common variants we obtained 39 and 44 aggregates and finally replicated 25 coding and 28 non-coding aggregates. Many important known Mendelian genes including *LDLR, APOB, PCSK9* were significant even after adjustment for known common variants and novel rare variant aggregates in *ABCA6*:plof-ds/missense and *RELB*:UTR were also associated with LDL-C and replicated at genome significant thresholds. **Conclusion:** In summary, we extend prior observations of rare non-coding variants near Mendelian lipid genes to novel genes without prior known common non-coding or rare variant coding evidence.

Proteome-wide Mendelian randomization identifies candidate causal proteins for cardiovascular diseases

Authors: C. Li[1], N. De Jay[1], S. Sharma[1], K. A. Catalano[1], V. Sridharan[1], Z. Wang[1], L. Zhao[1], J. D. Szustakowski[2], C-P. Chang[1], J. C. Maranville[3], E. M. Kvikstad[4], E. R. Holzinger[5]; [1]Bristol Myers Squibb, Cambridge, MA, [2]Bristol Myers Squibb, Pennington, NJ, [3]Bristol-Myers Squibb, Cambridge, MA, [4]Bristol-Myers Squibb, Brisbane, CA, [5]BMS, Cambridge, MA

Abstract:

Background: Cardiovascular diseases (CVD) remain the leading cause of death worldwide. Integration of human genetics and proteomics across different ancestries provides a novel, affordable, and systematic approach for target identification. Methods: Bi-directional Mendelian randomization approach was applied to unravel causal associations between 2,940 circulating proteins and 21 CVD. Genome-wide summary statistics for human plasma proteome were obtained from 46,595 participants from the UKB-PPP consortium for discovery, and 1,225 unrelated individuals from FinnGen Olink study for validation. Summary statistics for CVD were from UK Biobank and FinnGen for European ancestry and Biobank Japan for Asian ancestry. Forward and reverse causation were studied to distinguish targets and biomarkers, respectively. We further prioritized drug targets by integrating biological, clinical and population study evidence through cross-database annotations and literature review. Phenome-wide causality scan and single-cell enrichment analysis were performed to further understand target safety profile and mechanisms of action. Results: We found 221 candidate causal proteins that impacted risk of one or more CVD through forward MR, among which 112 were previously reported as associated with CVD or CVD-related traits, such as APOE, LPA, and PCSK9. There were 139 (62.9%) proteins replicated (FDR < 5%) using FinnGen Olink data. BTN2A1 was highlighted as a novel candidate gene for ischemic stroke, suggesting a crosstalk between immune modulation and stroke pathogenesis. Single cell integration further prioritized ADAM23 for cardiomegaly, PAM for stable angina pectoris and ventricular arrythmia and LPL for peripheral artery disease, whose transcriptional expressions were enriched in cardiomyocytes. Most candidate causal proteins (73.4%) identified are supported by strong literature evidence for a role in immune response, vascular remodeling, myogenesis or energy metabolism. Sixteen proteins were significant in reverse MR, whose expression levels were affected by CVD. Forward and reverse MR found largely non-overlapping proteins (only 2 overlapped: LGALS4 and MMP12), suggesting distinct proteomic causes and consequences of CVD. Conclusions: Our study identified potential therapeutic targets for CVD and distinguished them from biomarkers due to reverse causation. This study

provides human genetics-based evidence of novel candidate genes, a foundational step towards full-scale causal human biology-based drug discovery for CVD.

Characterising the role of 46 candidate genes in early-stage atherosclerosis using CRISPR/Cas9 and live fluorescence imaging

Authors: M. den Hoed[1], E. Mujica[1], A. Emmanouilidou[1], H. Zhang[1], E. Mazzaferro[1], C. Metzendorf[1], M. Bandaru[2], D. Djordjevic[3], S. Gry Vienberg[3], A. Larsson[1], J. Flannick[4], A. Allalou[1]; [1]Uppsala Univ., Uppsala, Sweden, [2]Uppsala Univ., Uppsala, Uppsala, Sweden, [3]Novo Nordisk A/S, Måløv, Denmark, [4]Boston Children s Hosp., Boston, MA

Abstract:

Aim: Genome-wide association studies identified hundreds of loci associated with coronary artery disease (CAD). For most loci, causal genes remain uncharacterised. We use CRISPR/Cas9 and fluorescence microscopy in zebrafish larvae to systematically characterise the role of genes in vivo. **Methods:** One human candidate gene at a time (n=46), we targeted all transcripts of zebrafish orthologues using CRISPR/Cas9, by microinjection at the single cell stage. We used fertilised eggs from zebrafish (AB) that carry transgenically expressed, fluorescently labelled macrophages and neutrophils or an oxidised LDL (oxLDL)-binding antibody. Larvae were overfed from day 5 to 10, before live imaging using semi-automated fluorescence microscopy. In each larva, vascular accumulation/co-localisation of moieties was subsequently quantified using deep learning-based neural networks for image analysis. On average, we imaged 102 affected larvae and 83 sibling controls per gene (n_{total} 9537). For each trait (n=17), the effect of gene perturbation was examined using linear regression. **Results:** For 21 genes, perturbation affects at least one vascular trait. Genes affecting at least one vascular trait in zebrafish larvae are enriched for common variant associations with CAD in humans, and for pathways like negative regulation of lipoprotein lipase activity and cellular response to oxLDL particle stimulus. For 14 of 21 genes, pLOF variants or drugs have been reported to influence atherosclerosis in humans or mice, with 11 genes (79%) showing directionally consistent results across species. One of these is *IL1B* - encoding the target of canakinumab - for which CRISPR/Cas9-induced mutations result in 2.8±1.1 SD units less vascular co-localisation of lipids and neutrophils (less vascular inflammation) in zebrafish larvae. **Conclusions:** Characterising candidate genes for a role in early-stage vascular inflammation and atherosclerosis in zebrafish larvae can bridge the gap between genetic discovery in humans and in-depth characterization of putative causal genes in larger animals.

Session 10: Advancements in Molecular and Cytogenetic Diagnostics

Location: Four Seasons Ballroom 2&3

Session Time: Wednesday, November 6, 2024, 8:00 am - 9:30 am

Somatic overgrowth and vascular malformations: Unveiling novel pathogenic variants and clinical utility through comprehensive genetic testing at Genetic Diagnostic Laboratory (GDL)

Authors: M. Limmina[1], S. Sheppard[2], D. Adams[3], A. Britt[3], J. Kalish[3], M. Deardorff[4], **A. Ganguly**[1]; [1]Genetics Diagnostic Lab., Dept. of Genetics, Univ. of Pennsylvania, Philadelphia, PA, [2]Div. of Translational Med., Eunice Kennedy Shriver Natl. Inst. of Child Hlth.and Human Dev., Rockville, MD, [3]Children's Hosp. of Philadelphia, Philadelphia, PA, [4]Children's Hosp. of Los Angeles, Los Angeles, CA

Abstract:

Introduction: Somatic overgrowth syndromes and vascular malformations (SOVM) stem from post-zygotic variants that activate genetic pathways, causing abnormal tissue growth and vascular defects. Identifying these somatic variants is crucial for targeted therapy and understanding disease heterogeneity. Detection is complicated by the many genes involved and appropriate tissue selection. GDL's 34-gene SOVM panel, with a 69% diagnostic yield, has identified numerous novel pathogenic variants, enhancing knowledge, and paving the way for more precise treatments. **Methods:** Patients who received the SOVM panel from 2018 to 2024 were included. DNA from tissue underwent NGS with a targeted read depth of 4000X. Variants were confirmed and classified per modified ACMG guidelines. Diagnostic yield was calculated as the rate of pathogenic variants. Informed consent was obtained, and data were anonymized. **Results:** 597 cases with diverse clinical presentations underwent SOVM testing, yielding a 62% diagnostic rate. Excluding germline-only samples (blood or saliva) increased the yield to 69%. We identified 132 unique variants across 23 genes (105 pathogenic and 27 VUS). Most variants were in pathways regulating cell growth, differentiation, and survival, aberrantly activating them. The largest group of variants was in the PI3K/AKT/mTOR pathway. Pathogenic *PIK3CA* variants accounted for 64% of positive cases, with 44 unique variants. Additionally, 14 unique pathogenic variants were found in *AKT1, AKT3, MTOR, PIK3R1,* and *PTEN*. The second largest group had pathogenic variants in the GPCR signaling pathway, with 9 unique variants in *GNA11, GNA14, GNAQ,* and *SMO*. The RAS/MAPK pathway included 17 unique variants in *HRAS, KRAS, NRAS, BRAF, RASA1, MAP2K1,* and *MAP3K3*. Pathogenic variants

outside these major pathways included 17 unique variants in *FLT4, TEK,* and *KRIT1*, which are related to angiogenesis and vascular integrity. Notably, 9 *TEK* cases had two pathogenic variants in exon 17. There were also 3 unique variants in *IDH1* and *IDH2* which have been shown to lead to production of oncometabolites that impact epigenetic regulation and cellular differentiation. Lastly, we identified 4 cases with pathogenic variants in two genes (e.g., *PIK3CA* and *GNA11*) indicating likely independent clonal evolution. Depending on tissue selection and DNA quality, we detected variants down to 0.5% allele frequency. **Conclusion:** GDL's panel improves diagnostic yield for SOVM, especially when an affected tissue is submitted. This panel has uncovered many novel variants, enhancing our understanding of the heterogeneity of these conditions, and improving the potential for targeted therapy.

Alport syndrome: Genetic, clinical features and renal transplant outcomes ⭐

Authors: A. Abid, A. Raza, T. Aziz, M. N. Zafar, S. A. Rizvi, A. Lanewala; Sindh Inst. of Urology and Transplantation, Karachi, Pakistan

Abstract:

Alport syndrome is the most common inherited kidney disease, with an incidence of one in 100 for autosomal dominant and one in 2000 for X-linked patterns. Clinical features range from isolated hematuria and proteinuria to end-stage kidney disease, often accompanied by extra-renal anomalies such as hearing loss, and retinopathy. This study describes the mutational landscape and its clinicopathological significance in a large cohort of Alport/Alport-like syndrome cases from 35 families, utilizing targeted next-generation DNA sequencing and subsequent bioinformatics analysis. The pathogenicity of identified variants was evaluated using in-silico tools and following ACMG guidelines. Family segregation was analyzed in available family members. Likely-Pathogenic/pathogenic sequence variants in the *COL4A3*, *COL4A4*, and *COL4A5* genes were detected in 77% of the cohort. A total of 24 different variants were identified in 25 solved cases; 15 variants were novel. A homozygous/heterozygous pathogenic variant in the *COL4A3* gene was detected in 12 cases, *COL4A5* variants in 10 cases, and *COL4A4* gene variants in 3 cases. Digenic inheritance was observed in two families. Three early-onset cases were diagnosed and treated as having steroid-resistant nephrotic syndrome. They showed partial remission to CNIs and were found to have no pathogenic variants in the *NPHS1* and *NPHS2* genes. We found pathogenic variants in the *COL4A3* gene in these children. Most patients in this cohort presented with ESRD, of which eighteen underwent live-related kidney transplants. The identified variants were associated with a spectrum of nephropathy, from microscopic

hematuria to progressive renal disease leading to ESRD, and extra-renal manifestations such as sensorineural deafness and ocular anomalies. This study presents a rapid and low-cost approach for genetic screening of Alport syndrome using targeted NGS, highlighting the importance of an accurate and affordable screening platform. The results of this study have broader clinical implications, particularly for our center, which has a large live-related renal transplant program and a big population of nephrotic syndrome. According to the current international screening guidelines, kidney donation from a person having a heterozygous pathogenic variant should be considered cautiously, as these individuals may develop hematuria and declining kidney function over time. Therefore, we perform genetic testing for nephropathies in familial cases for pre-transplant work-up. This approach helps ensure donor suitability and optimize long-term outcomes for both recipient and donor.

Assessing the utility of long-read genome sequencing in undiagnosed rare developmental disorders

Authors: L. Werren[1], P. Vats[1], M. Peracchio[2], C. King[2], E. Charnysh[1], P. Audano[3], P. Robinson[3], L. Kalsner[2,4], A. Matson[2,5,6], M. Kelly[1], M. Adams[3]; [1]The Jackson Lab. for Genomic Med., Farmington, CT, [2]Connecticut Children's Med. Ctr., Hartford, CT, [3]The Jackson Lab., Farmington, CT, [4]Univ. of Connecticut Sch. of Med., Farmington, CT, [5]Connecticut Children's Med. Ctr., Farmington, CT, [6]UConn Hlth., Farmington, CT

Abstract:

Developmental disorders exhibit high genetic heterogeneity, often with a constellation of non-specific syndromic features, posing major challenges for molecular diagnosis. While exome and genome sequencing efforts have improved diagnostic yield, more than half of individuals with rare disease and their families remain on diagnostic odysseys**.** The limitation of short-read technologies in assessing the full repertoire of disease variants, such as complex structural variants, likely contributes to the high negative result rate of genetic testing. Long-read whole genome sequencing (LR-WGS) using high quality PacBio HiFi read technology holds promise for uncovering missed etiologies. To assess this, we recruited affected children (and their unaffected parents) with undiagnosed rare developmental disorders with suspected genetic cause, and prior negative or inconclusive genetic testing. For all participants, genomic DNA extracted from fresh whole blood was sequenced using a PacBio Revio system to a target mean coverage of 30X. For analysis, we developed a robust pipeline that leverages both phased assembly-based and read-based tools to call a wide range of variant types including single nucleotide variants (SNVs),

structural variants (SVs), and repeat expansions. Phenotype-driven variant prioritization using Human Phenotype Ontology (HPO) terms was performed using a mixture of tertiary analysis solutions, including publicly (SvAnna) and commercially (Illumina's Emedgene Software) available tools. Using this approach, we were able to compare the efficacies of variant calling between assembly-based and read-based tools, as well as across variant prioritization tools. We observed an abundance of prioritized variants that are inherited with the same zygosity as unaffected parents, underscoring the need for trio data to identify and rule out false candidates. In addition, to rule out potential candidate SVs based on frequency in the general population, we relied on the LR-WGS population data from the Human Genome Structural Variation Consortium version 2 (HGSVC2), Human Pangenome Consortium, and Genome Answers for Kids Database. Several prioritized SVs by our pipeline were observed at an allele frequency higher than expected for disease in population databases, yet absent from short-read based SV population datasets such as gnomADv4, demonstrating the need for LR-WGS population databases. To date, we have identified reportable findings in ~17% of analyzed trios. Our findings highlight the limitations of phenotype-driven approaches and the difficulty of structural variant interpretation in rare disease.

Inherited metabolic disorders in critically ill patients: Results of genome sequencing of 1,000 consecutive patients from a single centre

Authors: E. ØStergaard[1], S. Gronborg[1], F. Wibrand[1], J. Ek[2], L. Nazaryan-Petersen[1], M. Duno[3], J. H. Svensmark[1], H. Karstensen[1], J. K. Jensen[1], R. Kjartansdottir[1], M. Bak[1], A. Lund[1]; [1]Copenhagen Univ. Hosp. RigsHosp.et, Copenhagen, Denmark, [2]Copenhagen Univ. Hosp., Copenhagen, Denmark, [3]UniHosp CPH, Copenhagen, Denmark

Abstract:

The implementation of whole genome sequencing in diagnostics has led to an increased diagnostic rate in patients with Inherited Metabolic Disorders (IMDs). Whereas biochemical testing was previously the first-line test, in many cases WGS and metabolic tests are performed simultaneously. We here investigate the contribution of IMDs in a group of critically ill patients and explore the significance of biochemical testing in diagnostics in this group.
Patients, mainly children, were referred for WGS from Denmark, Faroe Islands and Greenland in the period 2018 - 2024 with the indication of suspected IMD, inherited epilepsy, cardiac disease, progressive neurological disorder or severe malformation syndrome. IMDs were classified according to ICIMD. Biochemical testing was done using

mainly mass spectrometry and WGS was carried out using an Illumina platform. Data were analyzed for variants in a panel of ca 2,100 genes related to IMD and epilepsies or the mendeliome.
Out of the 1,000 patients, an IMD diagnosis was reached in 141 patients (ca 14%). The largest groups were disorders of amino acid metabolism (19 patients), disorders of complex molecule degradation (21 patients) and neurotransmitter disorders (20 patients). In 34 patients, biochemical tests pointed to a specific IMD or group of IMDs, including a positive newborn screening result for 7 patients. In additional 7 patients, biochemical testing was used to confirm the diagnosis, e.g. if variants identified by WGS were classified of uncertain significance (VUS), or the phenotype was atypical for the disease.
We found an incidence of IMD of 14% in a group of critically ill patients. In the majority of patients the initial diagnosis was achieved through WGS, whereas ca ¼ had a biochemical test result pointing to a specific disease or group of diseases. The study underscores the need for access to rapid WGS, but also highlights the importance of biochemical testing for rapid diagnostics and followup of genetic test results.

Advanced Chromosomal and Genomic Abnormality Detection in Hematological Malignancies: Leveraging Genomic Proximity Mapping as a Next-Generation Cytogenomics Tool

Authors: Y. Liu[1], H. Fang[1], M. Malig[2], M. Wood[2], E. Reister[2], A. Muratov[2], I. Liachko[2], S. Eacker[2]; [1]Univ. of Washington, Seattle, WA, [2]Phase Genomics, Seattle, WA

Abstract:

Chromosomal and genomic abnormalities are key markers in hematological malignancies, often analyzed through cytogenetics to aid diagnosis and treatment decisions. However, current methods like karyotyping, FISH, and chromosomal microarray analysis have non-overlapping limitations, requiring multiple tests and sometimes yielding low-resolution results. Newer techniques like long-read sequencing, optical genome mapping, and Genomic Proximity Mapping (GPM) are being explored to address these challenges. However, as karyotyping, long-read sequencing and optical genome mapping also are not suitable for malignancies tested with formalin-fixed paraffin-embedded (FFPE) tissue specimens. In our study, we investigated GPM as a solution to streamline testing and overcome these challenges. GPM captures structural variants and ultra-long range sequence contiguity using a specialized library preparation method compatible with various specimen types, including FFPE, and read-out by low-pass short read sequencing. We applied this method to >30 diagnostic samples from ten different types of leukemias

and lymphomas and compared the GPM findings to those of standard-of-care cytogenetics. GPM yielded high-quality data from over 96% of archival samples, accurately identifying all major diagnostic findings including instances of translocations, inversions, copy number variations, gene fusions, and copy neutral loss of heterozygosity except a single instance of aneuploidy. Additionally, it revealed clinically significant variants such as cryptic rearrangements of known driver genes (for example *BCL11B* rearrangements, deletion of *TP53, TP73*, and *CDKN2A*) and chromothripsis, which were missed by traditional methods. Importantly, GPM detected additional variants in over 85% of samples, showcasing its effectiveness in uncovering actionable insights in hematological malignancies and surpassing the limitations of current cytogenetic approaches, including its compatibility with FFPE specimens.

Living with your dynamic genome: T2T-CHM13 reference genome identifies Robertsonian translocation carriers in healthy newborn cohorts

Authors: A. Rhie[1], **J. Kim**[1,2], S. Solar[1], S. Koren[1], B. Pickett[1], B. Walenz[1], A. Guarracino[3], L. Gomes de Lima[4], M. Borchers[4], T. Potapova[4], E. Garrison[3], J. Gerton[4], A. Phillippy[1]; [1]NIH/NHGRI, Bethesda, MD, [2]Seoul Natl. Univ., Seoul, Korea, Republic of, [3]Univ. of Tennessee Hlth.Sci. Ctr., Memphis, TN, [4]Stowers Inst. for Med. Res., Kansas City, MO

Abstract:

The completion of the first telomere-to-telomere (T2T) human genome elucidated sequences not accessible before. The majority of the new sequences reside in centromeric, satellite repeats or segmental duplications, which were difficult to assemble and map before despite their important roles. Among the large, newly accessible regions, several reside in the acrocentric chromosomal p-arms. The draft human pangenome sequences revealed loci of high sequence homology and recombination rates, which were termed “pseudo homologous regions (PHR)”. Heterologous recombination has been observed between these acrocentric chromosomes, especially among Chrs 13, 14 and 21 which commonly share a large PHR. The inverted nature of the PHR on Chr 14 is predicted to facilitate the formation of Robertsonian translocations, resulting in a dicentric chromosome and the loss of two rDNA arrays along with the distal ends of the original short arms. Using a short-read genotyping pipeline specifically designed to detect these translocations, we successfully identified losses of two rDNA arrays in known Robertsonian cell lines. Extending this work to a large cohort of approximately 4,000 individuals sequenced with short reads, we identified 4 potential Robertsonian carriers, matching the previous report at about 1 in every 800 individuals. Linking the distal sequences of the

ribosomal DNA to the pairing q-arm was a long standing challenge in genome assembly. However, using the latest assembly methods and the pangenome, it is now possible to obtain the complete sequences of the most dynamic region, finally giving access to the biology behind it. As a future work, we will further investigate into the haplotypes of the forth-coming near T2T genomes, and comprehensively identify the polymorphic nature of the PHRs.

Session 11: All the Single Cells

Location: Room 501

Session Time: Wednesday, November 6, 2024, 8:00 am - 9:30 am

Rare and common genetic variants regulate single-cell expression of immune cells from 2,000 individuals ⭐

Authors: A. Cuomo[1,2,3,4,5], H. Tanudisastro[2,3,4,6], E. Spenceley[1,4], W. Zhou[7,8,9,10,11], A. Xue[1,4], M. Kanai[8,9,12,13,14], G. Chau[8,9], C. Krishna[12,13,14], R. Xavier[12,13,14], M. Daly[8,9,10,15], B. Neale[8,9,11,15], D. Neavin[1], M. Welland[2,3,4], B. Bowen[1,4], K. de Lange[2,3,4], A. Hewitt[16,17,18], G. Figtree[19,20], D. MacArthur[2,3,4], J. Powell[1,4,5]; [1]Garvan Inst. of Med. Res., Darlinghurst, Sydney, Australia, [2]Ctr. for Population Genomics, Garvan Inst. of Med. Res., Sydney, Australia, [3]Ctr. for Population Genomics, Murdoch Children's Res. Inst., Melbourne, Australia, [4]Faculty of Med. and Hlth., Univ. of New South Wales, Sydney, Australia, [5]UNSW Cellular Genomics Futures Inst., Univ. of New South Wales, Sydney, Australia, [6]Faculty of Med. and Hlth., Univ. of Sydney, Sydney, Australia, [7]Psychiatric and Neurodevelopmental Genetics Unit, Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [8]Stanley Ctr. for Psychiatric Res., Broad Inst. of MIT and Harvard, Cambridge, MA, [9]Program in Med. and Population Genetics, Broad Inst. of MIT and Harvard, Cambridge, MA, [10]Inst. for Molecular Med. Finland, Univ. of Helsinki, Helsinki, Finland, [11]Novo Nordisk Fndn. Ctr. for Genomic Mechanisms of Disease, Broad Inst. of MIT and Harvard, Cambridge, MA, [12]Infectious Disease and Microbiome Program, Broad Inst. of MIT and Harvard, Cambridge, MA, [13]Dept. of Molecular Biology, Massachusetts Gen. Hosp., Boston, MA, [14]Ctr. for Computational and Integrative Biology, Massachusetts Gen. Hosp., Harvard Med. Sch., Boston, MA, [15]Analytic and Translational Genetics Unit, Dept. of Med., Massachusetts Gen. Hosp., Boston, MA, [16]Menzies Inst. for Med. Res., Univ. of Tasmania, Hobart, Australia, [17]Dept. of Ophthalmology, Royal Hobart Hosp., Hobart, Australia, [18]Ctr. for Eye Res. Australia, Univ. of Melbourne, Melbourne, Australia, [19]Charles Perkins Ctr., The Univ. of Sydney, Sydney, Australia, [20]Kolling Inst. of Med. Res., Royal North Shore Hosp., Sydney, Australia

Abstract:

Understanding the genetic basis of gene expression can shed light on the regulatory mechanisms underlying complex traits and diseases. Single-cell resolved measures of RNA levels and single-cell expression quantitative trait loci (sc-eQTLs) have revealed genetic regulation that drives sub-tissue cell states and types across diverse human tissues.
Current sc-eQTL studies are still relatively underpowered: the largest dataset to date,

published recently by our lab, encompasses data from just under 1,000 individuals (OneK1K). Moreover, single-cell genetic studies at present focus only on the role of common genetic variation (minor allele frequency >5%), thus missing effects from rarer genetic variants that can play an important role in human biology and disease.
The TenK10K project is a new initiative performing whole genome sequencing (WGS) and single-cell RNA sequencing (scRNA-seq) on peripheral blood mononuclear cells (PBMCs) for 10,000 individuals, generating the largest set of paired human WGS and scRNA-seq data to date. Phase 1 of TenK10K comprises 28 cell types for 5,084,027 immune cells from approximately 2,000 individuals.
Using a Poisson mixed model, our newly developed tool SAIGE-QTL models single-cell expression profiles directly - rather than 'pseudobulk' aggregate counts. Interim results on OneK1K show that SAIGE-QTL has significantly more power than existing pseudobulk methods when mapping common variants' effects (48% more eGenes, i.e., genes with at least one eQTL). Using summary-based Mendelian randomisation analysis from four immune-mediated diseases we identified 25% more disease associations compared to pseudobulk-based sc-eQTLs, including genes likely involved in the aetiology of inflammatory bowel disease and rheumatoid arthritis.
The speed and scalability of SAIGE-QTL also makes the testing of trans-eQTLs - where a genetic variant affects the expression level of a distal gene, including on another chromosome - computationally tractable. For example, rs3924376 on chromosome 16 is a cis-eQTL in CD4+ T cells for SPNS1, and a trans-eQTL for MRPL32 on chromosome 7, with both genes involved in mitochondrial biology.
Here, we will present SAIGE-QTL results for the full phase 1 dataset, including set-based tests for rare variants captured by WGS. This updated map of common and rare variant effects from nearly 2,000 individuals offers unprecedented insight into genetic regulation at the single cell level. Integrating this with disease information from genome-wide association studies promise to reveal a deeper understanding of how each cell's function contributes to the genetic underpinnings of immune function and disease.

Population-scale single-cell RNA-seq across five countries reveal Asian-specific genetic architecture of alternative splicing and complex disease

Authors: B. Liu, C. Tian, Y. Zhang, Y. Tong; Natl. Univ. of Singapore, Singapore, Singapore

Abstract:

The nuances of human genomics are hindered by the fact that the most existing studies involves only European genetic ancestries. The lack of genetic diversity limits our fine-

grained understanding of population-specific genetic regulation and how these regulatory mechanisms contribute to disease. The Asian continent, home to 60% of the global population, boasts striking levels of genetic, phenotypic, linguistic, and cultural diversity. Here, we present the Asian Immune Diversity Atlas (AIDA), a multi-national single-cell RNA-sequencing (scRNA-seq) healthy reference atlas of human immune cells. AIDA comprises 1,265,624 circulating immune cells from 619 healthy donors, spanning 7 population groups across 5 countries, making it one of the largest and most diverse healthy blood datasets in terms of number of cells and population groups.We identified widespread sex-biased and ancestry-biased differential splicing events, with ancestry playing a significantly larger role than sex. A subset of ancestry-biased splicing events was driven by allele frequency differences. Splicing Quantitative Trait Loci (sQTL) analysis identified 11,577 independent *cis*-sQTLs, 607 *trans*-sQTLs, and 107 dynamic sQTLs whose allelic effects changed along the B cell developmental trajectory. In particular, colocalization between *cis*-eQTL and *trans*-sQTL revealed a cell-type-specific regulatory relationship between *hnRNPLL* and *CD45*. S-LDSC revealed a strong enrichment of *cis*-sQTL effects in autoimmune and inflammatory disease heritability. Specifically, we identified and experimentally validated an Asian-specific sQTL disrupting the 5' splice site of *TCHP* exon four, modulating the risk of Graves' disease in East Asian populations.AIDA provides fundamental insights into the relationships of human diversity with immune cell phenotypes, enables analyses of multi-ancestry disease datasets, and facilitates the development of precision medicine efforts in Asia and beyond. This resource will serve as a foundation for future studies investigating the complex interplay between genetics, immune function, and disease susceptibility across diverse populations.

Interindividual cellular and transcriptional regulatory changes in human aging

Authors: M. Matos[1,2], S. Ghatan[3], M. Suzuki[4], D. Reynolds[1], M. Isshiki[1], T. Thompson[1], R. Doña[1], K. Lundy-Perez[1], J. Stauber[1], A. Griffen[5], W. Oliveros Diez[6], S. Raj[1], T. Lappalainen[7], J. Greally[1]; [1]Albert Einstein Coll. of Med., Bronx, NY, [2]New York Genome Ctr., New york, NY, [3]New York Genome Ctr., New York, NY, [4]Texas A&M Univ., College Station, TX, [5]Albert Einstein Coll. of Med., San Diego, CA, [6]New York Genome Ctr., NEW YORK, NY, [7]SciLifeLab & NY Genome Ctr., New York, NY

Abstract:

Our goal is to understand how common genetic variation at noncoding regulatory regions influences susceptibility to age-associated cellular and molecular phenotypes. Polymorphisms at transcription factor (TF) binding motifs influence TF binding strength and affinity, ultimately affecting chromatin architecture and transcription, among other

molecular phenotypes. We hypothesize that age-associated cellular reprogramming and shifts in cell fate decisions are strongly influenced by the action of TFs, and that genetic variation significantly contributes to phenotypic differences in aging tissues. To test this, we integrated single cell RNA, bulk chromatin accessibility profiles, and genotypes from low-pass genome sequencing from CD4+ T lymphocytes derived from 364 Ashkenazi Jewish human donors ranging from 20-85 years old. First, we characterized the cell type composition and expression profiles from nearly 600,000 CD4+ T cells, discovering cell types and gene expression profiles that differ between the young and old age groups. Linking this to differential chromatic profiles revealed regulatory programs that associate to age in a cell-type specific manner, some of which have been implicated in age-related diseases. To understand effect of genetic variation across molecular layers, we mapped quantitative trait loci (QTL) associated with differential nearby gene expression (cis-eQTLs), chromatin accessibility (ca-QTLs) as well as inferred TF activity (TF-aQTLs). We discovered 10,939 cell type specific cis-eQTLs on pseudo-bulk expression of the three major CD4+ T cell subtypes and mapped their dynamic effects across continuous cell states. To infer TF activity from chromatin accessibility data, we fitted a linear regression model predicting binding strength of TF-binding motifs enriched within open chromatin regions. We then tested for genetic effects on TF activity and other chromatin accessibility traits (peak width, height) by mapping TF-aQTLs and context specific ca-QTLs. Finally, we observed colocalization of QTL variants GWAS loci of autoimmune diseases, with multiple layers of molecular data and cell type resolution unveiling functional heterogenous transcriptional regulatory mechanisms and cellular states influencing traits and disease in aging tissues.

Single-cell long-read sequencing analysis in endemic pemphigus foliaceus

Authors: T. Farias[1], V. Calonga-Solís[1], I. R. Wolf[1], G. A. Cipolla[2], D. Malheiros[2], M. Petzl-Erler[2], D. Augusto[1]; [1]The Univ. of North Carolina at Charlotte, Charlotte, NC, [2]Univ.e Federal do Paraná, Curitiba, Brazil

Abstract:

Pemphigus is a group of autoimmune skin diseases with a prevalence of 5.2 cases per 100,000 in the US. Pemphigus foliaceus (PF), a type of pemphigus, is endemic in certain regions of South America and Africa and reaches an astonishing prevalence of more than 3% in certain areas of Brazil, indicating a possible viral or environmental trigger. Single-cell transcriptomics enhances our understanding of cellular and molecular dynamics in health and disease. Coupled with long-read sequencing, it provides powerful information about splicing variants and differential transcript usage. Here, we employed cutting-edge single-cell RNA long-read sequencing on peripheral blood mononuclear cells in 8 individuals (4 PF

patients with active disease and 4 paired controls). Libraries were prepared using Parse Biosciences kits, and we used Oxford Nanopore technologies to sequence full-length transcripts. Quality control, data processing, and expression analysis were executed using established bioinformatics pipelines. Our analysis yielded 6,107 single cells, identifying nine major cell populations, including B cells, CD4+, and CD8+ T cells, and less common populations such as regulatory T cells. We identified differentially expressed genes in each cell type between patients and controls, which were later used to evaluate enriched biological processes using Gene Ontology (GO), KEGG (Kyoto Encyclopedia of Genes and Genomes), and Reactome databases. Two pathways were significantly overexpressed in patients' naïve $CD4^+$ T cells: i) NFκB signaling ($p^{corr} = 2x10^{-3}$), which regulates multiple aspects of innate and adaptive immunity, contributing to the pathogenic processes in inflammatory diseases. NFκB activation is critical for inflammatory responses against microorganisms and is a hallmark of most viral infections. ii) TNF signaling pathway ($p^{corr} = 3x10^{-4}$) is crucial in physiological processes such as proliferation, differentiation, apoptosis, immune response modulation, and inflammation induction. TNF signaling is important for autoimmune disease and is critical for protection against viruses; its overexpression could also result from viral infections. Finally, genes related to ribosome biogenesis ($p^{corr} = 2x10^{-3}$) were under-expressed in patients. Activation of ribosome biogenesis can produce cytokines and other immune mediators that are pivotal for antiviral responses. This study presents the first single-cell transcriptome analysis in this neglected disease affecting underrepresented populations, offering unprecedented insights that may contribute to developing more precise treatment and a better understanding of disease mechanisms.

Mapping the observable IBDverse: Identifying novel drivers of IBD susceptibility through population-scale, multi-tissue single-cell eQTL mapping

Authors: B. Harris[1], T. Alegbe[1], L. Ramirez-Navaro[1], M. Tutert[1], M. Krzak[1], M. Ozols[1], M. Ghouraba[1], M. Strickland[1], N. Wana[1], M. Hu[1], J. Ostermayer[1], R. McIntyre[1], C. Cotobal Martin[1], L. Fachal[1], G-R. Jones[2], T. Raine[3], C. Anderson[1]; [1]Wellcome Sanger Inst., Hinxton, United Kingdom, [2]Univ. of Edinburgh Ctr. for Inflammation Res., Edinburgh, United Kingdom, [3]Addenbrooke's Hosp., Cambridge, United Kingdom

Abstract:

Crohn's disease (CD) and ulcerative colitis (UC), the two most common forms of inflammatory bowel disease, are polygenic, immune-mediated conditions characterized by severe inflammation of the gastrointestinal tract. CD occurs most frequently in the

terminal ileum, while UC is most often found in the rectum. The etiology of both disease is incompletely understood and many cell types have been implicated in their pathogenesis. As IBD susceptibility loci are enriched in non-coding regions, previous efforts have sought to identify putative disease-causing, expression-linked genes (eGenes) by mapping expression quantitative trait loci (eQTL). These studies often rely on bulk gene expression from a single tissue of healthy samples, failing to holistically assess eQTL effects across the spectrum of disease status and cellular heterogeneity of all IBD-affected tissues. To better identify and characterize the universe of IBD driver genes, we created IBDverse - the largest collection of single-cell RNAseq data from the terminal ileum (nCD=142, nHealthy=310), rectum (nHealthy=330) and IBD patient blood (nCD=111) - a total of >2 million transcriptomes from 456 individuals. Generation of a multi-tissue cell atlas identified over 70 transcriptionally distinct clusters, spanning known stromal, immune and epithelial cell types. Genome-wide cis-eQTL mapping identified >16,000 eGenes (FDR<0.05), of which 40% were detected in a single tissue and 27% within a single cell-type annotation. Additionally, over 100 cis-eVariants were also associated with trans-gene expression (bonferroni-corrected FDR<0.05), >90% of which were detected in higher-level epithelial cell annotations. Importantly, colocalisation of eQTL effects with susceptibility GWAS nominated 214 genes as potentially causal drivers of CD or UC (PP.H4>0.8), spanning 108 (45%) of the 240 known susceptibility loci. A substantial proportion of these were detected in a single tissue (60%) or cell-type annotation (>70%), highlighting the importance of contextual characterisation in understanding the etiology of these diseases. Overall, this study serves as a step change in the understanding of cellular heterogeneity of these tissues, provides a compendium of genetic regulatory effects across constituent cell-types, and statistically quantifies the relevance of these effects to IBD susceptibility. This therefore showcases the additional value of eQTL mapping at higher resolutions, afforded by scRNAseq, and the potential for this to rapidly identify disease-effector genes at scale.

Single nucleus, multi-ancestry atlas of genetic regulation of gene expression in the human brain

Authors: B. Zeng[1], G. Hoffman[2], P. Roussos[3]; [1]Mount Sinai, New York, NY, [2]Icahn Sch. of Med. at Mount Sinai, New York, NY, [3]Panagiotis Roussos, New York, NY

Abstract:

Genetic risk variants for common diseases are often located in non-coding regulatory regions and act by modifying gene expression. Bulk tissue studies have yielded insight into

the mechanisms of shared genetic regulation affecting gene expression and disease risk. Yet disease mechanisms are cell-type specific, and characterizing the cell-type specificity of genetic regulation can yield further insight into molecular etiology. Here we present a large-scale atlas of genetic regulation of gene expression in the human brain from 6.3 million single nuclei of 1,494 donors. Due to the diverse ancestry of our cohort, including 443 non-European donors, our analysis has high resolution to perform statistical fine-mapping. We identify significant genetic regulation for 78.1% of genes tested containing 91.2% of eGenes detected in bulk, and analysis at multiple cellular resolutions from 8 classes, 27 cell types, and 64 subtypes reveals a hierarchy of cell type-specific regulatory effects. Colocalizing genetic regulatory effects with disease risk variants from genome-wide association studies of neuropsychiatric and neurodegenerative diseases identify risk genes, candidate causal variants, and the cell types they act in, highlighting excitatory neurons in Schizophrenia and microglia in Alzheimer's disease. We also observed genes, like EGFR, and CLU, whose expression in glial cells mediated their influence on Alzheimer's disease. Many of these are novel findings not previously identified in the analysis of bulk tissue. Our atlas of genetic regulation captures remarkable cell type specificity and offers novel mechanistic insight and testable molecular mechanisms underlying gene expression and disease risk.

Session 12: Beyond Genetic Discoveries: Novel Mechanisms of Neurodevelopmental Disorders

Location: Room 505

Session Time: Wednesday, November 6, 2024, 8:00 am - 9:30 am

Monoallelic *de novo* variants in DDX17 cause a novel neurodevelopmental disorder

Authors: E. Seaby[1,2,3], A. Godwin[4], V. Clerc[5], T. Fletcher[6], X. Grand[5], V. Meyer-Dilhet[5], L. Monteiro[5], DDX17 research consortium, D. Baralle[7], A. O'Donnell-Luria[2], H. Rehm[2], M. Guille[6], J. Courchet[8], C. Bourgeois[5], S. Ennis[1]; [1]Univ. of Southampton, Southampton, United Kingdom, [2]Broad Inst., Cambridge, MA, [3]Imperial Coll. London, London, United Kingdom, [4]Univ. of Portsmouth, European Xenopus Resource Ctr., Portsmouth, United Kingdom, [5]Univ. of Lyon, Lyon, France, [6]Univ. of Portsmouth, Portsmouth, United Kingdom, [7]Univ. of Southampton, Faculty of Med., Southampton, United Kingdom, [8]Univ. of Lyon, France, France

Abstract:

Introduction

DDX17 is an RNA helicase shown to be involved in critical processes during the early phases of neuronal differentiation. Globally, we identified 13 patients with neurodevelopmental phenotypes with *de novo* monoallelic variants in *DDX17*. All 13 patients had a neurodevelopmental phenotype, whereby intellectual disability, delayed speech and language, and motor delay predominated.

Materials and methods

We performed *in utero* cortical electroporation in the brain of developing mice, assessing axon complexity and outgrowth of electroporated neurons, comparing wild-type and Ddx17 knockdown. We then undertook *ex vivo* cortical electroporation on neuronal progenitors to quantitively assess axonal development at a single cell resolution. Homozygous and heterozygous *ddx17* crispant knockouts in *Xenopus tropicalis* were generated for assessment of morphology, performed behavioural assays, and neuronal outgrowth measurements. We further undertook transcriptomic analysis of neuroblastoma SH-SY5Y cells, to identify differentially expressed genes in DDX17-KD cells compared to controls.

Results

Knockdown of Ddx17 in electroporated mouse neurons *in vivo* showed delayed neuronal migration as well as decreased cortical axon complexity. Mouse primary cortical neurons revealed reduced axon outgrowth upon knockdown of *Ddx17 in vitro*. The axon outgrowth

phenotype was replicated in crispant *ddx17* tadpoles, including in a heterozygous model. Crispant tadpoles had clear functional neural defects and showed an impaired neurobehavioral phenotype. Transcriptomic analysis identified a statistically significant number of differentially expressed genes involved in neurodevelopmental processes in DDX17-KD cells compared to control cells.
Discussion
We have identified a new disease gene, *DDX17*, representing a rare cause of neurodevelopmental delay. We provide evidence for the role of the gene and mechanistic basis of dysfunctional neurodevelopment in both mammalian and non-mammalian species.

Maternally derived *de novo* variants in the non-coding spliceosomal snRNA *RNU4-2* are a frequent cause of syndromic neurodevelopmental disorders

Authors: Y. Chen[1], R. Dawes[1], H. Kim[1], A. Ljungdahl[1,2], S. L. Stenton[3,4], S. Walker[5], J. Lord[6], G. Lemire[3,4], A. C. Martin-Geary[1], V. S. Ganesh[3,4,7], J. Ma[3], J. M. Ellingford[8,5,9], E. Delage[10], E. N. Dsouza[1], S. Dong[1,2], RNU4-2 Consortium, J. L. Rubenstein[2], E. Markenscoff-Papadimitriou[2], S. M. Fica[1], D. Baralle[11,12], C. Depienne[13], D. G. MacArthur[14,15], J. M. M. Howson[16], S. J. Sanders[1,2], A. O'Donnell-Luria[3,4,17], N. Whiffin[1,3]; [1]Univ. of Oxford, Oxford, United Kingdom, [2]UCSF, San Francisco, CA, [3]Broad Inst. of MIT and Harvard, Cambridge, MA, [4]Boston Children's Hosp., Boston, MA, [5]Genomics England, London, United Kingdom, [6]Univ. of Sheffield, Sheffield, United Kingdom, [7]Brigham and Women's Hosp., Boston, MA, [8]Univ. of Manchester, Manchester, United Kingdom, [9]Manchester Univ. NHS Fndn. Trust, Manchester, United Kingdom, [10]Wellcome Sanger Inst., Hinxton, United Kingdom, [11]Univ. of Southampton, Southampton, United Kingdom, [12]Univ. Hosp. Southampton NHS Fndn. Trust, Southampton, United Kingdom, [13]Inst. für Humangenetik, UK Essen, Essen, Germany, [14]Garvan Inst., Darlinghurst, Australia, [15]Murdoch Children's Res. Inst., Melbourne, Australia, [16]Novo Nordisk Res. Ctr. Oxford, Oxford, United Kingdom, [17]Massachusetts Gen. Hosp., Boston, MA

Abstract:

Background: Around 60% of individuals with neurodevelopmental disorders (NDD) remain undiagnosed after comprehensive genetic testing, primarily of protein-coding genes. Increasingly, large genome-sequenced cohorts are improving our ability to discover new diagnoses in the non-coding genome.
Methods/results: Using a cohort of 8,841 probands with genetically undiagnosed NDD in

Genomics England (GEL), we identify the non-coding RNA *RNU4-2* as a novel syndromic NDD gene. *RNU4-2* encodes the U4 small nuclear RNA (snRNA), which is a critical component of the major spliceosome. We identify an 18 bp region of *RNU4-2* mapping to two structural elements in the U4/U6 snRNA duplex that is severely depleted of variation in the general population, but in which we identify heterozygous variants in 115 individuals with NDD across cohorts. This region is significantly enriched for variants in GEL NDD probands compared to individuals in the UK Biobank (OR=85.8; 95%CI:56.4-131.6; Fisher's $P=1.84x10^{-78}$). The majority of individuals with NDD (77.4%) have the same highly recurrent single base-pair insertion (n.64_65insT). Strikingly, for 54 individuals where we could determine the parent of origin of the identified de novo mutations, all 54 were on the maternal allele, pointing to a novel mutational mechanism. Individuals with *RNU4-2* variants have a severe NDD syndrome with global developmental delay, intellectual disability, speech abnormalities, microcephaly, hypotonia, short stature and seizures. Using blood RNA-sequencing data from five individuals, we show a systematic change in 5' splice site usage in individuals with *RNU4-2* variants compared to controls, consistent with the importance of this region of the U4 snRNA in correctly positioning the U6 ACAGAGA sequence to receive the 5' splice site. We demonstrate that *RNU4-2* is highly expressed in the developing human brain, in contrast to other U4 homologs, supporting *RNU4-2*'s role as the primary U4 transcript in the brain. Finally, we estimate that variants in this 18 bp region of *RNU4-2* explain 0.41% of individuals with NDD.
Conclusion: We identify variants in *RNU4-2* as one of the most common causes of NDD, underscoring the importance of non-coding genes in rare disorders. This work will provide a diagnosis to thousands of individuals with NDD worldwide and catalyse development of treatments for these individuals.

Biallelic inactivating variants in *DMAP1* underlie a syndromic neurodevelopmental disorder

Authors: D. Li[1], A. Sobering[2], DMAP1 Genetics Group, H. Hakonarson[1], Y. Song[3]; [1]Children's Hosp. of Philadelphia, Philadelphia, PA, [2]Augusta Unveristy / Univ. of Georgia Athens Med. Partnership, Athens, GA, [3]Children's Hosp. of Philadelphia, Philadelphia, PA

Abstract:

DNA Methyltransferase 1 Associated Protein 1 (*DMAP1*) encodes a versatile protein involved in different complexes responsible for maintenance of DNA methylation, DNA damage repair, regulation of histone acetylation and catalysis of exchange of histone H2A and H2A.Z. Despite DMAP1's essential roles in multiple transcriptional processes, it has

not been implicated in human disease. Through exome sequencing and subsequent reach out to the international matchmaking community, we identified 17 individuals from 15 families with a syndromic neurodevelopmental disorder carrying homozygous or compound heterozygous variants in *DMAP1*. Among these variants were three splice-altering or frameshift variants and nine missense variants residing in or around the SANT domain, suggesting they may affect interactions with DNA and/or histones. All 17 individuals have global developmental delay, intellectual disability, hypotonia, and craniofacial dysmorphisms, although the reported findings varied. Utilizing the Gal4-UAS system to perform neural-specific knockdown of the *Drosophila* ortholog *Dmap1*, we observed pupal lethality and structural defects in the mushroom body (MB), highlighting an underappreciated role of Dmap1 in MB development. These phenotypes can be rescued by the wild-type (WT) and the two missense variants of human *DMAP1*, allowed us to conduct the social space assay to further query the high order effect in social behavior. We found that WT DMAP1 restored proper fly distribution, whereas the two missense variants caused clustering and reduced inter fly distance, reflecting impaired social avoidance and conforming the pathogenicity of these missense variants. Additionally, WT DAMP1 but not the two missense variants partially rescued the seizures induced by mechanical stimulation in a sex dependent manner. Transcriptome analyses from *Drosophila* RNAi brains identified dysregulation of hundreds of genes implicated in transcription processing, neuronal function, and brain development. The decreased expression of two genes, *Cbl* and *SF1*, was confirmed by qPCR in *Drosophila* RNAi brains. Subsequently, overexpression of either *Cbl* or *SF1* in *Drosophila* RNAi could rescue lethality and MB defects, suggesting that *Cbl* and *SF1* are the potential effectors of DMAP1. In light of its involvement in DNA methylation, we performed an episignature analysis using blood-derived genomic DNA and identified a specific DNA methylation episignature. Taken together, we demonstrate that biallelic variants in *DMAP1* are associated with a novel neurodevelopmental disorder.

Loss-of-function of the Zinc Finger Homeobox 4 (*ZFHX4*) gene underlies a neurodevelopmental disorder

Authors: M. Pérez Baca[1,2], M. Palomares Bralo[3,4], M. Vanhooydonck[1,2], L. Hamerlinck[1,2], E. D'haene[1,2], S. Leimbacher[1,2], E. Jacobs[1,2], L. De Cock[1,2], E. D'haenens[1,2], Z. Malfait[1,2], L. Vantomme[1,2], F. Santos-Simarro[5], R. Lleuger-Pujol[6], S. García-Miñaúr[3,4], I. Losantos-García[7], B. Menten[1,2], A. Silva[8], K. Rooney[9], N. Ragge[10], ZFHX4 consortium, B. Sadikovic[8,9], B. Dermaut[1,2], E. Bogaert[1,2], D. Syx[1,2], S. Vergult[1,2], B. Callewaert[1,2]; [1]Ctr. for Med. Genetics, Ghent Univ. Hosp., Ghent, Belgium, [2]Dept. for Biomolecular Med., Ghent Univ., Ghent,

Belgium, [3]Inst. de Genética Médica y Molecular (INGEMM), Madrid, Spain, [4]ITHACA-European Reference Network, Madrid, Spain, [5]Unit of Molecular Diagnostics and Clinical Genetics, Hosp. Univ.ri Son Espases, Hlth.Res. Inst. of the Balearic Islands (IdiSBa), Palma, Spain, [6]Hereditary Cancer Program, Catalan Inst. of Oncology IDIGBI, Girona, Spain, [7]Dept. of Biostatistics, Hosp. Univ.rio La Paz, Madrid, Spain, [8]Dept. of Pathology and Lab. Med., Western Univ., London, ON, Canada, [9]London Hlth.Sci. Ctr., London, ON, Canada, [10]Clinical Genetics Unit, Birmingham Womens Hosp., Lavender House, Mindelsohn Way, Edgbaston, Birmingham, United Kingdom

Abstract:

The 8q21.11 microdeletions encompassing the transcription factor ZFHX4, have been associated with a syndromic form of intellectual disability, hypotonia, decreased balance and hearing loss. Here, we report on 55 individuals with protein truncating variants (n=34), (micro)deletions (n=20) or an inversion (n=1) affecting ZFHX4 with variable developmental delay and intellectual disability (85%), distinctive facial characteristics, morphological abnormalities of the central nervous system, short stature, hypotonia, and occasionally cleft palate and anterior segment dysgenesis. The phenotypes associated with the 8q21.11 microdeletions and ZFHX4 intragenic loss of function variants largely overlap identifying ZFHX4 as the main driver for the microdeletion syndrome, although leukocyte-derived DNA shows a mild common methylation profile in (micro)deletions patients only. We identified ZFHX4 as a nuclear protein that is increasingly expressed during human brain development and neuronal differentiation. In neural progenitor cells, ZFHX4 interacting factors suggest an important role for ZFHX4 in cellular and tissue developmental pathways, especially during embryonic and neural development. Accordingly, we observed that ZFHX4 interacts with the promoter regions of genes with crucial roles in embryonic, neuron and axon development. Since ZFHX4 loss-of-function associates with consistent dysmorphic features, we investigated whether the disruption of zfhx4 causes craniofacial abnormalities in zebrafish. First-generation (F0) zfhx4 crispant zebrafish, (mosaic) mutant for zfhx4 loss-of-function variants, have significantly smaller Meckel's cartilages and ethmoid plates compared to control zebrafish. Furthermore, behavioral assays showed a decreased movement frequency in the zfhx4 crispant zebrafish in comparison with control zebrafish larvae. Our in vivo work for zfhx4 suggests a role for facial skeleton patterning, palatal and neurodevelopment.

Biallelic *UGGT1* gene variants cause a congenital disorder of glycosylation

Authors: Z. Dardas[1], L. Harrold[2], C. Salter[3], T. Kikuma[4], K. P. Guay[5], B. Ng[6], K. sano[4], A. Saad[1], H. Du[1], R. Sangermano[7], S. Patankar, G[8], S. Jhangiani[9], S. Gürsoy[10], M. Abdel-

Hamid[11], M. Ahmed[12], R. Maroofian[13], R. Kaiyrzhavov[14], W. Jones[15], L. McGavin[16], M. Durkie[17], N. Wood[18], M. Zaki[19], H. Van Esch[20], J. Posey[1], O. Wenger[21], E. K. Scott[22], K. Bujakowska[23], R. Gibbs[9], D. Pehlivan[1], D. Marafi[24], J. Leslie[3], N. Ubeyratna[25], J. Costello[3], L. AlAbdi[26], F. Alkuraya[27], Y. Takeda[4], H. Freeze[6], D. N. Hebert[5], E. Baple[28], D. Calame[1], J. Lupski[1], A. Crosby[25]; [1]Baylor Coll. of Med., Houston, TX, [2]Level 4, RILD Wellcome Wolfson Med. Res. Ctr., RD&E (Wonford) NHS Fndn. Trust, Univ. of Exeter Med. Sch., Exter, United Kingdom, [3]Univ. of Exeter, Exeter, United Kingdom, [4]Ritsumeikan Univ., Shiga, Japan, [5]Univ. of Massachusetts, Amherst, MA, [6]Sanford Burnham Prebys Med. Discovery Inst., La Jolla, CA, [7]Massachusetts Eye and Ear Infirmary, Boston, MA, [8]Harvard Med. Sch., Boston, MA, [9]Baylor Coll. Med., Houston, TX, [10]S.B.Ü. Dr. Behçet Uz Children's Ed. and Res. Hosp., IZMIR, Turkey, [11]Natl. Res. Ctr., Giza, Egypt, [12]Natl. Res. Ctr., GIZA, Egypt, [13]Univ. Coll. London, London, United Kingdom, [14]Univ. Coll. London, LONDON, United Kingdom, [15]Great Ormond Street Hosp., London, United Kingdom, United Kingdom, [16]Univ. Hosp. Plymouth NHS Trust, Plymouth, United Kingdom, [17]Sheffield Children's NHS Fndn. Trust, Sheffield, United Kingdom, [18]Bradford Teaching Hosp. NHS Fndn. Trust, Bradford, United Kingdom, [19]Natl. Res. Ctr., Cairo, Egypt, [20]Univ. of Leuven, Leuven, Belgium, [21]New Leaf Ctr., Clinic for Special Children, Mount Eaton, OH, [22]The Univ. of Queensland, Brisbane, Australia, [23]Massachusetts Eye & Ear, Boston, MA, [24]Kuwait Univ., Jabriya, Kuwait, [25]Univ. of Exeter Med. Sch., Exeter, United Kingdom, [26]KFSHRC, RIYADH, Saudi Arabia, [27]King Faisal Specialist Hosp. & Res. Ctr., Riyadh, Saudi Arabia, [28]Univ. of Exeter, Exeter, Exeter, United Kingdom

Abstract:

Congenital disorders of glycosylation (CDG) comprise a large heterogeneous group of metabolic conditions due to defects in glycoprotein and glycolipid glycan assembly and remodelling, a fundamental molecular process with wide-ranging biological roles. Herein, we describe biallelic *UGGT1* variants in nineteen individuals (twelve unrelated families) of various ethnic backgrounds, as a cause of a distinctive CDG of variable severity, associated with normal isoelectric focussing of transferrin. The cardinal clinical features of *UGGT1*-related CDG involve severe developmental delay/intellectual disability [ID], seizures and craniofacial dysmorphology and microcephaly in the majority, with more severely affected individuals displaying congenital cardiac malformations, variable skeletal abnormalities including scoliosis, hepatic and renal involvement including polycystic kidneys mimicking autosomal recessive polycystic kidney disease. *UGGT1* encodes UDP-glucose:glycoprotein glucosyltransferase 1, an enzyme critical for maintaining quality control of N-linked glycosylation. Our functional molecularstudies implicate selected *UGGT1* variants as contributing to disease impairing UGGT1 catalytic activity, disrupting mRNA splicing, or inhibiting endoplasmic reticulum retention. Collectively, our data provide a comprehensive

clinical, genetic, and molecular characterization of *UGGT1*- related CDG, broadening the spectrum of N-linked glycosylation disorders. These findings not only enhance our understanding of this condition in these families but also facilitate its diagnosis and treatment globally, benefiting individuals affected by *UGGT1*- related CDG worldwide.

Unveiling the crucial neuronal role of the proteasomal ATPase subunit gene *PSMC5* in neurodevelopmental proteasomopathies

Authors: J. Stanton[1], S. Küry[2], G. van Woerden[3], T-C. Hsieh[4], C. Rosenfelt[5], M. Scott-Boyer[6], V. Most[7], T. Wang[8], D. Ortiz[9], A. Ziegler[10], R. Oegema[11], K. Steindl[12], K. McDonald[13], A. Dauber[14], S. Srivastava[15], C. Costin[16], J. Pappas[17], D. Niyazov[18], M. Malicdan[19], D. Babikyan[20], H. Al Saif[21], M. Hajianpour[22], G. Costain[23], D. Geneviève[24], T. Harel[25], A. Sabir[26], S. Weber[27], A. Ververi[28], A. Cueto-González[29], M. Rio[30], S. Jurgensmeyer[31], I. De Bie[32], S. Kamphausen[33], B. Isidor[34], PSMC5 Consortium, P. Hildebrand[35], E. Eichler[36], K. McWalter[37], P. Krawitz[38], A. Droit[39], Y. Elgersma[40], A. Grabrucker[41], F. Bolduc[42], S. Bezieau[43], F. Ebstein[44], E. Krüger[45]; [1]Univ. of Limerick, Limerick, Ireland, [2]Nantes Université, CHU Nantes, Service de Génétique Médicale, Nante, France, [3]Erasmus Med. Ctr., Rotterdam, Netherlands, [4]Univ. Hosp. of Bonn, Bonn, Germany, [5]Dept. of Pediatrics, Univ. of Alberta,, Edmonton, AB, Canada, [6]Ctr. de recherche du Chu de Quebec-Université Laval, Québec,, QC, Canada, [7]Inst. for Drug Discovery, Med. Faculty, Leipzig Univ., Leipzig, Germany, [8]Peking Univ., Beijing, China, [9]UPMC Children's Hosp. of Pittsburgh, Pitsburgh, PA, [10]Service de Génétique, CRMR AnDDI-Rares, CHU Reims, Reims, France, [11]UMC Utrecht, Utrecht, Netherlands, [12]Inst. of Med. Genetics, Univ. of Zürich,, Schlieren-Zurich, Switzerland, [13]Norton Children's Med. Group, Univ. of Louisville Sch. of Med., Louisville, KY, [14]Div. of Endocrinology, Children's Natl. Hosp. and Dept. of Pediatrics, The George Washington Univ. Sch. of Med. and Hlth.Sci., Washington, WA, [15]Rosamund Stone Zander Translational NeuroSci. Ctr., Boston Children's Hosp., Boston, MA, [16]Dept. of Genetics, Akron Children's Hosp., Akron, OH, [17]Clinical Genetic Services, Dept. of Pediatrics, NYU Sch. of Med., New York, NY, [18]Duke Univ. Sch. of Med., Chapel Hill, NC, [19]Med. Genetics Branch, Natl. Human Genome Res. Inst., NIH, Bethesda, MD, [20]Dept. of Med. Genetics, Yerevan State Med. Univ. after Mkhitar Heratsi, Yerevan, Armenia, [21]Dept. of Human and Molecular Genetics, Div. of Clinical Genetics, Virginia Commonwealth Univ. Sch. of Med., Richmond, VA, [22]Div. of Med. Genetics and Genomics, Dept. of Pediatrics, Albany Med. Coll., Albany, NY, [23]Div. of Clinical and Metabolic Genetics, The Hosp. for Sick Children, Toronto, ON, Canada, [24]Université Montpellier, Inserm U 1183, centre de référence maladies rares anomalies du développement, Service de génétique médicale, Hôpital Arnaud de Villeneuve, Montpellier, France, [25]Dept. of Genetics, Hadassah Med.

Organization, Jerusalem, Israel, [26]Dept. of Clinical Genetics, Lavender House, Birmingham Women's and Children's Hosp. NHS Fndn. Trust, Birmingham, United Kingdom, [27]CCA-AHU de génétique clinique et de neurogénétique, Service de Génétique et de Neurologie, CHU de Caen, Caen, France, [28]Dept. of Genetics for Rare Diseases, 'Papageorgiou' Gen. Hosp., Thessaloniki, Greece, [29]Univ. of Limerick, Limerick, Spain, [30]Service de Médecine Génomique des Maladies Rares, Hôpital Necker-Enfants Malades, Paris, France, [31]Div. of Genetics, Genomics and Metabolism, Ann & Robert H. Lurie Children's Hosp. of Chicago, Chicago, IL, [32]Div. of Med. Genetics, Dept.s of Human Genetics and Med., McGill Univ., Montreal, QC, Canada, [33]Inst. of Human Genetics, Univ. Hosp. Magdeburg, Magdeburg, Germany, [34]CHU Nantes, Nantes, France, [35]Inst. für Medizinische Physik und Biophysik, Univ. Leipzig, Medizinische Fakultät, Leipzig, Germany, [36]Univ of Washington, Seattle, WA, [37]GeneDx, 207 Perry Parkway, Gaithersburg, MD, [38]Univ. Bonn, Bonn, Germany, [39]CHU de Quebec Res. Ctr., Quebec, QC, Canada, [40]Erasmus MC, Rotterdam, Netherlands, [41]Bernal Inst., Univ. of Limerick, Limerick, Ireland, [42]Univ. of Alberta, Edmonton, AB, Canada, [43]Nantes Université, CHU Nantes, Service de Génétique Médicale, Nantes, France, [44]Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, Nantes, France, [45]Univ.smedizin Greifswald, Inst. für Medizinische Biochemie und Molekularbiologie, Greifswald, Germany

Abstract:

Neurodevelopmental proteasomopathies, a distinct subset of neurodevelopmental disorders (NDD) are driven by genetic variants in the 26S proteasome, a key complex contributing to the overall regulation of cellular protein homeostasis. In this novel study, 23 unique variants were uncovered in *PSMC5*, encoding the essential AAA-ATPase subunit PSMC5/Rpt6, resulting in syndromic NDD in 38 unrelated individuals. An accumulation of approaches and models were integrated to provide data from human cells, neuronal models, and animal studies, offering a multi-dimensional perspective on PSMC5-associated NDDs. Major findings have suggested these variants dramatically reshape human hippocampal neuron morphology and impair cognitive flexibility in Drosophila, in addition to resulting in a loss of excitatory synapses in rat neurons. Our findings further highlight the vulnerability of proteasomal subunit genes to genomic lesions, resulting in diverse and severe neurological phenotypes. Notably, *PSMC5* loss-of-function mutations lead to significant alterations in protein aggregation, derailing innate immune signalling, mitophagy, and lipid metabolism. A pivotal finding resulted from targeting key components of the integrated stress response, such as PKR and GCN2 kinases which ameliorated immune dysregulations in cells from affected individuals. Overall, these findings underpin the crucial role of *PSMC5* in neurodevelopment and draw compelling parallels with aging and neurodegeneration, suggesting broader therapeutic implications.

Session 13: Biobank Scale Genetic Data Resources for Studying Complex and Rare Human Diseases

Location: Mile High Ballroom 2&3

Session Time: Wednesday, November 6, 2024, 8:00 am - 9:30 am

100,000 Genomes of Europe: Unlocking genetic variability across Europe for science and health

Authors: A. Herzig[1], A. Vicente[2], H. Martiniano[2], E. Genin[1,3], H. Ray-Jones[4], J. van Rooij[4], A. Uitterlinden[4], GoE/1+MG consortium; [1]Inserm, Univ Brest, EFS, UMR 1078, GGB, Brest, France, [2]Inst. Natl. de Saúde Doutor Ricardo Jorge, Lisbon, Portugal, [3]CHRU Brest, Brest, France, [4]Lab. for Population Genomics, Erasmus MC, Rotterdam, Netherlands

Abstract:

Background/Objectives: As part of the 1 million genomes (1+MG) initiative, the Genome of Europe (GoE) project will create a pan-European reference database comprising whole-genome sequencing (WGS; 30x) data from >100,000 European citizens. Spanning 51 contributing partners from 29 countries providing existing and de novo datasets (short-read and long-read WGS), GoE presents enormous potential for understanding the genetic dimension to public health in Europe; including through interactions with the European Rare Disease Research Alliance (ERDERA). **Methods:** Challenges for GoE are evident across three axes: technological, statistical, and ethical, in order for GoE to provide wide-ranging utility. Here we present the scope of the “use-cases” of GoE, which will establish the envisaged applications of GoE. **Results:** The following pilot studies are outlined: (1) Building a fine-scale map of genetic structure across Europe to inform aggregating individual-level data. (2) Creating a reference database for individual variant look-ups; with particular attention on clinically relevant variant frequencies in actionable-disease and pharmacogenetic genes. (3) Building reference panels for ancestry-informed genetic imputation and the required secure informatics environments. (4) Deriving distributions of polygenic (risk) scores, calibrated for ancestry gradients in Europe; with a focus on cancer phenotypes in collaboration with the Can.Heal project. (5) Understanding the added value of long-read sequencing in the general population and exploring the ‘dark regions of the human genome’. **Conclusion:** We lay out here the methodological advances and adjustments necessary to best unlock the potential of GoE; particularly in the context of the prospective GoE federated data embedding in 1+MG.

The UAE Genome Program: Unique Genetic Insights from 43,608 Individuals

Authors: M. Olbrich[1], M. Mousa[1], I. Wohlers[2], A. Al-Aamri[1], A. Alsuwaidi[1], N. a. Marzouka[1], H. Alnaqbi[1], M. S. Alameri[3], S. A. Al-Marzooqi[3], W. M. Abdulrahman[3], D. Rute[4], J. Alazazi[5], T. Magalhaes[6], J. Mafofo[6], J. Quilez[6], M. Allam[7], M. S. Mohamad[7], N. Drou[8], Y. Idaghdour[9], R. Hamoudi[10], G. Tay[11], S. M. Ibrahim[1], F. Alkaabi[5], A. AlMannaei[3], H. Alsafar[1]; [1]Khalifa Univ. of Sci., Technology & Res., Abu Dhabi, United Arab Emirates, [2]Res. Ctr. Borstel, Borstel, Germany, [3]Abu Dhabi Dept. of Hlth., Abu Dhabi, United Arab Emirates, [4]Emirates ICT Innovation Ctr., Abu Dhabi, United Arab Emirates, [5]Genome Office, Abu Dhabi, United Arab Emirates, [6]M42, Abu Dhabi, United Arab Emirates, [7]United Arab Emirates Univ., Al Ain, United Arab Emirates, [8]New York Univ., Abu Dhabi, United Arab Emirates, [9]New York Univ. Abu Dhabi, Abu Dhabi, United Arab Emirates, [10]Univ. of Sharjah, Sharjah, United Arab Emirates, [11]Ctr. for Forensic Sci., The Univ. of Wes, Perth, Australia

Abstract:

The United Arab Emirates Genome Project (EGP) aims to comprehensively map Emirati nationals' genetic landscape. This study presents a detailed analysis of a subset comprising 43,608 individuals whose genomes were sequenced using Illumina technology. Our analysis identified 421,605,069 variants, 38% of which are previously unreported genetic variations. Of particular significance is the discovery that among the variants classified as common (43,491,009; 10%) within the studied population, 12% (5,296,683) were determined to be novel. While this cohort reflected a diverse ancestral background spanning European, Asian, and African populations, it also exhibited levels of homozygosity, particularly evident in long runs of homozygosity (ROHs). A significant difference ($p<0.001$) in the total lengths of ROHs was observed between third- and fourth-degree consanguineous marriages.

We observed high-impact variants with higher allele frequencies within our cohort than in global populations. These included rs532444320 in the *TMEM59* gene (0.02 vs. 0.001), rs753628430 in the *LACTBL1* gene (0.02 vs. <0.00001), and SNP rs775100038 in the *SOS1* gene (0.02 vs. 0.00043). This may be due to the high frequency of consanguineous marriages, as regions with ROH harbored high-impact variants within specific genes (*SLC22A1*, *ZAN*, *NPRL3*), representing deleterious alleles in 15-33% of the EGP cohort. Furthermore, our findings elucidated the association between consanguinity and chromosomal sexual and autosomal disorders, revealing that 90 out of 141 individuals (63.8%) with chromosomal disorders were consanguineously related, extending up to the fourth degree.

This subset represents the largest Middle Eastern cohort reported to date, providing an unprecedented opportunity to elucidate the genetic intricacies of the diverse Emirati

population. Additionally, our study integrates these findings, facilitating closer investigations into culture-bound kinship patterns. Overall, the EGP represents a significant improvement in the scale of genomics research within the Emirati population, paving the way for comprehensive understanding and targeted interventions in genetic health.

Structural variant discovery with GATK-SV in 97,940 short-read whole genomes from the *All of Us* Research Program

Authors: E. Pierce-Hoffman[1,2,3], M. Walker[1,2,3], C. Whelan[1,2,3], X. Zhao[1,2,4,5], R. Collins[1,2,6], S. Zaheri[1,2,3], K. Veeraraghavan[1,2,3], N. E. Kurtas[1,2], V. Jalili[1,2,3], All of Us Research Program, H. Brand[1,2,4,5], M. E. Talkowski[1,2,4,5]; [1]Program in Med. and Population Genetics, Broad Inst. of MIT and Harvard, Cambridge, MA, [2]Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [3]Data Sci. Platform, Broad Inst. of MIT and Harvard, Cambridge, MA, [4]Dept. of Neurology, Harvard Med. Sch., Boston, MA, [5]Stanley Ctr. for Psychiatric Res., Broad Inst. of MIT and Harvard, Cambridge, MA, [6]Div. of Med. Sci. and Dept. of Med., Harvard Med. Sch., Boston, MA

Abstract:

Structural variants (SVs), DNA rearrangements of ≥50 nucleotides, account for the majority of sequence divergence in the human germline and have been implicated in a variety of diseases. Scalable and accurate SV discovery from short-read whole-genome sequencing (srWGS) data has traditionally posed major technical challenges, limiting the integration of SVs into trait association studies.

We developed and applied GATK-SV to 97,940 srWGS samples from the *All of Us* Research Program (AoU). GATK-SV takes an ensemble approach to identify and jointly genotype SVs across samples with high sensitivity and specificity. The method resolves all classes of SVs accessible to srWGS, including 11 subclasses of complex SVs. We discovered 1,506,805 high-quality SVs in AoU, which represents the largest SV reference dataset in the field. This callset is highly sensitive, with a median of 9,686 SV sites per genome, and precise, with a confirmation rate of 88% in matched long-read data and an estimated false discovery rate of 3.4% for bi-allelic CNVs larger than 10 kilobases based on comparisons to microarrays. This callset comprises an estimated majority (53%) of individuals of non-European genetic ancestry that are historically underrepresented in genetics research. The high quality of this dataset showcases the accuracy and scalability of the GATK-SV method, which researchers can now easily apply to their own samples via a featured workspace in the Terra platform.

To demonstrate the utility of these SV data, we investigated whether SNVs and indels linked

to human traits via genome-wide association studies (GWAS) are in linkage disequilibrium (LD) with SVs that could be drivers of the association. We identified 3,258 common SVs (allele frequency [AF] >1% in AoU) in strong linkage disequilibrium ($R^2 \geq 0.8$) with at least one common SNV or indel (AF >1%) associated with at least one trait or disease in the NHGRI-EBI GWAS catalog. Importantly, 1,103 (34%) of the SVs in LD with a GWAS hit came solely from a non-European genetic ancestry group. Among these SVs were an inversion disrupting *ZNF257* associated with Type 2 diabetes in East Asians and new potential drivers including an *Alu* insertion in the promoter of *ANGPTL4*. This insertion is in LD with a GWAS hit for triglyceride levels, and after controlling for ancestry, sex, and age, we observed a significant (p=0.01) depletion of triglyceride levels in carriers (mean=115 mg/dL) vs. non-carriers (mean=124 mg/dL). Overall, these data demonstrate the exciting potential to incorporate sequence-resolved SVs into association studies and the importance of diversity in genetic analyses.

A complete telomere-to-telomere reference panel of 6404 human haplotypes improves imputation and phasing accuracy

Authors: J. Lalli[1], A. Bortvin[2], R. McCoy[2], D. Werling[3]; [1]Univ. of Wisconsin Madison, Madison, WI, [2]Johns Hopkins Univ., Baltimore, MD, [3]Univ. of Wisconsin-Madison, Madison, WI

Abstract:

The recent publication of the first complete human genome (T2T-CHM13) has opened up the possibility of assessing the impact of genetic variation in previously inaccessible regions of the genome. That promise cannot be fulfilled if T2T-CHM13 reference datasets are unavailable. For example, many statistical genetics techniques (imputation, phasing, array-based GWAS) require a reference dataset of haplotype-phased genomes. The 1000 Genomes Project (1kGP) has provided the largest, most diverse dataset of publicly accessible whole genome haplotypes in GRCh38 coordinates (“GRCh38 panel”). However, no such reference haplotype panel exists in T2T-CHM13 coordinates.
To meet this need, our group has utilized the SHAPEIT5 package to produce and benchmark a panel of 3202 phased whole genomes collected as part of the 1kGP from reads aligned to the T2T-CHM13 reference genome (“T2T panel”). Phasing was performed in a trio-informed and X chromosome ploidy-informed manner. Singleton variants were included in the final panel. As part of this effort, we have also produced and released the first telomere-to-telomere human recombination map.
To assess the accuracy of our phased panel, we took advantage of the fact that 39 of our

samples had been empirically phased by the Human Pangenome Reference Consortium (HPRC). These variants are available in GRCh38 and T2T-CHM13 coordinates. Using the HPRC haplotypes as a ground truth, we observed an empiric switch error rate (SER) of 0.434% in the GRCh38 panel and 0.363% in the T2T panel. This improvement in SER was especially apparent in chromosome X (GRCh38: 0.93% / T2T: 0.41%) and highly nonsyntenic chromosomes (e.g., chr21: 0.60% / 0.34%).
Using a T2T panel as a source of reference haplotypes when phasing pangenome variation resulted in an SER of 1.29%, a 0.94 percentage point reduction in switch error rate compared to the GRCh38 panel. Variation in regions nonsyntenic with GRCh38 had an SER of 3.65%. Imputation of Simons Genome Diversity Project variants downsampled to mimic the output of an Illumina Omni2.5 genotyping chip also benefited slightly from the use of the T2T panel (average T2T panel r^2: 0.9786 vs average GRCh38 panel r^2: 0.9754).
In summary, phasing and imputation appear to benefit from the reduction in reference error and increase in haplotype contiguity that is afforded by the T2T reference genome. We anticipate that the availability of this panel will facilitate and expedite more widespread utilization of the T2T-CHM13 reference. To that end, we have made this panel freely available on the CHM13 Github repository.

Diversity in the NHGRI-EBI GWAS Catalog: addressing disparities while promoting accessibility and data sharing

Authors: M. Cerezo[1], A. Abid[1], K. Bircan[1], P. Hall[2], S. John[1], N. Keller[2], E. Lewis[1], A. Mosaku[1], S. Ramachandran[1], E. Sollis[1], J. Morales[2], H. Parkinson[1], L. Harris[1]; [1]EMBL-EBI, Cambridge, United Kingdom, [2]NIH/NHGRI, Bethesda, MD

Abstract:

The NHGRI-EBI GWAS Catalog is a comprehensive updated repository of significant human GWAS findings, metadata and full genome-wide summary statistics (SumStats). To date, it contains 6868 publications, 101,577 GWAS (64.1% of them with SumStats) and 619,964 top associations. The GWAS Catalog provides detailed metadata on population descriptors that describe the samples giving rise to the GWAS results. Any descriptor the investigator uses is reported, along with an “ancestry label” indicating predicted or calculated similarity to a reference population. This provides important information for downstream analyses such as meta-analysis, fine mapping or development of polygenic scores, allowing researchers to harmonise between studies and monitor diversity and inclusion among study participants. We present an overview of the current landscape of ancestry group representation within the GWAS Catalog, highlighting efforts to enhance accessibility and

data sharing while addressing disparities in diversity. We show the different improvements in our website, harmonisation pipeline and community engagement. Despite multiple recent efforts, the European label is still overrepresented, being present in more than 63.2% of our studies and 92% of individuals. Acknowledging this imbalance, the GWAS Catalog has made efforts to promote the representation of all populations. This, for example, includes prioritising the curation of studies with ancestry labels other than European or working on flagging those studies with samples from several cohorts contributing to the European bias. A new feature this year is the inclusion of cohort metadata for each curated study. Including this data field has allowed us to determine the contribution of a small number of well-used cohorts to the observed bias. For example, the percentage of European-labelled individuals is reduced to 72.5% when UKBB studies are not included. Since 2020, we have encouraged authors to submit their data directly through our submission system even before the publication. We worked with the community to agree on a SumStats standard format that is also compatible with PLINK and makes data reusable. While the proportion of shared data from cohorts with non-European ancestry labels has grown in recent years, it remains low in comparison to the European group. Following the publication of the NASEM report on the use of population descriptors in genetics and genomics, our framework for describing the samples underlying GWAS is under review, and we welcome discussion with the community as to key metadata requirements for reuse and interoperability of future GWAS data.

Harmonizing the world's rare disease knowledge in Mondo

Authors: M. Munoz-Torres[1], **J. Berg**[2], M. Haendel[3]; [1]Univ. of Colorado Anschutz Med. Campus, Aurora, CO, [2]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [3]Univ. of North Carolina Chapel Hill, Brownsville, OR

Abstract:

Diagnosing and researching rare diseases is a global challenge hampered by fragmented, incompatible knowledge and data sources. Different communities worldwide utilize various sources in their diagnostic pipelines or to inform health policy decisions and research. This results in inequities in access to care, poorer diagnostic efficacy, and missed opportunities to find treatments. A large community of resources has worked together to overcome these challenges and create the Mondo Disease Ontology. Mondo unifies diverse disease knowledge sources into a coherent and interoperable resource, providing precise, computable, curated mappings with full provenance and attribution. Mondo is a comprehensive ontology that integrates rare disease knowledge and offers rich links to anatomical sites, genes, and phenotypes, enhancing our shared knowledge of rare

diseases and paving the way for more effective research and diagnosis. Our approach emphasizes a sustainable, community-driven, transparent process to integrate and sustain these global rare disease resources. Mondo's open-source terminological alignment system is robust and scalable, ensuring continuous synchronization and combining automated matching with collaborative curation workflows involving key stakeholders. Members of Orphanet, OMIM, MedGen, NORD, GARD, ClinGen, and many others, have collaborated to create Mondo, each contributing something different and synergistic: Orphanet, GARD, and NORD provide disease descriptions, diagnostic tests, orphan drug data, and information for patients and families. Exomiser leverages integrated phenotype data to identify pathogenic variants. ClinGen, OMIM, Monarch, and MedGen integrate genetic and phenotypic data, essential for understanding disease mechanisms and improving diagnostics. ICD is used worldwide by electronic medical record systems. By unifying diverse terminologies and providing rich connections to an enormous amount of data and knowledge, Mondo enhances the use of these independent resources while advancing rare disease analytics and mechanism discovery. This abstract is submitted on behalf of the Mondo Consortium; see a snapshot of the Mondo contributors at https://github.com/monarch-initiative/mondo/graphs/contributors.

Session 14: Cancer Risk: Novel Genes and Mechanisms

Location: Four Seasons Ballroom 4

Session Time: Wednesday, November 6, 2024, 8:00 am - 9:30 am

An Atlas of Pan-cancer Susceptibility Genes Revealed by Intronic Polyadenylation Transcriptome-wide Association Study

Authors: H. Chen[1], X. Zou[1], S. Zhang[1], T. Ni[2], L. Li[1]; [1]Shenzhen Bay Lab., Shenzhen, China, [2]Fudan Univ., Shanghai, China

Abstract:

Functional interpretation of human disease-associated non-coding variants remains challenging in the post-genome-wide association studies era. In our previous study, we identified 3'UTR alternative polyadenylation (APA) quantitative trait loci (3'aQTLs) and connected 3’UTR APA events with QTLs as a major driver of human traits and diseases. In addition to 3’UTR, APA has also been frequently identified in introns, which can lead to truncated mRNA or even truncated protein products. However, genetic determinants of intronic APA in various human tissues and their impact on human disease risk remain elusive. Here we described DiPars (Dynamics analysis of Intronic PolyAdenylation from multiple RNA-seq data) as the first computational method that infers dynamics intronic polyadenylation (IPA) events from population-scale RNA-seq data without relying on normal controls. DiPars demonstrates superior sensitivity and precision compared to traditional methods. We have applied DiPars to 24,078 RNA sequencing samples across 16,008 individuals and revealed a comprehensive list of genetically influenced IPA events. We further performed a pan-cancer intronic polyadenylation transcription-wide association study by integrating with 55 well-powered genome-wide association studies across 22 major cancer types. We identified 555 significant cancer susceptibility genes predicted to modulate cancer risk via intronic polyadenylation, 60.2% of which has been ignored by traditional gene expression, splicing, and 3’UTR-APA studies. Together, our study highlights the significant role of intronic polyadenylation in identifying new cancer susceptibility genes and provides a strong foundational framework for enhancing our understanding of the etiology underlying human cancers.

Exploring gene-by-environment interactions in colorectal cancer risk using massively parallel reporter assays ⭐

Authors: T. Fabo, R. Meyers, L. Kellman, Y. Zhao, S. Montgomery, P. Khavari; Stanford Univ., Stanford, CA

Abstract:

While GWAS is a powerful tool for identifying germline variants associated with risk for complex disease, they have fallen short in detecting gene-by-environment (GxE) interactions that contribute to disease. Colorectal cancer (CRC) is one such complex disease where GxE interactions remain understudied. Dietary metabolites such as butyrate, a fiber metabolite, and deoxycholate, a fat metabolite, have been shown to modulate CRC risk. In this work, we apply massively parallel reporter assays (MPRAs) to screen 2,265 CRC-associated SNVs and indels for differential transcriptional activity upon treatment with butyrate and deoxycholate as an experimental proxy for population-based GxE studies. MPRA identified 151 context-dependent differentially active SNVs (daSNVs) at FDR <0.1. Butyrate and deoxycholate tended to have opposite effects on transcriptional activity, suggesting these metabolites may modulate genetic risk through opposing regulatory mechanisms. One top scoring daSNV, rs6507875, exhibited both a butyrate and deoxycholate GxE effect, which respectively decrease and increase the allelic difference in transcriptional activity. rs6507875 is located near SMAD7, a previously nominated CRC GxE locus for diet, suggesting that our GxE approach can nominate population-level GxE effects. Integrating RNA- and ATAC-Seq data generated from metabolite-treated colon cells nominated transcription factors (TFs) mediating context-dependent regulatory effects at these variants. Motif analysis of metabolite-responsive elements found that butyrate “down” and deoxycholate “up” fragments had the strongest motif agreement, with AP-1 and Sp family motifs as top hits. rs72597431 is a daSNV exhibiting a deoxycholate-dependent effect that is predicted to disrupt binding motifs of multiple TFs, including ATF3, an AP-1 family member whose expression is also upregulated by deoxycholate. We are using EMSAs, luciferase assays, and CRISPRi experiments to understand the mechanistic underpinnings of MPRA hit SNVs and how they confer risk, including differential TF binding, context-dependent transcriptional activity, and impact on target genes. We are further validating our experimental approach to GxE discovery using population-scale biobanks to test MPRA hit loci for GxE effect enrichment. Our work has shown that butyrate and deoxycholate modulate functional activity of CRC risk variants; future work will interrogate the biological and epidemiological validity of these findings to nominate novel GxE loci for CRC.

Genetic regulation of *TERT* splicing contributes to reduced or elevated cancer risk by altering cellular replicative potential ⭐

Authors: O. Florez-Vargas[1], M. Ho[1], M. Hogshead[1], C-H. Lee[1], B. Papenberg[1], K. Forsythe[1], K. Jones[2], W. Luo[2], K. Teshome[2], C. Blauwendraat[3], K. J. Billingsley[3], M. Kolmogorov[4], M. Meredith[5], B. Paten[5], R. Chari[6], C. Zhang[2], J. Schneekloth[7], M. Machiela[8], S. Chanock[9], S. Gadalla[10], S. Savage[10], S. Mbulaiteye[11], L. Prokunina-Olsson[12]; [1]Lab. of Translational Genomics, Div. of Cancer Epidemiology and Genetics, Natl. Cancer Inst., Rockville, MD, [2]Cancer Genomic Res. Lab., Leidos BioMed. Res., Frederick Natl. Lab. for Cancer Res., Frederick, MD, [3]Ctr. for Alzheimer's and Related Dementias, Natl. Inst. of Aging and Natl. Inst. of Neurological Disorders and Stroke, Bethesda, MD, [4]Cancer Data Sci. Lab., CCR, Natl. Cancer Inst., Bethesda, MD, [5]UC Santa Cruz Genomics Inst., Santa Cruz, CA, [6]Genome Modification Core, Leidos BioMed. Res., Frederick Natl. Lab. for Cancer Res., Frederick, MD, [7]Chemical Biology Lab., CCR, Natl. Cancer Inst., Frederick, MD, [8]Integrative Tumor Epidemiology Branch, DCEG, Natl. Cancer Inst., Rockville, MD, [9]Lab. of Genetic Susceptibility, DCEG, Natl. Cancer Inst., Rockville, MD, [10]Clinical Genetics Branch, DCEG, Natl. Cancer Inst., Rockville, MD, [11]Infections and Immunoepidemiology Branch, DCEG, Natl. Cancer Inst., Rockville, MD, [12]Lab. of Translational Genomics, DCEG, Natl. Cancer Inst., Rockville, MD

Abstract:

Genome-wide association studies (GWAS) have identified multiple loci within chromosome 5p15.33 associated with reduced risk of some cancers but elevated risk of others. This region encodes the telomerase reverse transcriptase (TERT), which is critical in normal cells and carcinogenesis.
We investigated a locus marked by SNPs rs10069690 and rs2242652 within *TERT* intron 4, and identified a linked variable number tandem repeat within *TERT* intron 6 (VNTR6-1, 38-bp repeat unit, 24-66.5 repeat copies). In 544 phased long-read genome assemblies from 272 controls of diverse ancestries, we found more VNTR6-1 copies segregating with the rs10069690-T allele (p=2.53E-13) and rs2242652-A allele (p=9.52E-21) than with their alternative alleles. Based on the 1000 Genomes dataset, we constructed a custom imputation reference panel by adding the VNTR6-1 marker (Short allele, 24-27 copies and Long allele, 40.5-66.5 copies) inferred based on short-read whole-genome sequencing and targeted long-read PacBio sequencing.
rs10069690-T and VNTR6-1-Long alleles independently reduce *TERT* levels: rs10069690-T by increasing intron 4 retention and VNTR6-1-Long by expanding a G4 quadruplex in intron 6 (G4Q, 35-113 copies per allele). In UMUC3, a bladder cancer cell line, treatment with G4Q-stabilizing ligands decreased the ratio of the telomerase-producing full-

length *TERT* isoform, whereas VNTR6-1 deletion by CRISPR/Cas9 editing increased this ratio, cell growth and apoptosis.
To account for the combined effects of both variants on *TERT* splicing, we imputed a compound marker (V6.1rs100), with VNTR6-1-Short/Long and rs10069690-C/T alleles. In cancer-free individuals (UK Biobank, European ancestry, n=339,103), we observed a steeper age-related shortening of relative telomere length (rLTL) in peripheral blood leukocytes in those without the Short-C haplotype (marker, 5-years age groups interaction, p_{int}=1.39E-02). In a multi-cancer analysis (PLCO study, European ancestry, 73,085 cancer-free controls, and 29,623 cases), rs10069690-T, VNTR-6.1-Long, and the Long-T haplotype were comparably associated with reduced risk of cancers originating from tissues with low homeostatic proliferation, which maintains tissue self-renewal but high regenerative proliferation in response to environmental exposures and tissue damage (bladder and prostate). The same alleles were associated with elevated risk of cancers from tissues with no/low replicative potential (ovaries, thyroid, and brain). We conclude that the genetic regulation of *TERT* splicing by VNTR6-1 and rs10069690 contributes to differences in cancer risk by altering cellular replicative potential.

Immune surveillance and cancer risk

Authors: M. Saffern[1], C. Krishna[2], E. Olumuyide[1], E. Wilson[1], A. Tervi[3], H. Ollila[4], D. Chowell[1], R. Samstein[1]; [1]Mount Sinai, New York, NY, [2]Broad Inst. of MIT and Harvard, Cambridge, MA, [3]Inst. for Molecular Med. Finland (FIMM), Helsinki, Finland, [4]Inst. for Molecular Med., Finland, Univ. of Helsinki, Helsinki, Finland

Abstract:

Cancer risk factors include inherited mutations, environmental risks, and replication errors. These all point to mutational processes that can lead to cancer, but it is now appreciated that immune cells are constantly patrolling tissues for potentially transformed cells in a process known as cancer immunosurveillance. With the success of cancer immunotherapies, the idea that the immune system plays a role in cancer development has come to the forefront of cancer research, but the contribution of genetic variation in the immune system to risk of cancer is not well understood. Detection of foreign antigens presented on MHC molecules, encoded by the highly polymorphic and polygenic HLA locus, is critical to the interaction of cancer and the immune system. Since each HLA allele can bind a specific set of antigens, HLA heterozygosity confers a selective advantage against disease, which has been shown in the context of viral disease and immunotherapy. We hypothesize that germline heterozygosity at the HLA loci allows for binding of a greater repertoire of neoantigens that can be presented to the immune system, thereby decreasing

cancer risk. Using longitudinal clinical and genetic information from over 390,000 individuals in the UK Biobank and over 180,000 in FinnGen, we show that germline heterozygosity at HLA-II loci, but not HLA-I loci, is associated with decreased risk of lung cancer among smokers. Strikingly, this effect is independent of a genome-wide polygenic risk score (PRS) and HLA-II homozygosity increases lifetime risk of lung cancer in both current and former smokers with high PRS as compared to those with high PRS and HLA-II heterozygosity. Mechanistically, single-cell RNA sequencing analyses revealed that smoking increases pro-inflammatory macrophages in the lung, which express HLA molecules. We also show that loss of HLA-II on tumor cells is as prevalent in non-small cell lung cancer as loss of HLA-I, a known mechanism of immune escape, with preferential loss of alleles with larger neopeptide repertoires, highlighting the importance of HLA-II expression on epithelial cells. Ongoing studies in murine models of carcinogen-induced lung cancer are aimed at dissecting the relative roles of MHC-II on epithelial cells and immune cells as well as the overall importance of CD4 T cells in tumor immunosurveillance. These findings deepen our understanding of the role of the immune system in cancer risk, suggest the incorporation of HLA-II genotype in polygenic risk scores and stratification of HLA-II homozygous smokers for increased screening, and will inform immune-directed cancer prevention strategies.

Prevalence and effect of inherited chromosomally integrated human herpesvirus 6 in 735,434 human genomes

Authors: R. Dhindsa[1], L. Kida[2], F. Hu[3], O. Burren[4], Q. Wang[5], S. Petrovski[6], C. Lareau[7]; [1]Baylor Coll. of Med., Houston, TX, [2]Mem. Sloan Kettering Cancer Ctr.,, New York, NY, [3]AstraZeneca, Cambridge, United Kingdom, [4]Discovery Sci., R&D, AstraZeneca, Royston, United Kingdom, [5]AstraZeneca, Chapel Hill, NC, [6]Duke Univ, Durham, NC, [7]Mem. Sloan Kettering Cancer Ctr., New York, NY

Abstract:

Human herpesvirus 6 (HHV-6) is the only virus known to transmit through the human germline, which is mediated through chromosomal integration in telomeres. To date, the population prevalence of inherited chromosomally integrated HHV-6 (ici-HHV-6) and its impact on complex diseases have only been explored in small cohort studies. Leveraging comprehensive population-scale whole-genome sequencing (WGS) data, we examined occurrences of ici-HHV-6 in two large cohorts, the UK Biobank (UKB, *n*=490,040) and the United States-based All of Us (AOU; *n*=245,394) and associations with patient electronic health records (EHRs).We identified ici-HHV-6 individuals from these population

sequencing cohorts using WGS data, consistent with prior approaches.
In brief, unmapped sequencing reads were reprocessed and re-aligned to the HHV-6A and HHV-6B reference genomes using standard methods. Across both cohorts (n=735,434 individuals), we estimated an overall prevalence of ici-HHV-6 at 1.1% that varied between cohorts, observing that ici-HHV-6B was more common. Next, we conducted a phenome-wide association study (PheWAS) to identify associations between individuals with ici-HHV-6 and complex traits. Notably, we identified ici-HHV-6(B) with a strong association with basal cell carcinoma (BCC), the most frequently occurring form of cancer in the United States. We replicated the association between ici-HHV-6 and clinical screening for skin neoplasms in the AOU cohort (ICDX code: Z12.83 OR: 1.99; p-value: 2.4x10-6). In comparison to other risk alleles in the UKB, ici-HHV-6 represents the heritable allele with the single largest effect size. Our results corroborate prior work that detected recurrent HHV-6 viral DNA in BCC tumors but refines the association model where germline, rather than somatic, exposure predisposes individuals to BCC. Our identification of ici-HHV-6B as a common, heritable risk factor BCC motivates further work to understand the interplay between host genetics and the multifaceted exposure of this virus.

Session 15: Decoding Structural Variation at Scale

Location: Four Seasons Ballroom 1

Session Time: Wednesday, November 6, 2024, 8:00 am - 9:30 am

The contribution of linked structural variants to recent positive selection in humans

Authors: D. Radke, P. Hsieh; Univ. of Minnesota, Twin Cities, Minneapolis, MN

Abstract:

With genomic structural variants (SVs, such as large deletions and duplications that are >50 bp in length) historically being ignored in positive selection studies for technical reasons, there remains a question of the extent to which SVs may be underlying the signal of positive selection identified from single nucleotide variants (SNVs). To assess the contribution of SVs in positive selection in humans, we focused on putatively recent events to mitigate the complex history of human evolution. We devised a computational pipeline and identified highly differentiated SNVs in individual populations from the high-coverage short-read genomes of the 1000 Genome Project. We grouped these SNVs into contiguous loci based on strong linkage disequilibrium. Using this approach, we identified hundreds of loci in each population (median=488, s.d.=236), with a median length of 29,327 bp (s.d.=8,736 bp), including the previously reported selected inversion locus on chromosome 17q21.31 in Europeans. An overwhelming fraction of our identified loci overlap genes (median=59.8%, s.d.=3.9%), showing evidence consistent with the action of positive selection at highly functional regions. We quantified the extent of linked SVs to the highly differentiated SNVs at our identified loci using a published, high-quality SV dataset from the same collection. We estimate that the median fraction of these loci with linked SVs is 2.0% (s.d.=1.0%, range=0.8%-5.7%). Using a variety of simulations, we evaluate our estimates against expectations under null hypotheses and identify several SV candidates of recent positive selection in multiple populations. Contrary to the expectation of purifying selection removing SVs around genes, we do not find such a depletion in linked SVs to our candidate selected loci involving genic sequences (Fisher's exact test, two-sided p-value>0.05). Our findings highlight the potential role of SVs in recent positive selection in human populations, and that they may play an outsized role in positive selection loci relative to neutral expectations.

Haplotype-informed analysis of structural variation in 490,414 genomes and its effects on human health ⭐

Authors: S. Rubinacci[1,2], M. Hujoel[1,2], R. Mukamel[1,2], N. Kamitaki[1,2], D. Tang[1,2], P-R. Loh[1,2]; [1]Brigham and Women's Hosp. and Harvard Med. Sch., Boston, MA, [2]Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

Genome-wide association studies (GWAS) have uncovered numerous associations between genetic variants and human phenotypes. However, few such studies have considered structural variants (SVs), which can underlie SNP and indel associations and help identify molecular mechanisms and causal genes at GWAS loci.
To explore the phenotypic impact of structural variation throughout the human genome, we developed a new, haplotype-informed method to efficiently extract and denoise signatures of SVs from hundreds of thousands of whole-genome sequences. The approach works by (i) quantifying SV-informative signals (split reads, discordant reads, and read-depth) across all sequenced individuals; and (ii) leveraging haplotype-sharing to denoise these signals (similar to genotype refinement in low-coverage sequencing). Importantly, this approach generates cohort-wide SV genotype dosage estimates that can be directly tested for association with phenotypes, and it produces these data for both previously-ascertained SVs and common SVs discovered within the cohort. The approach is scalable, requiring only ~15 CPU-minutes of computation per genome.
We applied this method to analyze SVs in 490,414 UK Biobank (UKB) participants and 3,202 1000 Genomes Project (KGP) participants, using KGP data to evaluate SV-calling performance. We observed high genotyping agreement at SVs present in both our data set and a previous KGP SV call set (Byrska-Bishop et al.) (mean r^2=0.78 in GBR samples). Many SVs—in particular, multi-allelic tandem repeats—were present only in our data set, and even among biallelic SVs, we observed a 23% increase in discovery of long-read-validated SVs (based on high-coverage ONT data available for 7 GBR samples).
We performed a preliminary analysis of SVs on chromosome 20 (together with imputed SNPs and indels) for association with 57 quantitative traits in UK Biobank. After restricting to independent associations ($p < 5 \times 10^{-8}$ in conditional analyses), 55 of the 57 phenotypes had at least one significant SV association (totaling 570 associations), with SVs accounting for 7% of the top five associations per phenotype (21 SV associations out of 285). A haplotype containing a common 335bp Alu insertion polymorphism (AF = 89%) ~150kb upstream of *EDN3* was the lead association for both systolic and diastolic blood pressure ($p = 1.1 \times 10^{-38}$ and $p = 2.4 \times 10^{-53}$, respectively), suggesting that this SV may be a distal regulator of endothelin 3 expression, similar to the *EDN1 / PHACTR1* locus for vascular

diseases.
These results demonstrate the potential of haplotype-informed methods for SV analysis in GWAS, and we are now extending association analyses to genome-wide SVs.

A phenome-wide association study of tandem repeat variation in 168,554 individuals from the UK Biobank

Authors: C. Manigbas, B. Jadhav, P. Garg, M. Shadrina, W. Lee, G. Altman, A. Martin Trujillo, A. J. Sharp; Icahn Sch. of Med. at Mount Sinai, New York City, NY

Abstract:

SNP-based GWAS are a prevalent tool for identifying risk loci associated with human traits and diseases. However, most genetic association studies focus on binary variants, typically ignoring other variant types that could contribute to the "missing heritability" of the genome. We hypothesized that common length polymorphism of tandem repeats (TRs), a multi-allelic variant not assayed in standard genetic studies, can modulate phenotypic variation of human traits. To investigate this, we used direct genotypes for >50,000 TRs and performed phenome-wide association studies for >30,000 traits in 168,554 individuals of European ancestry with whole genome sequencing data from the UK Biobank.
We identified 47 TRs that showed causal associations with 73 traits based on fine-mapping and conditional analysis. We replicated 23 of 31 (74%) of these causal associations with matching traits in the All of Us cohort. While this set included several known repeat expansion disorders, novel associations we identified were attributable to common polymorphic variation in TR length, rather than rare expansions. Causal TRs were strongly enriched for functional repeats that impact local gene expression and DNA methylation, supporting their causality and providing insights into the molecular mechanism by which TRs impact the associated phenotype.Notable causal associations include:
1. A very highly polymorphic coding polyhistidine motif in HRCT1 with risk of hypertension ($p = 4.1 \times 10^{-24}$). While HRCT1 is a gene of unknown function, it shows highest expression in aorta. The associated TR also shows significant association with both expression level of HRCT1 ($p = 7.2 \times 10^{-25}$), and local DNA methylation levels ($p = 8.8 \times 10^{-12}$).
2. A poly(CGC) motif within the 5'UTR of GNB2 that associates with pulse rate ($p = 9.6 \times 10^{-14}$). Consistent with this finding, mutations in GNB2 cause sick sinus syndrome 4, a disorder characterized by atrioventricular conduction defects.
3. A poly(AC) motif in the 3'UTR of WNT9A associated with standing height ($p = 5.2 \times 10^{-16}$). WNT9A plays key roles in regulating synovial joint formation and chondrocyte differentiation.

Our analysis provides one of the first comprehensive evaluations of the impact of TR variation on human traits, highlighting TRs as a key candidate for the missing heritability problem and a crucial variant to consider in future genetic studies.

Contribution of Copy Number Variation in Disease Related Phenotypes Risk in 23andMe Research Cohort

Authors: S. Saini[1], J. Shi[1], 23andMe Research Team, A. Auton[2], S. Pitts[3], S. Shringarpure[2], X. Wang[2]; [1]23andMe, Inc., Sunnyvale, CA, [2]23andMe, Sunnyvale, CA, [3]23andMe, South San Francisco, CA

Abstract:

Copy Number Variants (CNVs) are one of the largest sources of all rare loss-of-function events and contribute to numerous genomic disorders including neuropsychiatric diseases. Due to the rarity of CNVs in general populations, and ability to ascertain only larger CNVs (tens of kilobases or longer) using traditional methods that utilize SNP-array data generated by large biobanks, phenotypic effects of these CNVs are largely understudied. We used a recently developed CNV detection method, Hi-CNV, to detect CNVs in 5 million European ancestry research participants from the 23andMe, Inc. cohort. After extensive QC and validation of the CNV calls, we observed 26 CNVs on average per sample, with a median CNV length of 8.0 kilobases. The resulting callset contained 80 million deletions and 60 million duplications across all samples. We further performed genome-wide association study analysis of several disease related binary phenotypes using SPAtest, a score-test-based method that uses saddlepoint approximation to estimate the distribution of the test statistic, across three association models: (1) the mirror model assessing the additive effect of each additional copy, (2) the duplication-only model assessing the impact of a duplication while disregarding deletions, and (3) the deletion-only model assessing the impact of a deletion while disregarding duplications. Upon testing more than 400 disease related phenotypes, we identified 3393, 2337 and 2575 genome-wide significant associations in the mirror, deletion-only and duplication-only models respectively, many of which were rare and high effect size associations. Due to the ultra rare frequency of many of the associations, 60% of these were not previously tagged by common SNPs. We confirm multiple well-established height-CNV associations such as *1q21.1* (Beta=2.5cm/CN, p<10^{-15}), *16p11.2* (Beta=3.2cm/CN, p<10^{-15}), *22q11.21* (Beta=0.6cm/CN, p<10^{-13}), and pLOF CNVs in *UHRF2* (Beta=0.7cm/CN, p<10^{-8}). We additionally identify several novel high effect height-CNV associations, including *DNM3* (Beta=4cm/CN, p<10^{-15}), *EFEMP1* (Beta=-4cm/CN, p<10^{-12}), and

IGF1 (Beta=9.7cm/CN, p<10^{-20}). We show the robustness of our CNV calling and association testing methodology by replicating associations linked to previously known disease associated genes, including *PCSK9*, *LDLR*, *APOA1*, *LIPC* for lipid traits, and *PRKN* deletions for Parkinson's disease. These results show the importance of studying rare CNVs in determining common disease susceptibility within the general population.

Pangenome-derived copy number variation maps with global diversity and association analysis in BioBank scale data with ctyper

Authors: W. Ma, M. Chaisson; Univ. of Southern California, Los Angeles, CA

Abstract:

Genetic analysis of copy number variations (CNVs), especially in complex regions, is challenging due to reference bias and ambiguous alignment of Next-Generation Sequencing (NGS) reads to repetitive DNA. Consequently, aggregate copy numbers are typically analyzed, overlooking variation between gene copies. Pangenomes contain diverse sequences of gene copies and enable the study of sequence-resolved CNVs. We developed a method, ctyper, to discover sequence-resolved CNVs in NGS data by leveraging CNV genes from pangenomes. From 118 public assemblies, we constructed a database of 3,351 CNV genes, distinguishing each gene copy as a resolved allele. We used phylogenetic trees to organize alleles into highly similar subgroups (HSS) that revealed events of linked small variants due to stratification, structural variation, conversion, and duplication. Saturation analysis showed that new samples share an average of 98.0% CNV alleles with the database. The ctyper method traces individual gene copies in NGS data to their nearest alleles in the database and identifies allele-specific copy numbers using multivariate linear regression on k-mer counts and phylogenetic clustering. Applying ctyper to 1000 Genomes Project (1kgp) samples showed Hardy-Weinberg Equilibrium on 99.2% of alleles and a 95.3% F1 score on genotypes based on 641 1kgp trios. Leave-one-out analysis on 40 assemblies matched to 1kgp samples showed that 95.9% of variants in query sequences match the genotyped allele. Genotyping 1kgp data revealed 226 population-specific CNVs, including a conversion of *SMN2 to SMN1*, potentially impacting Spinal Muscular Atrophy diagnosis in Africans. Our results revealed two models of CNV: recent CNVs due to ongoing duplications and polymorphic CNVs from ancient paralogs missing from the reference. To measure the functional impact of CNVs, after merging HSS copies, we conducted genome-wide Quantitative Trait Locus analysis on 451 1kgp samples with Geuvadis rRNA-seqs. Using a linear mixed model, our genotyping enables the inference of relative expression levels of paralogs within a gene family. In a global

evolutionary context, 470 out of 26,436 alleles were found with significant alternative gene expression ($p < 0.05$ after multiple test corrections), including 2.44% of novel gene duplicates, versus 1.43% of orthologs were differentially expressed, suggesting divergent expression from original genes. Specific examples include lower expression on the converted *SMN* and increased expression on translocated *AMY2B* (GTEx pancreas data). Our method enables large cohort studies on complex CNVs to uncover hidden health impacts and overcome reference bias.

Novel short tandem repeats on the Telomere-to-Telomere reference genome are associated with Alzheimer's disease neuropathology

Authors: A. Lee[1], B. Hu[1], J. Lu[1], D. Bennett[2], B. Vardarajan[1]; [1]Columbia Univ., New York, NY, [2]Rush Univ. Med. Ctr., Chicago, IL

Abstract:

Background: Alzheimer's Disease (AD) is characterized by genetic heterogeneity, and no single model fully explains its genetic inheritance. Short tandem repeats (STRs) which are hyper-mutable sequences in the human genome could account for some of the missing heritability in LOAD. STRs are implicated in several neuro-degenerative disorders. We systematically evaluated the impact of genome-wide STRs, sequenced from the Telomere-to-Telomere (T2T) and GRCh38 reference genomes, on neuropathological features of LOAD. **Methods**: We built the STR reference and genotyped STRs in whole-genome sequencing (WGS) data in 1,142 participants of from Religious Orders Study (ROS) and Rush Memory and Aging project (MAP) cohorts using our in-house algorithm LUSTR against the Telomere to Telomere consortium (T2T-CHM13) and GRCh38 reference genomes. We then tested the association of STRs with a) neuropathological LOAD status, b) clinical AD diagnosis, c) beta-amyloid levels, and d) neurofibrillary tau tangles (NFT). Regression models adjusting for age, sex and first three principal components. Subsequently, we examined if STRs regulated cis-gene expression in dorsolateral prefrontal cortex (DLPFC) after adjusting for age, sex and RNA-sequencing specific confounders. **Results**: We identified STRs in 176,624 loci using T2T-CHM13 and 168,928 loci using GRCh38 reference genomes. In T2T-CHM13 reference genome, average repeat count of TTAT repeat in *SIPA1L3* was associated with increased amyloid levels (p=6.60E-06). Variation in AC repeat in *ANKMY1* (p=8.43E-06), *TTTTTG* repeat in *GALNTL6* (p=1.55E-06), *TTCCT* repeat in *MYO1E* (p=2.84E-06) and *AAAC* repeat in *TMEM232* (p=3.78E-06) were associated with NFT and these repeats regulated cis-gene expression. We identified 1089 cis-eSTRs at genome-wide significance levels (p<1E-08). Amongst AD GWAS loci, 2,060 genes harbored

significant cis-eSTRs. In addition, amongst known pathogenic STRs affecting neurological conditions, CGG repeat in *FMR1* (Fragile X syndrome), and GAA repeat in *FXN* (Friedreich ataxia) significantly regulated cis-gene expression. **Conclusions**: We demonstrate that STRs influence the underlying gene expression in brain and are associated with neuropathological endophenotypes of AD. This suggests that STRs could explain some of the missing heritability in LOAD. Additional future directions include investigating STRs associated with single-nucleus RNA sequencing and protein expression in brain.

Session 16: I See Ghosts: Archaic DNA in Our Genomes

Location: Room 405

Session Time: Wednesday, November 6, 2024, 8:00 am - 9:30 am

A refined analysis of Neanderthal-introgressed sequences in modern humans with a complete reference genome

Authors: L. Chen[1], S-A. Liang[1], T. Ren[2], J. Zhang[1], J. He[1], X. Wang[2], X. Jiang[2], Y. He[1], R. McCoy[3], Q. Fu[4], J. Akey[5], Y. Mao[2]; [1]Fudan Univ., Shanghai, China, [2]Shanghai Jiao Tong Univ., Shanghai, China, [3]Johns Hopkins Univ., Baltimore, MD, [4]Chinese Academy of Sci., Beijing, China, [5]Princeton Univ., Princeton, NJ

Abstract:

Leveraging long-read sequencing technologies, the first complete human reference genome, T2T-CHM13, corrects the assembly errors in prior references and addresses the remaining 8% of the genome. While the studies on archaic admixture in modern humans so far have been only relying on the GRCh37 reference due to the version of the archaic genome data, the impact of T2T-CHM13 in this field remains unknown. Here, we refined the analysis of alignment and genetic variant calling of the high-quality Altai Neanderthal and Denisovan genomes in GRCh38 and T2T-CHM13 respectively. Compared with GRCh37, we found T2T-CHM13 has a significant improvement of read mapping quality in the archaic samples. We then applied IBDmix to identify the Neanderthal introgressed sequences in 2,504 individuals from 26 geographically diverse populations based on different genome references. Compared to GRCh38, we discovered ~51 Mb T2T-CHM13-unique Neanderthal sequences. In particular, these sequences are predominantly located in the genomic regions where the variants that are different between the GRCh38 and T2T-CHM13 assemblies emerge. Besides, we replicated, refined and observed novel instances of population-specific archaic introgression in diverse populations with T2T-CHM13, covering genes involved in metabolism, olfactory-related, and icon-channel, such as *FUT8*, *OR14A16*, and *KCNK2*. Finally, we integrated the introgressed sequences and adaptive signals with all reference genomes into a visualization database website, called ASH (www.arcseqhub.com), which facilitates evolutionary and genomic studies associated with archaic alleles in modern humans. Our work highlights that utilizing the T2T-CHM13 reference can provide novel insights of determining variation in archaic ancestry and further elucidating its functional, phenotypic and evolutionary significance in archaic admixture studies.

Patterns of genomic and morphological variation in the mid-19th century burial remains of the Liberated Africans from St. Helena Island in the South Atlantic

Authors: G. Johnson[1], A. Lisi[2], J. Liu[2], A. Galloway[3], M. Madrona[4], M. Hjorth[4], A. Ramsøe[4], C. Gaunitz[4], F. Demeter[4], E. Willerslev[4], C. M. Lee[1], M. C. Campbell[2]; [1]Howard Univ., Washington, DC, [2]Univ. of Southern California, Los Angeles, CA, [3]Univ. of California Santa Cruz, Santa Cruz, CA, [4]Univ. of Copenhagen, Copenhagen, Denmark

Abstract:

The remote island of St. Helena, located nearly 1,250 miles off the west coast of Angola, is home to a rare assemblage of skeletal remains of formerly enslaved Africans dating to the mid-19th century. Between 1840 and 1872, during the British Royal Navy's campaign to abolish slavery, over 26,000 indigenous Africans on seized slave ships were diverted to St. Helena Island where they were eventually liberated by the Vice-Admiralty courts and became known as the "Liberated Africans of St. Helena Island." During the construction of St. Helena's first airport, the skeletal remains of some of the Liberated Africans were unearthed. More specifically, burial sites containing ~8,000 individuals were found in an area of the island known as Rupert's Valley. This discovery provides a unique opportunity to gain further insights into the biological history of people who were targets of the Transatlantic slave trade. To unravel the geographic origins of the Liberated Africans as well as their patterns of genomic and phenotypic variation, petrous bones, teeth, metacarpals and metatarsals were collected from 100 individuals. We then extracted DNA from the petrous bones of 92 individuals and applied next-generation sequencing methods to generate sequence data. We further integrated these data with publicly available genomic data from >3,000 Africans from across the continent. In addition, we examined the dental morphology of the Liberated Africans to estimate their age and assess their overall health status. Based on our analyses, we inferred that the Liberated Africans originated from a geographic region intermediate between West Central and South Africa. Furthermore, we uncovered a striking pattern of population structure in the Liberated Africans and evidence of admixture from surrounding populations. We also explored patterns of adaptive evolution using a haplotype-based approach and observed long-range haplotypes around derived alleles across the genome, some of which are associated with several complex traits (such as cardiovascular and respiratory diseases as well as inflammatory bowel disease). Finally, our morphological analysis of the dental remains revealed that individuals in our sample were under the age of 20 years, and intriguingly some of them exhibited dental modifications. We also determined that the health status of our Liberated Africans was high. Overall, this study enhances our understanding of the Transatlantic slave trade

based on biological information, giving voice to those individuals who can no longer speak for themselves.

Deciphering genetic contribution of ancient hunter-gatherer Jomon in Japanese Populations

Authors: K. Yamamoto[1], S. Namba[2], K. Sonehara[2], K. Suzuki[2], N. P. Cooke[3], the Biobank Japan Project, K. Matsuda[2], T. Gakuhari[4], T. Yamauchi[2], T. Kadowaki[5], S. Nakagome[6], Y. Okada[7]; [1]Osaka Univ., Suita, Japan, [2]The Univ. of Tokyo, Tokyo, Japan, [3]Max Planck Inst., Leipzig, Germany, [4]Kanazawa Univ., Kanazawa, Japan, [5]Toranomon Hosp., Tokyo, Japan, [6]Trinity Coll. Dublin, the Univ. of Dublin, Dublin, Ireland, [7]Osaka Univ. / The Univ. Tokyo / RIKEN, Osaka, Japan

Abstract:

While advancements in paleogenomics have shed light on the ancestral admixture of modern humans and its influence on present-day traits, detailed insights remain elusive in non-European populations. Understanding these ancestral contributions is crucial for deciphering the genetic basis of complex traits and diseases. In Japan, the long-standing model of the origin of its people has been a dual ancestral structure based on morphology, with indigenous Jomon hunter-gatherers and continental ancestors from East Asia. However, the recent ancient DNA study proposed a tripartite ancestral structure for the genetic origin of modern Japanese, including additional Northeast Asia source. In this study, we conducted biobank-scale analyses to investigate this tripartite model in diverse Japanese populations. We analyzed Biobank Japan (BBJ; n = 171,287) as a dataset of modern Japanese individuals incorporating genetic data from ancient Japanese and Eurasian genomes (n = 22). Our analyses supported the applicability of the tripartite ancestral model to Japanese populations throughout the Japanese archipelago, with variations in ancestral proportions. The Individual proportions of Jomon ancestry strongly correlated with genetic diversities determined by principal components ($|R|$ = 0.61, $P < 1.0\times10^{-300}$ in PC1), indicating a substantial contribution of Jomon hunter-gatherer to the population structure of today. In terms of phenotypic impact, the positive association between Jomon ancestry and body mass index (BMI) suggests the contribution of ancient hunter-gatherer component to an increased risk of obesity in modern humans. Genome-wide association analysis with rigorous adjustments for geographical and ancestral substructures identifies 132 genetic markers showing evidence of recent natural selection pressure (P = 0.008). These variants tag significantly longer haplotypes than those from non-selected variants, suggesting their origin from Jomon ancestry. The predictive power of

these variants for individual Jomon ancestry was validated using independent Japanese cohorts (Nagahama cohort, *n* = 2,993; the second cohort of BBJ, *n* = 72,695). We then used the prediction model to detect individuals with Jomon ancestry in a separate East Asian population from UK Biobank (*n* = 200) and replicated the positive association between Jomon ancestry and BMI (Beta = 2.2, *P* = 0.03). Our extensive analysis of over 250,000 ancient and modern genomes provides valuable insights into the genetic contributions of ancient hunter-gatherers in contemporary populations.

Characterize the nature of ghost archaic introgression in African populations

Authors: H. Wang, S. Sankararaman; UCLA, Los Angeles, CA

Abstract:

Although DNA from archaic hominins has yet to be recovered in Africa, multiple lines of evidence have pointed towards the existence of a ghost introgression event in the history of African populations. Recent studies have noted the possibility of an archaic origin of African unique variations; however, this hypothesis depends on the ghost admixture event occurring after the out-of-Africa event. Using analysis of the conditional site frequency spectrum (CSFS) on both African and non-African populations, and by using an approximate Bayesian computation framework, we confidently timed the ghost admixture event to have occurred before the most recent out-of-Africa event. To further investigate the fine-scale distribution of archaic ghost ancestry in African and non-African populations, we introduce ArchIE2, a robust reference-free local archaic ancestry inference model extending the capabilities of its predecessor, ArchIE. Compared to ArchIE, ArchIE2 addresses ArchIE's issues by reducing the feature set to ten curated elements, overcoming challenges related to changing sample sizes and overfitting due to correlated features. Comparative analysis confirms the robustness of ArchIE2 across a spectrum of parameters, from varying mutation and recombination rates, to differing population histories. Leveraging the San population of Africa as the outgroup, ArchIE2 generated comprehensive genome-wide maps of archaic ancestries across diverse African and non-African populations. We observed a significant positive correlation between the archaic ancestry fraction vectors of African and non-African populations, suggesting that the ghost introgression event predates the out-of-Africa migration, which is consistent with our CSFS results.

Indirect inheritance of archaic ancestry in modern Peruvians

Authors: S. Gutierrez; Univ. of Michigan, Ann Arbor, MI

Abstract:

Archaic introgression events introduced novel genetic variation to modern human lineages, evidently shaping modern human population history and genetic diversity. The genetic ancestry of modern Native American populations can be modeled as an admixture between Asians via shared ancestry, Europeans via post-colonial contact, and Africans via shared modern human ancestry and potential recent admixture. The ancestry contribution of non-Africans indirectly introduced archaic genetic variants to the South American gene pool. Consequently, archaic loci have been captured in modern Americans despite the expected absence of direct contact with archaic humans. Here, we perform a comprehensive scanning of positive selection in Peruvians that revealed 20 loci with strong evidence of adaptive introgression in Peruvians by both Neanderthal and Denisovan. We leveraged whole-genome sequences from modern, ancient, and archaic humans to trace the allele frequency trajectory of archaic alleles in candidate adaptive introgression regions. We applied comprehensive selection scans and ARG methods to describe the source population, the timing of first introduction, and the strength and timing of selection within modern Peruvians. These methods reveal archaic variants of Denisovan-origin inherited from ancestral East Asian populations, and others of Neanderthal-origin introduced upon contemporary admixture with Europeans, both of which are under positive selection within modern Peruvians. Through this work, the selection history of archaic variants is informative of the key timelines of the initial settlement of Peruvians and their recent admixture history, which help refine the demographic model for modern Peruvians. Further, this study underscores the ability of archaic introgression to introduce standing variation that is ultimately inherited by populations without direct contact, that then becomes the basis for adaptation in contemporary environments long after initial introduction to the modern human gene pool.

Archaic introgression in Samoans: population structure, genetic admixture, and health associations

Authors: C. Liu[1], P. F. Reilly[1], R. L. Minster[2], D. E. Weeks[2,3], S. T. McGarvey[4,5], T. Naseri[6], S. Viali[7,8], R. Polimanti[9,10,11], S. Tucci[1,12]; [1]Dept. of Anthropology, Yale Univ., New Haven, CT, [2]Dept. of Human Genetics, Sch. of Publ. Hlth., Univ. of Pittsburgh, Pittsburgh, PA, [3]Dept. of Biostatistics, Sch. of Publ. Hlth., Univ. of Pittsburgh, Pittsburgh, PA, [4]Intl. Hlth.Inst., Dept. of Epidemiology, Brown Univ. Sch. of Publ. Hlth., Providence, RI, [5]Dept. of Anthropology,

Brown Univ., Providence, RI, [6]Naseri & Associates Publ. Hlth.Consultancy Firm & Family Hlth.Clinic, Apia, Samoa, [7]Oceania Univ. of Med., Apia, Samoa, [8]Dept. of Epidemiology (Chronic Disease), Yale Univ. Sch. of Publ. Hlth., New Haven, CT, [9]Dept. of Psychiatry, Yale Univ., New Haven, CT, [10]Cooperative Studies Program Clinical Epidemiology Res. Ctr. (CSP-CERC), Veteran Affairs Connecticut Hlth.care System, West Haven, CT, [11]Wu Tsai Inst., Yale Univ., New Haven, CT, [12]Dept. of Ecology and Evolutionary Biology, Yale Univ., New Haven, CT

Abstract:

Samoans represent a distinctive case among human groups due to their history of complex migration and multi-way admixture leading to diverse ancestral sources of archaic introgression. However, our understanding of the genetic structure of Samoans and its potential impact on health outcomes is still limited. Here, we leveraged 1,250 high-coverage Samoan whole genome sequences generated by the Trans-Omics for Precision Medicine (TOPMed) Program to characterize genetic variation across Samoa, to investigate signatures of Denisovan and Neanderthal introgression and recent admixture events, and to detect population-specific associations with anthropometric traits and cardiometabolic outcomes. By comparing Samoan genomes with 1,000 Genomes Project (1KG) reference populations, we observed strong genetic similarities among Samoan participants enrolled from rural and urban areas, and identified evidence of gene flow in Samoan genomes originating from East Asians and very recent admixture with European populations. We then generated genome-wide maps of Neanderthal and Denisovan introgression using SPrime and found that Samoans harbor more archaic introgressed genomic sequence (median 130.7 Mbp per individual) than worldwide populations included in the 1KG dataset (median 63.5-106.9 Mbp per individual). Furthermore, we explored the impact of archaic introgression on anthropometric and cardiometabolic traits in modern Samoans, and identified three associations with diastolic blood pressure. Of these, associations at the *NRG1* and *CDRT8* loci are on Neanderthal introgressed haplotypes, and the association at the *LINC02052* locus is on a Denisovan introgressed haplotype. These archaic introgressed haplotypes at *NRG1*, *LINC02052*, and *CDRT8* arose on genetic backgrounds of European, Oceanic, and East Asian ancestry, respectively, in line with the complex demographic history contributing to modern Samoan genomes. Overall, our study provides new insights into the population history of Samoans and our understanding of the genetic basis of human phenotypic variation.

Session 17: Creative Community Engagement: Gathering Data for Better Participatory Research

Location: Room 405

Session Time: Wednesday, November 6, 2024, 10:15 am - 11:45 am

From Barbershop to Biopsy: Improving Access to Genetic Screening through the Cleveland African American Prostate Cancer Project

Authors: F. Schumacher[1], K. Austin[2], R. D. Miller[2], A. Goldenberg[1], H. Hoban[3], A. D. Zimmer[3], C. L. Neben[3], **E. S. Trapl**[1]; [1]Case Western Reserve Univ., Cleveland, OH, [2]Case Comprehensive Cancer Ctr., Cleveland, OH, [3]Color Hlth., Burlingame, CA

Abstract:

Genetic screening has been shown to be beneficial in detecting potential genetic risk factors for disease, however, marginalized groups have low participation due to lack of access and awareness, mistrust and distrust of the hospital system, care navigation, and reluctance to participate. There is a need for intentional collaboration with community partners to ease concerns, and allow community voices to actively participate in the design and implementation of programs.
In the Cleveland African American Prostate Cancer Project (CAAPP), academics, clinicians, industry leaders, and trusted community partners developed a community-level engagement plan pairing genetic screening and testing with a prostate cancer (PrCa) screening program. CAAPP partners with local barbershops, which serve as community hubs, to train barbers as lay health advisors on providing Black men ≥40 access to PrCa and genetic screening. These “town-hall” style discussions, called Listening Tours, invited residents to discuss their experiences with local health care and provide suggestions for enhancing PrCa and genetic screening outreach. The main themes from the listening tour centered on the need for relationship building, education, and community-based outreach to reach those who lack access to screening information, generating the following expressed needs: (1) an infographic communicating relevant findings in an accessible way; (2) return of result sessions; (3) education sessions from a local Black physician discussing the importance of screening; (4) tailored barbershop training based on community need and gaps in knowledge; (5) educational materials; (6) in-shop videos; and (7) a network of participant support from the point of screening through follow-up care. By leveraging our existing clinical and institutional relationships and partnering with an industry leader of genetic testing, Color Health, we were able to address concerns focused on accuracy and reliability, cost and resource allocation, and data ownership. Color provided

comprehensive genetic counseling thus limiting a common obstacle of healthcare systems. When paired with the expertise of our Community Navigators and participating barbers, we were able to design a program that provides end-to-end support across CAAPP, our clinical partners, and Color genetic counselors.
These initiatives can increase trust, awareness, and participation in genetic screening and other preventive health measures. By addressing these concerns comprehensively, genetic screening in a community setting can be implemented in a manner that maximizes benefits while minimizing risks and ethical dilemmas.

Co-Creating a story-based video collection to engage LGBTQIA+ community members with the *All of Us* Research Program: An engagement marketing and human entered design approach

Authors: J. Uhrig[1], A. Jordan[1], D. Puckett[1], K. Baker[1], C. Johns[1], A. Corbo[1], J. Brown[1], D. Moretti[2], A. Rescate[2], M. A. Lewis[1]; [1]RTI, Research Triangle Park, NC, [2]PRIDEnet, Stanford Univ. Sch. of Med., Stanford, CA

Abstract:

The *All of Us* Research Program is a national effort to drive innovations in biomedical research and precision medicine. Engaging participants from diverse backgrounds is a core value of the program. We are using engagement marketing and human centered design principles to co-create digital solutions with community members to support engagement with *All of* Us. We conducted eight problem validation and solutioning workshops with 48 LGBTQIA+ community members. Community members validated barriers to engagement or enrollment in a research program like *All of Us*, including lack of awareness, representation of LGBTQIA+ communities in existing assets, and intentional outreach to LGBTQIA+ communities; concerns about privacy and data security, distrust, participant burden; and unclear value or benefit related to participation. Workshop participants brainstormed and generated 47 ideas for potential digital solutions to overcome barriers. We developed the ideas for potential solutions into 27 concepts (descriptive text and visual storyboards) and assessed acceptability, appropriateness, and feasibility in a set of 10 concept testing workshops with 57 community members. The highest rated concept to increase engagement with *All of Us* was a video series about how diversity in research matters (mean receptivity 4.37 out of 5). We developed a story-based video collection that is community-driven and research-based to support engagement efforts. After filming testimonials with community members and researchers, we reviewed the transcripts and raw footage to map video content back to the barriers identified by the community to

develop concepts for potential videos. We held three virtual workshops with advisory group and LGBTQIA+ community members to assess the extent to which the videos address the barriers; feature strategies, tactics and best practices recommended by the community; and are accessible, appropriate, engaging, and address short-term outcomes. Most workshop participants indicated the videos were designed to reach the LGBTQIA+ community, featured diverse members of the LGBTQIA+ community, featured personal stories, and will raise awareness of *All of Us*. Workshop participants were less likely to indicate that the videos made clear who is eligible to participate in *All of Us* or showed how easy it is to participate. Across all workshops, community members indicated that the process of engaging them demonstrated integrity, competence, dependability, trust, and collaboration; fostered a sense of connection to *All of Us*; and will enhance future engagement with *All of Us*. Future videos will address barriers not fully addressed in the current collection.

Breaking Barriers: Project GIVE's Tele-Genetic Initiative for 100 Children with Rare Diseases at the Texas-Mexico Border

Authors: B. Vuocolo[1], R. Sierra[1], D. Brooks[1], A. Iness[1], J. Chang[1], K. Carter[2], K. Rodriguez[3], L. Berry[4], A. Hernandez[3], C. Holder[1], L. Urbanski[1], J. Gamez[5], S. Mulukutla[6], S. Lizardo[7], H. Hidalgo[3], A. Allegre[3], J. Bernini[8,9], S. Magallan[3], S. Rodriguez[3], J. Gibson[3], H. Dai[1], C. Soler-Alfonso[1], B. Lee[1], S. Lalani[10]; [1]Baylor Coll. of Med., Houston, TX, [2]UT Hlth.San Antonio, San Antonio, TX, [3]UT at Rio Grande Valley, Edinburg, TX, [4]UT at Rio Grande Valley, McAllen, TX, [5]Driscoll Children's Hosp. Rio Grande Valley, Edinburg, TX, [6]DHR Hlth., Edinburg, TX, [7]Mercedes Children's Clinic, Mercedes, TX, [8]Vannie Cook Children's Cancer Clinic, McAllen, TX, [9]Texas Children's Hosp., Houston, TX, [10]Baylor Coll. Med., Houston, TX

Abstract:

Identifying genetic diagnoses in children is crucial for improving health outcomes, yet genomic disparities exist for individuals who are non-White and/or Hispanic or those with lower socioeconomic status. The Rio Grande Valley (RGV) along the Texas-Mexico border lacks consistent access to a full-time genetics provider. Over 94% of the RGV population identifies as Hispanic/Latino, and between 30-40% of children in the four counties live in poverty. Many families travel between 150-350 miles to access genomic services. Project GIVE (Genetic Inclusion by Virtual Evaluation) is an NIH-funded research study at Baylor College of Medicine that leverages Consultagene, a cutting-edge telehealth platform, to provide timely virtual genetic evaluation and whole genome sequencing (WGS) to children in the RGV with suspected genetic diseases. Physicians, developmental therapists, and

other healthcare professionals in the RGV can refer patients directly to Project GIVE through the Consultagene portal. Families accepted into the study meet with the study's bilingual research coordinator at the local study site for a virtual genetic evaluation with the BCM genetics providers ("Visit 1"). Buccal samples for CAP/CLIA WGS are sent to Baylor Genetics and return of results counseling is provided ("Visit 2). Patients are longitudinally followed for 1 year. Clinical Sequencing Evidence-Generating Research (CSER) surveys are utilized at all study visits to collect demographic information and assess study outcomes. Project GIVE received 224 Consultagene referrals from 21 community partners between February 2022 and January 2024, surpassing its target of 100 referrals. Of these, 184 (~82%) were accepted in the study, and 78 families have completed Visit 1. Most families identify as Hispanic/Latino (98%), and 82% of families live below 200% of the federal poverty line. WGS results have been returned to 72 families. 27 children received a diagnosis or partial diagnosis (37%). Of the children who received a diagnosis, 59% had changes to their medical management. We are exploring potential new gene-disease associations for three of our participants with negative WGS results. Preliminary results from surveys show that families feel satisfied with the use of telemedicine for the genetics evaluation and return of results. Findings from Project GIVE support the use of virtual genetic evaluation to improve access to genomic services in underserved pediatric populations with rare diseases. We believe that our model of integrating community engagement and using an advanced virtual platform can be replicated in other under-resourced areas to improve genomic health of children.

Assessing diverse communities' perspectives of precision research participation: the Precision rEsearCh pArticipatioN (PECAN) study

Authors: J. A. Williams[1], B. J. Wolf[1], S. T. Karim[2], L. A. Ueberroth[2], L. H. Moultrie[3], Q. Quet[4], R. Werner[2], M. A. Cunningham[2], D. L. Kamen[2], C. G. Allen[1], P. S. Ramos[1,2]; [1]Med. Univ. of South Carolina, Dept. of Publ. Hlth.Sci., Charleston, SC, [2]Med. Univ. of South Carolina, Dept. of Med., Charleston, SC, [3]Lee H. Moultrie and Associates, North Charleston, SC, [4]Gullah/Geechee Nation, St. Helena Island, SC

Abstract:

Historically marginalized groups remain underrepresented in research, hindering the generalization of results and the achievement of health equity in precision medicine. This lack of participation might be due to unique cultural, socioeconomic, or physical barriers. The objective of the *Precision rEsearCh pArticipatioN* (PECAN) study is to identify factors that positively or negatively influence perceptions about precision health research

participation among a diverse sample of participants living in the coastal Low Country of South Carolina, with the goal of identifying actionable steps for improving understanding and reducing barriers to participation. Between December 2023 and April 2024, 154 participants completed a 20 min survey to assess: 1) familiarity with precision medicine, 2) perceptions of factors associated with personal health, 3) knowledge, values, and beliefs about genetics research, 5) values considered important when deciding to participate in a research study, and 6) importance of various factors to be considered in a policy on precision health research. Descriptive statistics were determined and responses by race, sex, age, and education were compared using Chi-square tests. Most participants self-identified as African American or Black (45.5%) and white (40.3%), as female (70.8%), with ages younger than 46 (57.1%, vs. 41.6% ≥ 46), and as not a college graduate (53.9%, vs. 45.5% college graduate). The proportion of participants familiar with precision medicine terminology was higher for college graduates compared to non-college graduates, for females compared to males, and for participants 46 or older compared to younger than 46 ($P < 0.05$). When deciding to participate in a research study, Black individuals reported higher levels of importance for: (i) donating to non-profit vs for-profit researchers ($P < 0.05$), (ii) individual, familiar, or community benefits ($P < 0.01$), (iii) access, time, biospecimen, and compensation ($P < 0.01$), and (iv) privacy and confidentiality ($P < 0.01$). Black respondents also placed a higher importance on community engagement in developing a policy on precision health research ($P < 0.01$). Results from the PECAN study are expected to inform culturally-centered interventions that increase precision research participation and lead to a reduction in health disparities for all the diverse populations of South Carolina and beyond.

Genetic Services in Africa: Evidence-Based Recommendations for Policymakers and Healthcare Organizations ⭐

Authors: K. Kengne Kamga, P. Marlyse; Limbe Regional Hosp., Limbe, Cameroon

Abstract:

The incorporation of genetic services into African healthcare systems is a complex endeavor with numerous challenges and potential avenues for progress. This study aims to provide evidence-based recommendations tailored to policymakers and healthcare stakeholders, in order to facilitate the seamless integration of genetic services into African healthcare systems.
To conduct this study, we utilized a comprehensive scoping review methodology, meticulously examining a corpus of peer-reviewed studies spanning from 2003 to 2023.

These studies were sourced from prominent databases including PubMed, Scopus, and Africa-wide repositories. Our analysis focused on eight seminal research studies conducted between 2016 and 2023, each addressing a spectrum of genetic issues across six African nations: Cameroon, Kenya, Nigeria, Rwanda, South Africa, and Tanzania.
The synthesized findings from the reviewed studies underscore a myriad of challenges impeding the widespread implementation of genetic services across African healthcare systems. These challenges encompass deficiencies in disease awareness and education, logistical barriers to genetic testing, resource constraints, ethical dilemmas, and complexities related to follow-up and patient retention. However, the studies also illuminate promising opportunities and strategies conducive to effective integration, emphasizing proactive measures such as robust community engagement initiatives, targeted advocacy efforts, and the cultivation of supportive networks within local healthcare ecosystems.
In conclusion, while the integration of genetic services in Africa presents significant potential for improving healthcare outcomes, it is not devoid of challenges. However, these challenges also serve as catalysts for innovation and present fertile ground for growth. Healthcare and biotechnology enterprises are encouraged to seize upon these opportunities by investing in educational initiatives, forging strategic partnerships with local institutions, and harnessing the power of digital platforms to disseminate information and foster dialogue. Moreover, the establishment of online forums represents a pivotal step towards fostering collaboration and knowledge exchange within the African healthcare landscape.

Balancing Constitutional Protections in Genetic Research: Addressing Concerns of Minoritized Communities in Non-consented Tissue Reuse

Authors: T. Dye[1], N. Cardona Cordero[2], Z. Quiñones Tavarez[1], I. Rivera[1], J. Menikoff[3]; [1]Univ. of Rochester Sch. of Med., Rochester, NY, [2]Univ. of Puerto Rico Comprehensive Cancer Ctr., San Juan, Puerto Rico, [3]Natl. Univ. of Singapore, Singapore, Singapore

Abstract:

Introduction. The U.S. Constitution is increasingly used to decide cases where the public feels their rights have been violated regarding repurposing their DNA in research. Understanding public concerns related to Constitutional issues provides insight about genetic research (“GR”) hesitancy and Constitutionally-derived rights and limits in research tissue reuse. **Methods.** Our nested, analytical cross-sectional study of social determinants of GR included participants age 21+ with ancestral origins in Latin America

and the Caribbean ("Hispanic," n=1,370), non-Hispanic whites (n=270), and non-Hispanic other races (n=77), residing in the USA, the jurisdiction for which the U.S. Constitution governs. We created a "DNA Constitutional Concerns Score" (D-CCS) to reflect community concerns that could map to Constitutional protections. The D-CCS includes attitudes toward data privacy, protection of family history, data storage, genetic discrimination, misuse of genetic information, deprivation of freedom, and invalid consent. We measured interest in GR participation, using DNA for purposes other than the original study, and sharing DNA with private companies using 4-point Likert scales. We examined D-CCS with demographic, sociocultural, and research participation variables. **Results.** The additive scale ranged from 0 (no concerns) to 8 (all concerns), with a Cronbach's alpha=0.75, indicating good internal validity. D-CCS was significantly correlated with race and ethnicity, highest among Hispanics (mean=3.9; 95% CI: 3.9, 4.0) and lowest among non-Hispanic whites (mean=3.4; 95% CI: 3.4, 3.6), with non-Hispanic other races in between (mean=3.6; 95% CI: 3.0, 4.2). D-CCS was inversely correlated with intention to participate in GR (r=-0.35; p<0.001), allowing biospecimens to be used for other purposes (r=-0.36; p<0.001), and sharing data with private companies (r=-0.41; p<0.001). D-CCS also correlated positively with Victoroff's Oppression Score (r=0.11; p<0.001) and the Perceived Stress Score (r=0.06; p=0.02). **Discussion. M**inoritized communities are concerned about issues that map to potential violations of their Constitutional rights, no doubt informed by the lived experience and cultural knowledge of past research abuses. These concerns suggests the need for evaluation of legal protections and transparency in how genetic data is collected, stored, and used in research. Discriminatory practices based on genetic data could unequally affect minoritized groups. These concerns about Constitutional protections are significant barriers to participation in GR and addressing them may strengthen trust between researchers and communities.

Session 18: Machine Learning and AI Applications in Human Genetics

Location: Four Seasons Ballroom 2&3

Session Time: Wednesday, November 6, 2024, 10:15 am - 11:45 am

CellPhenoX: An eXplainable Cell-specific machine learning method to predict clinical Phenotypes using single-cell multi-omics

Authors: J. Young, J. Inamo, F. Zhang; Univ. of Colorado Denver Anschutz Med. Campus, Aurora, CO

Abstract:

As the scale of single-cell datasets expands, the task of linking cell phenotypic alterations with relevant clinical phenotypes becomes increasingly complex. Here, we introduce CellPhenoX, an eXplainable machine learning method to identify Cell-specific Phenotypes that influence clinical phenotypes. CellPhenoX generates low dimensional representations of single-cell multi-omics data as features for classification models (e.g., random forest) to predict clinical phenotypes of interest. To quantify the contribution of each cell to the clinical phenotype, CellPhenoX generates a predictive score for each cell based on SHAP (SHapley Additive exPlanations) values while accounting for covariates. Then, CellPhenoX identifies the salient low dimensions and calculates interpretable scores with the discriminative power for the model prediction alongside the corresponding markers. We benchmarked XCell on simulated data, which achieved an average AUC of 0.7. We applied CellPhenoX on a COVID proteomic dataset with >412,800 cells (PMID: 33879890), which uncovered an activated monocyte phenotype whose expansion level increases with disease severity from mild, moderate, to severe COVID, adjusting for technical batch, sex, age, and disease duration. Furthermore, CellPhenoX is able to identify an inflammatory fibroblast phenotype unique to inflamed gut compared with non-inflamed (cross-validated AUC 0.82) using an ulcerative colitis single-cell transcriptomic dataset. In summary, we demonstrate that CellPhenoX presents superior performance by providing interpretable scores for the pathogenic cell phenotypes that are key predictors to distinguish multi-class clinical variables and are associated with interaction effects in large single-cell datasets. We expect CellPhenoX to be powerful in determining clinic-related cell phenotypes in single-cell datasets with multiple sources of variation, including complex clinical groups, and technical and demographic variables. CellPhenoX promises broad applicability to numerous single-cell sequencing modalities, including exome/whole genome sequencing and ATAC-seq, making it relevant to genetic research interests.

scPrediXcan: leveraging single-cell data for transcriptome-wide association studies at cell-type level through transfer learning

Authors: Y. Zhou[1], T. Adeluwa[2], S. Gona[1], S. Sumner[1], L. Zhu[1], F. Nyasimi[1], R. Madduri[3], M. Chen[1], H. Im[1]; [1]Univ. of Chicago, Chicago, IL, [2]The Univ. of Chicago, Chicago, IL, [3]Argonne Natl. Lab., Naperville, IL

Abstract:

Transcriptome-wide association studies (TWAS) have been key in identifying genes linked to complex traits and diseases but often fail to pinpoint disease mechanisms at the cellular level. Whereas TWAS approaches use tissue-level prediction models, our association method (scPrediXcan) employs a deep learning model, scPred, trained on a reference genome and single-cell RNAseq data from 51 cell types across three datasets. scPred is a lightweight multi-layer perceptron that uses an existing sequence-to-epigenomics model (Enformer) as a feature extractor and predicts the gene expressions at single-cell pseudobulk level. scPred learns gene regulatory grammar (i.e., intra-genome variance) typically ignored by linear models, and it predicts gene expression with high accuracy: Pearson correlations R=0.75-0.89 in all 51 cell types tested. Moreover, scPred surpasses current linear TWAS gene expression prediction models in predicting pseudobulk expression levels across individuals when comparing the absolute Pearson correlations. Next, we generated in-silico personalized gene expressions using genotype data from 1000G dataset and linearized the deep learning models to enable computationally feasible TWAS using GWAS summary statistics. We compared the performance of scPrediXcan and the canonical TWAS method, PrediXcan, in type 2 diabetes (T2D) and systemic lupus erythematosus (SLE). We used the curated T2D gene set from Common Metabolic Diseases Knowledge Portal as a silver standard T2D gene list to calculate power (i.e., recall) of each method; scPrediXcan (power=0.429, precision=0.107) outperformed PrediXcan trained on two datasets separately: GTEx bulk (power=0.122, precision=0.108) and T2D pseudobulk (power=0.112, precision=0.115). For SLE, scPrediXcan identified 176 candidate causal genes in T cells distributed at 23 genomic loci, whereas PrediXcan trained by GTEx whole blood only identified 54 candidates at 14 genomic loci. Our results suggest scPrediXcan not only identifies more putative causal genes than traditional TWAS methods, but also explains more GWAS loci and gives cell-type-specific insights. Overall, our results demonstrate that scPrediXcan offers a significant advance in the precision and relevance of gene nominations in TWAS, promising to deepen our understanding of the cellular mechanisms underlying complex diseases.

MILTON: Disease prediction with multi-omics and biomarkers empowers case-control genetic discoveries in UK Biobank

Authors: M. Garg[1], M. Karpinski[1], D. Matelska[1], L. Middleton[1], O. S. Burren[1], F. Hu[1], E. Wheeler[1], K. R. Smith[1], M. Fabre[1,2], J. Mitchell[1], A. O'Neill[1], E. A. Ashley[3], A. Harper[1], Q. Wang[4], R. S. Dhindsa[5,4], S. Petrovski[1,6], D. Vitsios[1]; [1]AstraZeneca, Cambridge, United Kingdom, [2]Univ. of Cambridge, Cambridge, United Kingdom, [3]Stanford Univ., Stanford, CA, [4]AstraZeneca, Waltham, MA, [5]Baylor Coll. of Med., Houston, TX, [6]Univ. of Melbourne, Melbourne, Australia

Abstract:

The emergence of biobank-level datasets with rich phenotype and multi-omics data offers new opportunities for human disease prediction and novel biomarker discovery. Here, we present MILTON, an ensemble machine-learning framework based on XGBoost classifiers, that utilizes a range of blood, urine and proteomics-based biomarkers to predict 3,213 diseases available in UK Biobank (UKB). Leveraging UKB's longitudinal health record data, MILTON could predict incident disease cases that were undiagnosed at time of recruitment. Firstly, using 67 blood and urine-based biomarkers, MILTON achieved an AUC>=0.70 for predicting 1,091 ICD10 codes across cohorts of four different ancestries (with 384 ICD10 codes achieving AUC>=0.80). These models largely outperformed UKB standard polygenic risk scores in disease prediction ($AUC_{67\ traits}$ vs AUC_{PRS} = 0.71±0.12 vs 0.66±0.07), except for a few phenotypes, such as breast cancer, prostate cancer and melanoma. Remarkably, inclusion of ~3,000 UKB protein expression-level data in MILTON prediction models considerably boosted performance for 52 ICD10 codes (ΔAUC>=0.1). In a separate analysis, we validated that originally undiagnosed individuals (at time of recruitment), who were later diagnosed for a given disease in future UKB phenotype refreshes, were preferentially predicted by MILTON as putative novel cases (median odds ratio=3.27, p<0.05). Next, we generated augmented case-control cohorts, including MILTON novel predicted cases with already diagnosed ones, to check for putative novel associations in phenome-wide and exome-wide association studies (PheWAS/ExWAS) across 484,230 whole genome sequences. This resulted in improved signals for 88 known ($p<1x10^{-8}$) gene-disease relationships alongside 182 gene-disease relationships that did not achieve genome-wide significance in the ICD10-based baseline cohorts. We also observed 9,882 putative novel variant-disease associations from ExWAS on MILTON augmented cohorts out of which 61.94% had support from baseline cohorts (p<0.05). Furthermore, in our pan-ancestry PheWAS analysis on protein truncating variants, we identified 9 associations that did not achieve genome-wide significance in ancestry-

specific analyses. We validated 54.76% of putative novel associations in FinnGen biobank ($p<0.05$) as well as cross-overlapped with gene-disease predictions generated by two orthogonal machine-learning methods: Mantis-ML 2.0 and AMELIE. We report all disease prediction results, the most predictive biomarkers per disease and extracted gene/variant-disease associations in a publicly available portal: http://milton.public.cgr.astrazeneca.com.

Human data, machine learning, and mouse models demonstrate that SPEN plays a critical role in cardiac development

Authors: B-J. Kim, A. Hernandez-Garcia, P. Luna, C. Shaw, D. Scott; Baylor Coll. of Med., Houston, TX

Abstract:

Deletions of chromosome 1p36 affect approximately 1 in 5000 newborns and are associated with a high incidence of congenital heart defects (CHD). Using deletion/phenotype mapping, we have identified several 1p36 CHD critical regions suggesting that more than one gene contribute to the development of these defects. During cardiac development, mesenchymal cells are generated when the endocardial cells in the atrioventricular (AV) canal undergo endocardial-to-mesenchymal transition (EndMT). These mesenchymal cells proliferate to fill the AV cushions and then form part of the atrioventricular septum. We have previously shown that *RERE*—which encodes a nuclear receptor co-regulator—is a 1p36 gene that causes CHD in humans and mice though its effects on EndMT. Using a machine learning algorithm trained using genes known to play a role in EndMT, we identified *SPEN* as a 1p36 candidate gene for CHD.
Pathogenic *SPEN* variants have now been shown to cause a new neurodevelopmental disorder, Radio-Tartaglia syndrome which appears to cause CHD in a subset of individuals. To confirm the role of SPEN in the development of CHD, we harvested *Spen*$^{-/-}$ embryos at various timepoint during cardiac development. At E10.5, the number of mesenchymal cells in the AV cushions of *Spen*$^{-/-}$ embryos was significantly reduced compared to controls. In addition, the proliferation of mesenchymal cells was decreased in the AV cushions of these embryos at E12.5. These mesenchymal cell defects lead to the development of hypoplastic AV cushions. At E15.5, *Spen*$^{-/-}$ embryos were found to have ventricular septal defects (VSDs), double outlet right ventricle (DORV), and myocardial hypoplasia. We went on to demonstrate that tissue-specific ablation of *Spen* in the endocardium is sufficient to cause VSDs at E15.5. In RT-qPCR studies, we found that *Spen* is a downstream target of RERE. These results provide evidence that SPEN deficiency causes CHD in humans and mice and

suggest that RERE and SPEN function within the same pathway to regulate EndMT in the developing AV canal.

Using machine learning to predict noncoding variant associations with sulcal patterns in congenital heart disease

Authors: E. Mondragon-Estrada[1], J. W. Newburger[2], S. DePalma[3], M. Brueckner[4], J. Cleveland[5], W. K. Chung[2], B. D. Gelb[6], E. Goldmuntz[7], D. J. Hagler Jr[8], H. Huang[9], P. McQuillen[10], T. A. Miller[11], A. Panigrahy[12], G. A. Porter Jr[13], A. E. Roberts[2], C. K. Rollins[2], M. W. Russell[14], M. Tristani-Firouzi[15], E. Grant[2], K. Im[2], S. U. Morton[2]; [1]Boston Children's Hosp., Boston, MA, [2]Boston Children's Hosp., Harvard Med. Sch., Boston, MA, [3]Harvard Med. Sch., Boston, MA, [4]Yale Univ. Sch. of Med., New Haven, CT, [5]Univ. of Southern California, Los Angeles, CA, [6]Icahn Sch. of Med. at Mount Sinai, New York, NY, [7]Children's Hosp. of Philadelphia, Philadelphia, PA, [8]Univ. of California San Diego, San Diego, CA, [9]Children's Hosp. of Philadelphia, Philadelphia, PA, [10]Univ. of California, San Francisco, San Francisco, CA, [11]Primary Children's Hosp., Salt Lake City, UT, [12]Children's Hosp. of Pittsburgh, Pittsburgh, PA, [13]Univ. of Rochester Med. Ctr., Rochester, NY, [14]C.S. Mott Children's Hosp., Ann Arbor, MI, [15]Univ. of Utah Sch. of Med., Salt Lake City, UT

Abstract:

Neurodevelopmental impairments associated with congenital heart disease (CHD) may arise from perturbations in brain developmental pathways, including the formation of sulcal patterns. While genetic factors contribute to sulcal features, the association of noncoding *de novo* variants (ncDNVs) with sulcal patterns in people with CHD remains poorly understood. To determine if there is a relationship between the predicted functional impact of ncDNVs and sulcal pattern features among participants with CHD, we identified ncDNVs in 30 participants with available genome sequencing and brain MRI data. Leveraging deep learning models we examined the predicted functional impact of ncDNVs on gene regulatory signals of participants with CHD. Cortical surfaces were reconstructed from brain MRIs, and sulcal patterns were characterized as graphs. Correlations between ncDNV scores and sulcal folding patterns were explored using Weighted Correlation Network Analysis (WGCNA). To identify functional prediction and sulcal patterns relevant to CHD, we included an additional 62 participants with CHD who had MRI data only and 1,784 participants with CHD who had genome sequencing data only. We compared the CHD ncDNV scores and sulcal patterns with two cohorts without known congenital anomalies or neurodevelopmental diagnoses with genome sequencing (n = 1,611) and brain MRI data (n = 96), respectively. We compared the CHD ncDNV scores and sulcal

patterns with two cohorts without known congenital anomalies or neurodevelopmental diagnoses with genome sequencing (n = 1,611) and brain MRI data (n = 96), respectively. Among the highest-scored ncDNVs, neural H3K9me2 regulation, associated with transcriptional silencing, had a greater predicted impact in participants with CHD compared to those without CHD (p = 3.26e-07, θ = 0.51). The right parietal sulcal patterns in the CHD cohort were abnormal compared with the non-CHD (FDR-adjusted p = 0.0013). In the WGCNA, ncDNVs predicted to impact H3K9me2 modification were associated with larger disruptions in right parietal sulcal patterns in the CHD cohort (r = -0.70, p = 2.65e-05). The ncDNV from the CHD cohort with the highest predicted impact on H3K9me2 regulation were predicted to regulate genes enriched for functions related to neuronal development. Our results highlight the potential of deep learning models to generate hypotheses about the role of noncoding variants in brain development.

Whole exome sequencing shows cystic fibrosis risk variants confer a protective effect against inflammatory bowel disease

Authors: M. Yu; Broad Inst., Cambridge, MA

Abstract:

Genetic mutations that produce defective Cystic fibrosis transmembrane regulator (CFTR) protein cause cystic fibrosis (CF), a lethal autosomal recessive Mendelian disorder. The role of CFTR mutations in inflammatory bowel disease (IBD) has been widely debated, with inconclusive evidence supporting both risk and protective effects.
Here, we generated the largest IBD exome sequencing dataset to date, comprising 38,744 cases and 69,570 controls in the discovery stage, and 27,883 cases and 168,355 controls in the follow-up stage, from the International IBD Genetics Consortium. Our analyses support two independent lines of evidence establishing a protective role of CF-risk variants against IBD.
Firstly,we found that delF508, the predominant CF-causing variant that accounts for 70% of all CFTR mutations observed in CF patients, has a significant protective effect against IBD (p=1.7E-10, beta=-0.30, se=0.048). This association was successfully replicated in the follow-up dataset (p=3.1E-05, beta=-0.16, se=0.038).
Second, we conducted an unweighted gene burden test of CFTR, restricted to all rare (MAF<0.001) variants annotated as “CF-causing” in the Clinical and Functional Translation of CFTR (CFTR2) database, excluding delF508. This revealed a significant association between CF-causing variants and IBD (p=1.1E-6) with a slightly smaller protective effect (beta=-0.23, se=0.047). Another burden test across all rare (MAF<0.001) nonsynonymous

non-CF-causing CFTR variants showed no significant association (p=0.4).
In addition,we assessed the performance of AlphaMissense (AM), the leading variant pathogenicity predictor, at prioritising missense variants for CFTR burden tests. Our results indicate that although AM performed reasonably well at separating CF-risk and non-CF-risk variants, a higher AM score threshold does not always improve the statistical power of the burden test.
In summary, we report convincing evidence suggesting CFTR mutations confer a protective effect against IBD. It is shown that CFTR serves as epithelial receptor for S. Typhi transluminal migration and that heterozygous deltaF508 mice translocated significantly fewer S. typhi into the gastrointestinal submucosa than wild-type CFTR mice. Therefore, it is plausible that the protective effect of CFTR in IBD may stem from similar interactions with yet unidentified bacteria. Moreover,our analysis shows that to effectively incorporate pathogenicity scores in burden tests, such as AM, thresholds for selecting variants need careful consideration, and further improvements on in silico variant annotations are needed for it to match the accuracy of clinical annotations.

Session 19: Mapping the Brain in Health and Disease

Location: Room 501

Session Time: Wednesday, November 6, 2024, 10:15 am - 11:45 am

Common variants for migraine tested in 1,138,261 Europeans implicate biological processes with specific effects on head pain severity symptoms

Authors: F. Wendt, R. Wang, J. Otto, N. Banarjee, A. Zhang, GHS-RGC DiscovEHR Collaboration, MAYO-RGC Project Generation, Colorado Center for Personalized Medicine - RGC Collaboration, UCLA-RGC ATLAS Collaboration, A. Baras, for the Regeneron Genetics Center, M. Cantor, E. Stahl, G. Coppola, N. Parikshak; Regeneron Genetics Ctr., Tarrytown, NY

Abstract:

Migraine is a type of headache often accompanied by nausea, vomiting, hypersensitivity to light and sound, and lasting for 4 to 72 hours. Between 8-15% of the general population experienced a migraine in the past year with little relief offered by current therapeutics. We performed the largest common variant association study of migraine and used genetics to understand its relationship to head pain symptoms.
Migraine was defined as answering “yes” to recurrent/bad headache questionnaire and/or a G43 ICD-10 code resulting in 129,681 cases and 1,008,580 controls of European ancestry. We associated common variants (minor allele frequency > 0.05%) with migraine using logistic regression including demographic and batch effect covariates, principal components of ancestry, and a polygenic score capturing relatedness and population structure. LD Score Regression and Genomic Structural Equation Modeling (gSEM) were used to characterize common variants across 11 symptoms (e.g., light sensitivity, frequency, throbbing, etc.) using the 1000 Genome Project European reference genome. The most significant association was the known protective intronic variant *LRP1*-rs11172113 (OR=0.91, 95% CI: 0.90-0.92). After multiple testing correction (FDR < 5%), migraine genes were enriched for ‘blood vessel morphogenesis’ (3.7-fold), ‘up-regulation of regulatory T-cells’ (6.8-fold), and ‘human midbrain GABAergic neuron’ (2.4-fold) ontologies. We used gSEM to learn how 127 independent migraine variants affect symptoms through a latent factor. Migraine was genetically correlated (r_g) with 11 migraine symptoms (r_g=0.23 for ‘visual changes at headache onset’ to r_g=0.98 for ‘bad/recurrent headaches ever’). The pattern of r_g was consistent with 3 head pain factors (chi-squared=125.1, p=6.1e-10; Akaike Information Criterion=195, comparative fit index=0.99, standardized root mean square residual=0.09): ‘severity’, ‘vision’, and ‘chronicity’. Migraine loaded exclusively onto the

'severity' factor (loading = 0.93). 52% of migraine SNPs affected head pain through only the 'severity' factor. SNPs associated with greater risk for 'severity' symptoms were enriched for phosphatidylinositol-mediated signaling (2.4-fold) while SNPs associated with lower risk for 'severity' symptoms were enriched for TGF-beta signaling (8.04-fold). Consistent with the loading of migraine on only the 'severity' factor, no migraine SNPs were associated with the 'vision' or 'chronicity' factors.
Common variants associated with migraine point to several biological pathways in the brain as potential causal mechanisms for disease that may be unique to the 'severity' factor.

Genetic regulation of the gene expression in fetal and adult brains explains GWAS signals from the East Asian population

Authors: C. Han[1], Y. Chen[2], Q. Liang[1], M. Guo[1], H. Yang[1], C. Chen[1], C. Liu[3]; [1]Central South Univ., Changsha, China, [2]Broad Inst., boston, MA, [3]SUNY Upstate Med. Univ., Syracuse, NY

Abstract:

Large-scale GWAS studies have identified hundreds of risk loci for neuropsychiatric disorders like schizophrenia (SCZ), but their biological basis remains unclear. Genetic regulation of gene expression during fetal brain development may influence susceptibility to these disorders. Previous research has focused mainly on early- and mid-gestation brain development in European, populations, lacking data from late-gestation and East Asian populations. This study aims to construct eQTL maps in mid- and late-gestation East Asian brains, considering developmental stages and sexes, to elucidate GWAS signals in these populations.We collected 217 adult and 105 fetal brain samples (43 females and 62 males) from East Asian populations. eQTLs were mapped for the second (47) and third trimesters (58), as well as for adult and combined fetal stages. Specific eQTLs at each stage were further annotated with chromatin states and psychiatric disorder-related genes. The proportion of SCZ heritability mediated by cis-genic components was compared across four stages. Co-localization analyses were performed to fine-map SCZ GWAS signals using eQTLs from both adult and fetal brains. To investigate sexual dimorphism in developmental brains, we conducted sex-stratified and sex-interaction eQTL analyses and compared gene co-expression networks between sexes.Out of 143,146 fetal brain eQTLs, 45,660 (32%) overlapped with adult brain eQTLs, showing a significant effect size correlation (r = 0.72, p = 2.2e-16). Fetal-specific eQTLs (68%) were enriched for those associated with ASD, BD, and SCZ. Heritability explained by the third trimester (2.9%) closely matched the second trimester (2.6%). Fetal brain eQTLs attributed higher heritability (11.3%) compared to adult

brain (9.4%). Co-localization of fetal brain eQTLs and SCZ GWAS revealed four novel prenatal-specific risk genes: *ZNF391, NDUFA6, SMDT1, and TBL1XR1-AS1* (PPH4 > 0.95). Approximately 23% of eGenes exhibited sex-biased eQTLs in the fetal brain, with 162 eQTLs showing potential sex interactions (FDR < 0.1). At FDR < 0.05, TSPAN14 showed a significant negative correlation of effect sizes in sex-stratified eQTLs (r = -0.14, p = 2.2e-16). Most co-expression modules were highly conserved between sexes, while female-specific developmental modules (z-summary < 10) were enriched in immune-related pathways.This study constructed the first eQTL map across developmental stages in East Asian populations, identifying four novel prenatal-specific risk genes for SCZ in East Asians. It also highlighted sex dimorphism in the genetic regulation of early development, occurring before the influence of sex hormones.

Single-cell atlas of transcriptomic vulnerability across multiple neuropsychiatric and neurodegenerative diseases

Authors: D. Lee[1], M. Koutrouli[2], N. Y. Masse[1], S. Kinrot[1], M. Pjanic[1], T. Clarence[1], F. Tsetsos[1], S. P. Kleopoulos[1], Z. Shao[1], S. Argyriou[1], M. Alvia[1], C. Casey[1], A. Hong[1], X. Wang[1], P. N.M.[1], D. Mathur[1], K. Therrien[1], D. A. Bennett[3], V. Haroutunian[1], L. J. Jensen[2], S. Finkbeiner[4], D. Wang[5], K. Girdhar[1], G. Voloudakis[1], G. E. Hoffman[1], J. Bendl[1], J. F. Fullard[1], P. Roussos[1]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Univ. of Copenhagen, Copenhagen, Denmark, [3]Rush Univ. Med. Ctr., Chicago, IL, [4]Gladstone Inst.s, San Francisco, CA, [5]Univ. of Wisconsin - Madison, Madison, WI

Abstract:

Neuropsychiatric and neurodegenerative diseases impose a significant societal and public health burden. However, our understanding of the molecular mechanisms underlying these highly complex conditions remains limited. To gain deeper insights into the etiology of different brain diseases, we used specimens from the human dorsolateral prefrontal cortex of 1,494 unique donors to generate an expansive single-cell transcriptomic atlas, comprising over 6.3 million individual nuclei. The sample cohort includes neurotypical controls as well as donors affected by eight representative disorders, including Alzheimer's disease (AD), diffuse Lewy body disease (DLBD), vascular dementia (Vas), tauopathy (Tau), Parkinson's disease (PD), frontotemporal dementia, schizophrenia, and bipolar disorder. Analysis of this population-scale dataset showed that a substantial portion of gene expression variation in the human brain is due to inter-individual differences. When comparing cell type composition and transcriptomic variation across these diseases, we discovered a universal disease signature enriched in basal cellular functions such as

mRNA splicing and protein localization. After discounting shared components, trait-trait pairwise correlations revealed that AD, DLBD, Vas, and PD show strong concordance between heritability and transcriptomic similarity, with their common molecular mechanisms driven by inhibitory neurons and mural cells. We performed a detailed phenotypic analysis of AD by correlating transcriptomic changes with clinical, cognitive, and neuropathological measures of disease severity. We also explored the neuropsychiatric symptoms (NPS) that frequently accompany AD, showing that a specific category of NPS, including weight loss and psychomotor agitation, may involve the proliferation of excitatory neurons. We built non-linear AD trajectories to interpret transcriptional changes related to AD progression and show that initial activation of immune function and compensatory upregulation of synaptic function occur in the early stages of the disease, followed by a decline of both neuronal and cortical vascular function in later stages. Our atlas provides an unprecedented perspective on the transcriptomic landscape of neuropsychiatric and neurodegenerative diseases, which sheds light on shared and distinct disease-related changes and how the interplay between the neuro-immune-vascular systems contributes to disease progression.

A nucleotide-scale map of brain cell effects of neurodegenerative GWAS variants reveals distinct and shared causal disease mechanisms

Authors: K. Fletez-Brant, T. Bhangale; Genentech, South San Francisco, CA

Abstract:

Genome-wide association studies (GWAS) have identified hundreds of loci associated with neurodegenerative disorders, yet assigning effector genes and cell-types of action to these loci have remained challenging. Here we first apply a deep learning approach ChromatinHD (CHD doi:10.1101/2023.07.21.549899) to derive cell-type-specific SNP to gene maps based on multiome data in CNS cell types, and use these to identify candidate genes and corresponding effector cell-types using fine-mapping results for five neurodegenerative traits. We focus on high-confidence GWAS (fine-mapped posterior inclusion probability (PIP) > 0.25, and in credible sets (CS) with less than 5 variants) variants in: Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS), multiple sclerosis severity (ARMSS), multiple sclerosis (MS), and Parkinson's disease (PD). We integrate these variants with the cell type-specific information of the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD: 500,000 cells in 28 donors, including 17 healthy individuals, in 18 cell types) 10X Multiome dataset using CHD. CHD learns the predictive contribution ('predictivity' of gene expression) for chromatin accessibility of each base-pair

in a window up- and downstream of a gene's transcription start site, without peak calling, enabling the linking of SNPs to genes. Using SEA-AD healthy individuals' data, we train CHD models for oligodendrocytes (OG), oligodendrocyte precursor cells (OPC), microglia (MG), astrocytes (AS) and three sets of GABAergic neurons (expressing SNGC, LAMP5 or SST) separately, training on variants 100kb up- or downstream of the TSS . Genome-wide, we observe nominally significant correlations between per-CS average predictivity and PIP for ARMSS and AS (ρ; = 0.99, p = 0.01), MS and OPCs (ρ; = 0.21, p = 0.012), and PD with MG (ρ; = 0.32, p = 0.0015), OG (ρ; = 0.27, p = 0.006), AS (ρ; = 0.23, p = 0.015) and LAMP5 (ρ; = 0.27, p = 0.008) and SST (ρ; = 0.26, p = 0.042). At the level of individual GWAS CS, CHD-based SNP evaluation proposes mechanistic roles for disease-associated variants with cell type specificity, and recovers known phenotype-gene associations. Examples include PD and LRRK2 (OPC, MG, OG, AS, LAMP5), AD and ABCA7 (LAMP5, AS), PICALM (SNGC, OG, OPC, AS), MS and TAGAP (OPC, AS), TNFRSF1A (OG). Overall, we identify 164 unique GWAS CS-gene effects for AD, 33 for ALS, 4 for ARMSS, 217 for MS and 112 for PD, highlighting the discovery power of our integrative approach.

Cell-Cell Communication Patterns in Alzheimer's Disease Dementia and Mild Cognitive Impairment Vary by Cortical Layers

Authors: Q. Dai[1], S. Tang[1], J. Hu[1], P. De Jager[2], D. Bennett[3], A. Buchman[3], J. Yang[1], M. Epstein[1]; [1]Emory Univ., Atlanta, GA, [2]Columbia Univ Med Ctr, New York, NY, [3]Rush Univ., Chicago, IL

Abstract:

Background: Current cell-cell communication (CCC) studies of Alzheimer's disease (AD) with snRNA-seq data from the prefrontal cortex fail to consider that the prefrontal cortex consists of six layers and white matter, each with distinct patterns of cell types and gene expression. To bridge this gap, we used spatial transcriptomic reference data to annotate cortical layers in a large-scale snRNA-seq dataset of AD and subsequently conducted layer-informative CCC studies of AD using our recently developed Supervised Tensor Analysis tool (STACCato).
Method: Using the CeLEry tool with LIBD spatial transcriptomic reference data, we annotated cortical layers for 1.5M cells in the snRNA-seq data profiled from the prefrontal cortex of 415 participants in ROSMAP studies, including 151 No Cognitive Impairment, 109 Mild Cognitive Impairment (MCI), and 155 AD dementia (ADD). We employed STACCato to estimate the effects of AD and MCI on CCC events, while adjusting for confounding covariates, including age, gender, education, APOE and TOMM40 variants, and recent

medication usage. We identified significant CCC events as those with the top 5% largest AD/MCI effect magnitudes and p-values < 0.001. We compared the layer-informative CCC results to layer-independent results that did not consider layer information.
Results: We identified ~2K out of ~40K CCC events to be significantly associated with AD/MCI in each layer-specific analysis. The identified significant CCC patterns were distinct across the layers. Results of layer 3 (the main information integration layer) were most similar to the layer-independent results, likely due to its highest cell counts. In contrast, layer 5 (L5, the main output layer) demonstrated a pattern unique from other layers where Microglia (MG) cells actively interacted with other cell types more in AD patients. Specifically, in L5 of AD patients, ligand C3 in MG actively interacted with receptors LRP1, GRM7, CD81, and CD46 in Astrocytes (AST), Excitatory Neurons (EX), Inhibitory Neurons (IN), Oligodendrocytes (OL), and OPCs. Meanwhile, receptors NOTCH2 and CD74 in MG actively interacted with ligand APP, a precursor molecule whose proteolysis generates amyloid beta and is known as a key player in AD, in AST, EX, IN, OL, and OPCs. Effect size directions are consistent for CCC events significant for both AD and MCI.
Conclusion: Our layer-informative CCC study revealed that the associations between CCC patterns and AD/MCI differ across cortical layers, with L5 exhibiting distinctive interaction patterns in the Microglia cells of AD patients, indicating that the underlying biological mechanisms of AD could differ across cortical layers.

Mechanisms of presenilin2 driven neuroinflammation: Impact of *PSEN2*-N141I variant on microglial response to Alzheimer's disease-relevant stimuli

Authors: A. Reid[1], S. Mamde[2], L. N. Cochoit[1], S. Jayadev[3], J. Young[3]; [1]Univ. of Washington, Seattle, WA, [2]Univ. of California San Diego, San Diego, CA, [3]Univ of Washington, Seattle, WA

Abstract:

Autosomal dominant variants in PSEN1, PSEN2, and APP cause nearly 100% penetrant early-onset familial Alzheimer's Disease (EOFAD), sharing many clinical and pathological features with late-onset AD (LOAD). The efficacy of newly trialed drugs for EOFAD and LOAD has been modest at best, driving an impetus to explore novel therapeutic avenues. Microglia have emerged as pivotal contributors to LOAD pathogenesis. However, our understanding of microglial states and their regulation in EOFAD is inadequate, due in part to a scarcity of dedicated studies in this population. A challenge to translating data from human brain omics studies into mechanistic understanding is the static nature of autopsy

tissues. We aimed to leverage a microglia-neuron coculture model derived from human-iPSC with a causative EOFAD mutation, PSEN2-N141I, to elucidate the phenotypic and functional repercussions of the variant on microglia and how AD-relevant stimuli alter EOFAD microglia physiology and interactions with neurons.Microglia were differentiated from PSEN2-N141I and isogenic control hiPSC lines and cocultured with wildtype hiPSC-derived mixed cortical neurons and astrocytes. Cocultures were exposed to healthy-neuron conditioned media, UV-injured neuron conditioned media, or fibrillized Aß, both in acute and chronic settings. Transcriptomic analysis yielded 7 distinct microglial phenotypes. PSEN2-N141I exhibited specific enrichment in phagocytic and lysosomal processing pathways. Inferred gene regulatory network analysis revealed PSEN2-N141I specific regulons at baseline and after stimulation. Cell-chat analysis nominated microglia-astrocyte and microglia-neuron communication pathways that were either genotype or treatment specific. Functional assays showed increased lysosomal processing and enzymatic activity in mutant microglia consistent with function implicated by gene expression data. These results show how AD PSEN2 N141variant may exaggerate the typical microglial response to AD-associated injury, potentially further exacerbating the immune response to AD pathology. These studies underscore the utility of hiPSC-derived microglia-neuron cocultures in identifying changes in microglial physiology implicated in supporting neuronal health and potentially contributing to disease pathogenesis.

Session 20: Moving Polygenic Risk Scores Closer to Clinical Implementation

Location: Room 401

Session Time: Wednesday, November 6, 2024, 10:15 am - 11:45 am

Implementation of breast cancer polygenic risk scores in a personalized screening trial

Authors: K. Fergus[1], R. Heise[2], L. Sabacan[1], S. Kapoor[1], A. Fiscalini[1], A. Blanco[1], K. Ross[1], D. Goodman-Gruen[3], M. Scheuner[1], J. Tice[1], L. Madlensky[4], E. Ziv[1], L. Van 'T Veer[1], L. Esserman[1], Y. Shieh[1], Athena/WISDOM Network Collaborators and Advocate Partners; [1]Univ. of California San Francisco, San Francisco, CA, [2]Weill Cornell Med., New York, NY, [3]Univ. of California Irvine, Irvine, CA, [4]Univ. of California San Diego, San Diego, CA

Abstract:

Background: The Women Informed to Screen Depending On Measures of risk (WISDOM) Study is a first-in-kind randomized trial of the safety and efficacy of personalized breast cancer screening. Participants in the intervention arm undergo risk stratification via a clinical risk model modified by a polygenic risk score (PRS). Results are used to inform the starting age, interval, and modality of screening. As such, WISDOM is one of the first large-scale, prospective uses of the PRS to inform screening. We examined associations between the PRS and known risk factors, as well as its effect on changing screening recommendations. **Methods**: We analyzed participants undergoing personalized screening in WISDOM with available PRS. Genotyping was performed using a targeted capture panel from Color Genomics. We implemented separate PRS for each of four self-reported racial or ethnic groups: Asian, Black, Hispanic, and White. Each PRS contained genome-wide significant SNPs from European ancestry GWAS, as well as SNPs discovered in race- or ethnicity-specific GWAS. The resulting PRSs contained 118-126 SNPs. The PRS was used to modify the 5-year risk from the Breast Cancer Surveillance Consortium (BCSC) model, a validated clinical risk model. We analyzed between-group differences in PRS using Kruskal-Wallis tests and correlations using Pearson's r. **Results**: Our analysis included 21,670 participants with available PRS, with 5% self-reporting as Asian, 5% Black, 9% Hispanic, and 78% White. The PRS was right-skewed, with a mean of 1.04 (SD=0.54) and median of 0.92 (IQR 0.67-1.28). Higher PRS was associated with combined first- and second-degree family history of breast cancer (1.13; SD=0.59), followed by first-degree only (1.08; SD=0.55), second-degree only (1.05; SD=0.56), compared to no family history (1.00; SD=0.51), ($p < 0.001$). PRS was also positively associated with breast density ($p<0.001$).

Using the PRS to modify the BCSC risk score reclassified 836 of 6615 (13%) women aged 40-49 to receive a screening recommendation (versus no screening) in year one of enrollment. Among women who initially received a screening recommendation, 443 of 1995 (22%) were recommended decreased screening frequency or no screening at all until age 50 based on PRS. **Discussion**: The breast cancer PRS provided additional risk stratification when combined with a validated clinical risk model. We are awaiting the availability of trial outcomes in 2025 to directly evaluate the performance of the PRS, but the association between PRS and family history and mammographic density suggests that it should predict breast cancer.

All of Us diversity and scale improve polygenic prediction contextually with greatest improvements for under-represented populations

Authors: Y. Wang[1], K. Tsuo[2], Z. Shi[3], T. Ge[1], R. Mandla[4], K. Hou[4], Y. Ding[5], B. Pasaniuc[3], A. Martin[1]; [1]Massachusetts Gen. Hosp., Boston, MA, [2]Broad Inst. of MIT and Harvard, Cambridge, MA, [3]UCLA, Los Angeles, CA, [4]Univ. of California, Los Angeles, Los Angeles, CA, [5]Dana-Farber Cancer Inst., Boston, MA

Abstract:

Recent studies have demonstrated that polygenic risk scores (PRS) trained on multi-ancestry data can improve prediction accuracy in groups historically underrepresented in genomic studies, but the availability of linked health and genetic data from large-scale diverse cohorts remains limited. To address this need, the All of Us research program (AoU) generated whole-genome sequences of 245,388 individuals who collectively reflect the diversity in the US. Leveraging this resource and the UK Biobank (UKB) of a half million participants, we developed PRS trained on multi-ancestry and multi-biobank data with up to ~750,000 participants for 32 common, complex traits and diseases across a range of genetic architectures. We held out a subset of ancestrally diverse participants for testing in AoU. We compared effects of ancestry, PRS methodology, and genetic architecture on PRS accuracy across ancestrally diverse AoU participants. Due to the more heterogeneous study design of AoU, we found lower heritability on average compared to UKB (0.075 vs 0.165), which limited the maximal achievable PRS accuracy in AoU. Overall, we find that the increased diversity of AoU significantly improved PRS performance in some participants in AoU, especially underrepresented individuals, across multiple phenotypes. Notably, we observe that for the African ancestry test group, maximizing sample size by combining discovery data across AoU and UKB is not the optimal approach for predicting less polygenic phenotypes; rather, using data from only AoU for these traits resulted in the

greatest accuracy. This was especially true for less polygenic traits with large ancestry-enriched effects such as neutrophil count (e.g., DARC, R2 :0.041 vs 0.009) and white blood cell count (e.g., HBB, R2 :0.058 vs 0.005). Lastly, we calculated individual-level PRS accuracies rather than grouping by continental ancestry, a critical step towards interpretability in precision medicine. Individualized PRS accuracy linearly decays as a function of ancestry divergence, and notably the slope was smaller when using multi-ancestry GWAS compared to using European GWAS only. Our results highlight the potential of biobanks with more balanced representations of human diversity to facilitate more accurate PRS for individuals least represented in genomic studies.

Polygenic risk score enriches for clinically significant prostate cancer in a screening program - the BARCODE 1 study results

Authors: R. Eeles[1], E. Bancroft[2], J. McHugh[1], E. Saunders[1], M. Brook[1], E. McGrowder[1], S. Wakerell[1], D. James[1], E. Page[1], A. Osborne[1], N. Kinsella[2], S. Sohaib[2], D. Cahill[2], S. Hazell[2], S. Withey[2], I. Rafi[3], P. Kumar[2], N. James[1], S. Benafif[4], N. Pashayan[5], Z. Kote-Jarai[1]; [1]The Inst. of Cancer Res., Sutton, United Kingdom, [2]The Royal Marsden NHS Fndn. Trust, Sutton, United Kingdom, [3]St Georges Univ., London, United Kingdom, [4]Univ. Coll. London, Sutton, United Kingdom, [5]Univ. of Cambridge, London, United Kingdom

Abstract:

Background: Incidence of prostate cancer (PCa) is increasing, but there is no internationally agreed population screening program. Studies using an age-based PSA approach show a high rate of false-positive results as well as over-diagnosis of indolent PCa. Genome wide association studies identify common germline variants to calculate a polygenic risk score (PRS) associated with PCa risk. The BARCODE1 study used PRS to target PCa screening to those at higher risk based on genotype. **Methods:** European men aged 55-69yrs were recruited via Primary Care in the UK. PRS was constructed by summing weighted risk alleles for 130 PCa risk variants using germline DNA from saliva samples via mailed kits. Men with a PRS > 90th centile were invited for PCa screening using MRI and 12-core transperineal biopsy (including MRI fusion to target additional lesions where identified) irrespective of PSA result. **Results:** Invitation letters were sent to 40,292 men. 8,953 (22%) expressed an interest; 8,014 were eligible and sent a saliva kit. 6,644 consented; 6,393 were genotyped; 251 failed QC. A total of 6,142 participants had PRS calculated: 745 (12.1%) had a PRS > 90th centile and were invited to screening. 558/745 participants attended screening (121 declined, 66 excluded on health grounds). 551 underwent MRI and 468 had prostate biopsy resulting in 187 (40.0%) diagnoses of PCa,

overall PCa detection rate 2.8%. Mean age at diagnosis 64.1yrs (range 57-73; median 64). Using NCCN criteria (2023) 103/187 (55.1%) of cancers were Intermediate or High Risk; 40/187 (21.4%) were Intermediate Unfavourable/High/Very High Risk. 119/187 (63.6%) men had a PSA ≤3.0ug/L; PPV of biopsy for PSA > 3.0ug/L was 49.6%. PPV of MRI (presence of PI-RADS 3-5 lesion) 60.4%. PPV of PRS alone 40%. 103/187 (55.1%) had Gleason ≥7; compared with 360/1014 (35.5%) p < 0.001 in the PSA directed ERSPC study. **Conclusions:** A population PCa screening program using PRS risk-stratification enriches for clinically significant PCa requiring treatment. It detects a high proportion of clinically significant disease compared with PSA or MRI based screening programs and MRI missed a significant proportion (17-67%) of cancers found on biopsy. This is the first study to assess if this approach will be useful in population screening programs.

Genetic and metabolomic determinants of disease in the UK Biobank

Authors: J. Barrett[1], K. Schut[2], S. Kerminen[3], K. Alasoo[4], R. Tambets[4], I. Rahu[4], P. Palta[5], E. Abner[4], J. Kronberg[4], U. Võsa[4], P. Wurtz[1], L. Jostins-Dean[6]; [1]Nightingale Hlth., Helsinki, Finland, [2]Nightingale Hlth.Plc., Helsinki, Finland, [3]Nightingale Hlth.Plc, Helsinki, Finland, [4]Univ. of Tartu, Tartu, Estonia, [5]Universit of Tartu, Tartu, Estonia, [6]Univ. of Oxford, Oxford, United Kingdom

Abstract:

We recently completed metabolomic profiling of all half a million UK Biobank participants. Having both genetic and metabolomic data available for all participants allows focused integrative omics analyses that were previously restricted by sample size. Here, we highlight key results from initial analyses.
We built and validated combined polygenic and metabolomic risk scores for 30 chronic diseases and cancers. The metabolomic scores are more strongly associated than polygenic scores for all diseases tested except common cancers, and the metabolomics tracked observed changes in risk profile across time in longitudinal samples. Our multi-omic scores increased predictive accuracy when added to standard clinical screening scores for most common diseases (delta AUC ranging from 0.01 to 0.12), and for four diseases (myocardial infarction, colon cancer, alcoholic liver disease and liver cirrhosis) multi-omic scores alone outperformed existing clinical scores (delta AUCs of 0.01, 0.02, 0.12, 0.05). We validated our scores for 8 disease endpoints in individuals of non-European ancestries, showing attenuated but significant predictive accuracy. Differences in the distributions of metabolomic risk profiles by ancestry mirrored differences in population risk for diabetes and heart attack though these explain only a minority of differences in observed disease incidence rates between ancestry groups in the UK.

We were able to study rarer diseases and shorter time-frames. This included rare cancers, revealing strong potential to predict 10-year risk of liver cancer and multiple myeloma using multi-omic scores (hazard ratios of 1.84 and 1.62 per standard deviation of the score). We found considerably higher risks for individuals within a year of the blood sampling for many diseases, including vascular dementia and many cancers, which may in part reflect prevalent cases who had not yet been diagnosed. The full dataset also enables GWAS of serum metabolites at unprecedented scale. Within Europeans, we found 771 independent autosomal regions with genome-wide significant metabolite associations, with between 29 and 340 associations per metabolite. We were also able to test for new metabolite associations for rare mutations, including a novel association between Noonan Syndrome 1-causing mutations in *PTPN11* and serum fatty acid levels. Our complete metabolomic dataset, will be available for approved users of UK Biobank projects from Autumn 2024. Summary statistics for genome-wide association studies and metabolite-phenotype associations will be made publicly available at the same time.

Performance of contemporary polygenic risk scores for atherosclerotic cardiovascular disease in the All of Us Workbench ⭐

Authors: J. L. Smith[1], K. Norland[2], M. Hamed[1], Y. Yu[1], J. Na[1], O. Dikilitas[1], D. J. Schaid[1], I. Kullo[1]; [1]Mayo Clinic, Rochester, MN, [2]deCode genetics/Amgen Inc., Reykjavik, Iceland

Abstract:

Background: Polygenic risk scores (PRS) for atherosclerotic cardiovascular disease (ASCVD) phenotypes are trained on primarily European (EUR) cohorts, limiting portability. Recently developed PRS need to be benchmarked in an external dataset to identify the optimum and most portable PRS for ASCVD phenotypes. The All of Us Workbench provides access to a large diverse cohort with electronic health record data for whole genome sequenced individuals, ideal for PRS validation. **Objectives:** Evaluate the predictive performance of 12 contemporary PRS for the following ASCVD phenotypes: coronary heart disease (CHD), cerebrovascular disease (CVD), abdominal aortic aneurysm (AAA), and peripheral artery disease (PAD), and PRS for risk factors including hypertension (HTN), low-density lipoprotein (LDL), obesity, and type 2 diabetes (T2D). **Methods:** Validations of PRS for ASCVD phenotypes in 245,388 All of Us participants were performed to compare risk estimates in EUR, Admixed American (AMR), and African (AFR) populations. PRS for CHD, CVD, AAA, PAD, and risk factors (HTN, LDL, obesity, and T2D) were assessed for portability across genetic ancestry groups using hazards ratios (HR) per SD, C-statistics, and calibrations across ancestry specific populations. We also integrated the PRS with two

conventional risk equations, the pooled cohort equation (PCE) and the newly developed PREVENT equation, across all populations. **Results:** For CHD, the PRS by Patel et al. had the strongest association (HR[95% CI]), in all ancestry groups (EUR: 1.72[1.67-1.78], AMR: 1.48[1.37-1.59], AFR: 1.24[1.18-1.31]). For PAD, the PRS by Norland et al. performed best in EUR (1.16[1.12-1.20]) and AFR (1.10[1.05-1.15]) and the best performing PRS in AMR (1.09[1.02-1.16]) was developed by Prive et al. For CVD, the best performing PRS was from Abraham et al. in AFR (1.12[1.06-1.17]) and AMR (1.11[1.04-1.19]), and from the Norland et al. PRS in EUR (1.16[1.12-1.20]). Best performing PRS for AAA across ancestry groups was from Roychowdhury et al. (EUR: 1.68[1.59-1.78], AFR: 1.29[1.13-1.48], AMR: 1.30[1.06-1.60]). PREVENT had higher accuracy than PCE equations for conventional 10-year ASCVD risk estimation and integrating PRS led to an increase in C-statistic on average across all traits and ancestry groups. **Conclusions:** Integrating PRS enhances ASCVD risk estimations. Inclusion of multi-ancestry cohorts in PRS development improves risk prediction for ASCVD phenotypes across ancestrally diverse and admixed individuals. Results highlight the need for larger training sets for AFR and all genetic ancestry groups for CVD, AAA, and PAD.

PGS Browser: a comprehensive analysis of 3,168 polygenic score models across 400,000 Finns ⭐

Authors: N. Kolosov[1,2,3], M. Reeve[4,5], T. Sipilä[4], V. Llorens[4], M. Aavikko[4], S. Ripatti[4], A. Palotie[4,5,6], M. Daly[4,5,6], FinnGen, M. Artomov[2,1,3]; [1]The Ohio State Univ. Coll. of Med., Columbus, OH, [2]Nationwide Children's Hosp., Columbus, OH, [3]Inst. of Molecular Med. Finland (FIMM), Helsinki, Finland, [4]Inst. for Molecular Med. Finland (FIMM), HiLIFE, Univ. of Helsinki, Helsinki, Finland, [5]Broad Inst. of MIT and Harvard, Cambridge, MA, [6]Massachusetts Gen. Hosp., Boston, MA

Abstract:

Polygenic scores (PGS) serve as an individual metric for disease susceptibility and are broadly utilized for risk prediction, risk stratification, or uncovering shared genetic etiology across phenotypes. The Polygenic Scores Catalog (PGS Catalog) is a primary resource for accessing published PGS models for individual phenotypes. However, the overall utility of this catalog has not been explored in a context of large-scale multi-phenotypic databases, such as biobanks. Here we integrated PGS Catalog with a dataset of 392,649 participants from the FinnGen project. Through this, we were able to perform several unique types of analyses which were not feasible in a single-phenotype context. In particular, we comprehensively evaluated 3,168 PGS models from the PGS catalog on the entire multi-

phenotypic cohort of FinnGen participants. For each PGS we conducted a Phenome-Wide Association Study, testing associations with 4,995 endpoints from FinnGen, and identified **652,865 significant associations**. Further, we highlight 81 phenotypes for which risk prediction can be significantly improved by combining multiple PGS. Improvement reached 0.5%-8% for the area under the receiver operating characteristic curve. Ultimately, we developed the "PGS Browser," a web-based application that provides access to the aforementioned experimental results, along with the ability to interactively filter, visualize and download the data. Furthermore, the **browser enables the use of models, trained on the entire FinnGen, for external individuals to infer age of the disease onset and lifetime risks for 77 phenotypes**. PGS browser addresses the need for more accessible and clinically integrated tools, making cutting-edge research results from FinnGen available to the broader research community. We envision that this type of analysis and novel interface will bridge the gap between the development of PGS models and their practical application in clinical settings.

Session 21: Multimodal Approaches to Interpreting the Non-Coding Genome: Evolution, Functional Genomics, and Machine Learning

Location: Mile High Ballroom 2&3

Session Time: Wednesday, November 6, 2024, 10:15 am - 11:45 am

Evolutionary conservation and functional analysis of neuronal regulatory elements in mammals

Authors: B. Rogers[1], A. Anderson[1,2], E. A. Barinaga[1], J. Loupe[1], E. WaMaina[1], L. Rizzardi[2], G. Cooper[1], R. Myers[1], J. Cochran[1]; [1]HudsonAlpha Inst. for Biotechnology, Huntsville, AL, [2]Univ. of Alabama at Birmingham, Birmingham, AL

Abstract:

Understanding the gene regulatory mechanisms underlying brain function is crucial for advancing knowledge of the genetic basis of neurologic diseases. Cis-regulatory elements (CREs) play a pivotal role in gene regulation, and their evolutionary conservation across mammals can offer valuable insights into their functional importance. This study leverages numerous genomic techniques, including single nucleus multiomics (snRNA+snATAC-seq), bulk and NeuN+ RNA-seq, NeuN+ ATAC-seq, NeuN+ ChIP-seq of key histones, and Hi-C, to analyze post-mortem cortex tissue from cow (*Bos taurus*), horse (*Equus caballus*), cat (*Felis catus*), mouse (*Mus musculus*), rat (*Rattus rattus*), rabbit (*Oryctolagus cuniculus*), and human (*Homo sapien*). By generating comprehensive datasets across these species, we aim to elucidate the conservation and divergence of brain-specific candidate CREs (cCREs) and 3D genome interactions. To determine the transcriptional activity of nominated cCREs, we are performing a massively parallel reporter assay (MPRA) of cCREs that are either conserved across all species or unique to humans in neural precursor cells (NPCs) and excitatory neurons (KOLF2.1J-hNGN2), providing a comprehensive view of their activity and relevance. Additionally, using CRISPR interference (CRISPRi) in induced pluripotent stem cell (iPSC)-derived neurons (KOLF2.1J), we functionally validated the target genes of 22 cCREs that are highly conserved across mammals, as well as zebrafish. Overall, this research has the potential to identify conserved regulatory elements critical for fundamental neural processes and uncover human-specific elements that could contribute to unique cognitive functions and disease susceptibilities. By integrating data from diverse mammalian species and utilizing genomic and functional validation techniques, this study provides a deeper understanding of the evolutionary dynamics of

brain regulation, which will inform novel contributors to and therapeutic targets for neurologic diseases. Ultimately, this work aims to bridge the gap between evolutionary biology and translational research, fostering a better understanding of how conserved regulatory mechanisms influence brain physiological function and disease.

Uncovering gene regulatory differences between human and chimpanzee neural progenitors

Authors: J. Song[1], A. C. Carter[2], E. M. Bushinsky[1], S. G. Beck[1], J. Petrocelli[2], M. E. Greenberg[2], C. Walsh[1]; [1]Boston Children's Hosp., Boston, MA, [2]Harvard Med. Sch., Boston, MA

Abstract:

Although comparisons of human and non-human primate brains have identified thousands of molecular differences, it has been difficult to identify the human-specific sequence variants that underlie the dramatic modifications to brain size, connectivity, and function found in humans. One hurdle is that current comparative approaches cannot distinguish *cis*-regulated genes, which change in expression due to nearby sequence variants on the same DNA molecule, from *trans*-regulated genes, which change in expression due to changes in diffusible factors in the cellular environment (like the levels of *cis*-regulated transcription factors (TFs)). To distinguish *cis* from *trans* changes, we generated human-chimpanzee tetraploid stem cell lines as a genetic model where the human and chimpanzee genomes are in the same cellular environment and only *cis*-regulated changes are observed. We have now used this system to profile *cis*- and *trans*-regulated genes and open chromatin regions in neural progenitor cells (NPCs), in order to identify genetic changes that underlie the expansion in size and neuron number in the human brain. Genes that are more highly expressed in humans are enriched for processes related to mitosis, consistent with increased neurogenesis in humans. We identify *cis*-regulated TFs, including *FOSL2* and *MAZ*, whose motifs are enriched at *trans*-regulated open chromatin peaks, suggesting that these TFs may be major drivers of epigenomic and transcriptomic rewiring between human and chimpanzee NPCs. To identify human-specific variants that underlie *cis*-regulated gene expression changes, we linked *cis*-regulated open chromatin peaks that contain derived sequence changes in humans to nearby *cis*-regulated genes. A CRISPR inhibition screen of 106 *cis*-regulated peaks identified species-specific enhancers, including one near *TNIK*. Further characterization of *cis*-regulated TFs and non-coding regions in NPCs, along with the application of this model to additional cell types and paradigms, will advance our understanding of how human-specific sequence

changes contribute to increased brain size, as well as other phenotypes that have arisen in the human lineage.

Cross-species variant-to-function analyses implicate insomnia effector genes and reveal a highly conserved regulatory architecture at the *MEIS1* locus

Authors: A. Zimmerman[1,2], M. Pahl[2], F. Doldur-Balli[1], B. Keenan[1], E. Almeraya Del Valle[1], J. Palermo[3], A. Chesi[1,2], S. Sonti[2], E. Brown[3], J. Pippin[2], A. Wells[1,2], O. Veatch[4], D. Mazzotti[4], P. Gehrman[1], A. Keene[3], A. Pack[1], S. Grant[1,2]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Children's Hosp. of Philadelphia, Philadelphia, PA, [3]Texas A&M Univ., College Station, TX, [4]Univ. of Kansas Med. Ctr., Kansas City, KS

Abstract:

Genome-wide association studies (GWAS) have identified hundreds of insomnia loci, but the functional mechanisms by which they confer their effects remain largely uncharacterized. GWAS signals often implicate non-coding variants, typically annotated to the nearest protein-coding gene; however, via chromatin folding, non-coding elements can act in cis on multiple genes, influencing their expression in a cell- and developmental-specific manner. We deployed our variant-to-gene mapping approach in iPSC-derived neural progenitor cells to implicate candidate causal insomnia variants that reside in non-coding regions, which form chromatin contacts with accessible promoters of putative effector genes. Next, we screened the nominated effector genes through a high-throughput, neuron-specific *Drosophila* RNAi approach followed by CRISPR-Cas9 mutagenesis in F0 zebrafish larvae. We compared behavioral measures across flies and fish for insomnia-associated effector genes and revealed distinct behavioral fingerprints characterized by hyperactivity at night and altered daytime sleep duration, which are characteristics of human insomnia. Phenotypic clustering by sleep traits revealed loss of *MEIS1* orthologs in fish and flies produced a similar behavioral profile marked by shortened sleep duration and fragmented sleep at night. Given *MEIS1* is also significantly associated with restless legs syndrome (RLS), we analyzed the linkage disequilibrium relationship between insomnia and RLS variants, revealing an independent GWAS signal associated only with insomnia risk sentinel SNP, rs1519102. Using our variant-to-gene approach across 11 human iPSC and ESC-derived brain cell types, we implicated proxy variant ($r2$>0.8) rs13033745, which resides in the promoter of an antisense gene, *MEIS1-AS3*, and forms multiple chromatin contacts with the *MEIS1* promoter and other distal regulatory elements. rs13033745 colocalizes with an eQTL where the insomnia risk allele increases expression of *MEIS1-AS3* in the cerebellum. We compared this locus across

species and found an antisense-oriented transcript upstream of the *MEIS1* orthologs in fish and flies. These antisense transcripts show spatio-temporal restriction of expression to neurons involved in cerebellar and hindbrain development that oppose temporal patterns of *MEIS1* expression, suggesting the antisense transcript acts to control spatio-temporal regulation of *MEIS1* in these cells. Together, this variant-to-function analysis reveals a discrete and highly conserved regulatory element at the *MEIS1* locus that implicates a specific role for the cerebellum/hindbrain in the development of insomnia.

Constructing cell type-specific enhancer-promoter regulatory interaction networks with massively parallel reporter assays

Authors: W. DeGroat, A. Kreimer; Rutgers, The State Univ. of New Jersey, Piscataway, NJ

Abstract:

Enhancers are cis-regulatory elements, non-coding sequences of DNA pivotal to cell type-specific gene regulation. While a consensus that enhancers are hubs for disease-associated variants has been reached, little is known about the mechanisms through which these elements mediate their effects on gene expression. Additionally, the map of these enhancers' location and their target genes remains partial. Computational models, which extrapolate epigenetic markers to regulatory activity, have proven immensely successful in predicting enhancer-promoter interactions (E-P-Is). Still, there are steps to be taken to improve these model's ability to capture cell type specificity and generate more accurate E-P-I predictions.

Massively parallel reporter assays (MPRAs) are a cutting-edge technique for assessing the functionality of regulatory elements and their perturbations under varied conditions (e.g., cell types). Yet, computational methods for analyzing MPRA data in a cell-type-specific manner are still lacking. MPRA offers a single, comparable means of computing the regulatory activity of enhancers across E-P-I networks. Utilizing a convolutional neural network trained on MPRAs from K562, HEPG2, and iPSCs in tandem with epigenetic datasets, we defined and scored enhancer regions and linked them to target genes. The K562 E-P-I network was benchmarked against existing models utilizing CRISPRi as a "gold standard." We performed a series of analyses on these three cell type-specific networks, mapping eQTLs and GWAS variants, dissecting regulatory substructures, and integrating TF interactions into our network. Overall, we saw improvements in our framework's ability to predict cell type-specific E-P-Is compared to existing approaches.

The usage of MPRA in E-P-I prediction models could allow for the creation of high-accuracy models reliant on fewer multi-omic datasets. The circulation of our model could incentivize

the more widespread generation and usage of MPRA, a breakthrough technology, across the scientific community. Importantly, this approach is generalizable and can be utilized for different cellular contexts.

Unbiasedly partitioning the heritability of scRNA-seq data reveals that the vast majority of cell type-specific gene regulation lies in trans

Authors: M. Chen[1], L. Krockenberger[2], B. Balliu[2], X. Liu[1], A. Dahl[1]; [1]Univ. of Chicago, Chicago, IL, [2]UCLA, Los Angeles, CA

Abstract:

It is likely that most genetic effects on complex traits are mediated through the regulation of gene expression. While bulk RNA studies like GTEx have successfully identified many expression quantitative trait loci (eQTLs) driving GWAS signals, most GWAS hits remain unexplained. One key limitation of these efforts is that they primarily identify large-effect *cis*-eQTLs that are shared across cell types, a class of eQTL that is depleted of biological relevance. Recently, large-scale single-cell RNA-seq (scRNA-seq) data have enabled the identification of eQTLs at cell type resolution, which has succeeded in explaining additional GWAS signals. Nonetheless, single-cell eQTLs are still biased toward large-effect *cis* eQTLs that are shared across cell types. In particular, cell type-specific and *trans* regulation remains poorly understood. To address this gap, we developed a novel Gene-by-Cell Type linear Mixed Model (GxCTMM) for scRNA-seq data to partition the heritability of gene expression into cell type-specific and -shared components. Extensive simulations show that GxCTMM is powerful and unbiased in realistic settings. We applied our method to scRNA-seq data from blood cells in the OneK1K cohort (N=982). First, we show that the median *cis* heritability is ~3% across the transcriptome, consistent with estimates from bulk RNA-seq. We further show that *cis* effects are highly shared across cell types (60% specific). Next, we find that the median *trans* heritability is ~25%, though estimates are noisier. Excitingly, we find that *trans* effects are mostly cell type-specific (86%) in contrast to *cis* effects (p_diff=0.0001); importantly, GxCTMM is not biased by power differences in *cis* vs *trans*. Functional genomic analyses show that genes with more specific *cis* regulation are more evolutionarily constrained, have complex enhancer structures, and are more connected in gene expression networks. Additionally, we find that conserved enhancers are enriched for cell type-specific effects. Finally, LDSC shows that genes with more specific *cis* regulation enrich heritability for 4/7 immune diseases, including 3 missed by ordinary differential expression across cell types. We then studied scRNA-seq data from CLUES (N=264). We find similar *cis* heritability estimates and

replicate the correlations between cell type-specificity and all three functional genomic annotations. We also find that GxCTMM results correlate strongly across populations within CLUES. Overall, we have rigorously partitioned gene regulation across loci and cell types in scRNA-seq data, and our results suggest that most of the regulation in complex traits lies in *trans*, cell type-specific eQTLs.

Predicting and interpreting functional non-coding regulatory variants with base-resolution deep learning models of chromatin accessibility

Authors: A. Pampari[1], A. Shcherbina[1], S. Kundu[1], E. Kvon[2], M. Kosicki[3], B. Zhang[1], K. Kuningas[4], Z. Chen[1], S. Deshpande[1], A. Patel[1], G. Marinov[1], E. Kotler[1], K. Alasoo[4], L. Pennacchio[3], A. Kundaje[1]; [1]Stanford Univ., Stanford, CA, [2]Univ. of California, Irvine, Irvine, CA, [3]Lawrence Berkeley Lab, Oakland, CA, [4]Univ. of Tartu, Tartu, Estonia

Abstract:

Genome-wide association studies (GWAS) have identified thousands of non-coding loci associated with diverse traits and diseases. Identifying the causal variants in these loci that disrupt transcription factor (TF) binding, chromatin state and gene expression in disease-relevant cellular contexts is challenging. Molecular QTL mapping can provide insights into regulatory and transcriptional impact of variants but is limited to estimating common variant effects in accessible cell types and tissues from typically underpowered cohorts lacking genetic diversity. Predictive models of regulatory DNA trained on reference molecular profiles across diverse cell contexts, offer a scalable alternative for prioritizing regulatory variants via in-silico mutagenesis.

Here, we present ChromBPNet, a compact deep learning sequence model of base-resolution chromatin accessibility profiles. Coupled with advanced interpretation methods, ChromBPNet can (1) assess the impact of variants on chromatin accessibility and TF footprints, and (2) identify the causal sequence features and higher-order syntax affected by these variants.

We evaluate ChromBPNet's performance against a state-of-the-art model called Enformer on carefully curated benchmark datasets spanning molecular QTLs (bQTLs, caQTLs, dsQTLs) across multiple cell types (LCLs, microglia, smooth muscle cells) and populations (European and African ancestry), reporter experiments, CRISPR screens, and fine-mapped GWAS variants (cardiovascular diseases, blood traits, colorectal cancer). ChromBPNet consistently matches or outperforms Enformer in prioritizing functional variants and predicting their effects, even after correcting critical flaws in Enformer's variant scoring method. ChromBPNet’s performance is also consistent across different ancestry groups.

Additionally, ChromBPNet models trained on scATAC-seq data from mouse vasculature can predict and interpret human variant effects in coronary artery disease GWAS loci. Models using scATAC-seq profiles from fetal heart and brain samples help prioritize rare non-coding variants and their target TFs, genes, and cell types in rare congenital heart diseases and neurodevelopmental disorders. We validate several predictions using CRISPR-Cas9 experiments in iPSC-derived cell types and in vivo tissue-resolved enhancer activity assays in transgenic mice. ChromBPNet models of diverse cell types and tissues also enable prioritization and interpretation of putative functional variants in ancient human genomes.
In summary, ChromBPNet is a powerful new tool for non-coding variant prioritization and interpretation.

Session 22: New Frontiers in Multi-ancestry Methods for Complex Traits

Location: Four Seasons Ballroom 1

Session Time: Wednesday, November 6, 2024, 10:15 am - 11:45 am

Multi-trait and multi-ancestry genetic analysis of comorbid lung diseases and traits improves genetic discovery and polygenic risk prediction

Authors: Y. He[1], W. Lu[2], Y. Jee[3], Y. Wang[4], K. Tsuo[5], J. Byun[6], C. Amos[6], M. Cho[7], M. Moll[8], A. Martin[4]; [1]Harvard Med. Sch., Boston, MA, [2]Broad Inst., Cambridge, MA, [3]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA, [4]Massachusetts Gen. Hosp., Boston, MA, [5]Broad Inst. of MIT and Harvard, Cambridge, MA, [6]Baylor Coll. of Med., Houston, TX, [7]Brigham and Women s Hosp., Duxbury, MA, [8]Brigham & Women s Hosp., Boston, MA

Abstract:

Respiratory diseases like COPD and asthma are leading causes of morbidity and mortality. While these diseases share many risk factors, most studies to date investigate single traits in predominantly European ancestry groups. Here, we develop and validate PRS-xtra (cross TRait and Ancestry), a multi-trait and multi-ancestry polygenic risk score approach, for predicting respiratory diseases. PRS-xtra jointly models the genetic correlation of 8 traits-COPD, asthma, lung cancer, FEV1, FVC, FEV1/FVC, smoking status and quantity-and the linkage disequilibrium and allele frequencies across African (AFR), Admixed American (AMR), East Asian (EAS), and European (EUR) ancestry groups to improve polygenic prediction.

We first conducted the largest meta-analysis of lung function in EAS using data from the Korean Cancer Prevention Study II and Taiwanese Biobank (N=132K), then leveraged these results with the largest and most diverse existing GWAS from the Global Biobank Meta-analysis Initiative, UK Biobank, and GSCAN Consortium, to conduct multi-trait analysis of GWAS (MTAG) in each population. Next, we implemented PRS-CSx to model LD across populations, deriving 39 scores. We constructed PRS-xtra using regularization to penalize individual scores and their effects in N=289K from All of Us and evaluated PRS-xtra against trait and ancestry specific polygenic risk scores (PRS) in a held out set (N=124K).

In conducting the largest meta-analysis for lung function in EAS, we identified 44, 73, and 31 independent loci for FEV1, FVC, and FEV1/FVC, respectively, of which 25 are novel. By modeling genetic correlations across traits through MTAG, we identified 609 total new loci related to respiratory traits across populations. While PRS-xtra and PRS were significantly correlated with each other (r=0.444, 0.185, and 0.120 for asthma, COPD, and lung cancer,

respectively), PRS-xtra showed substantial predictive improvements, especially in non-EUR populations. For example, PRS-xtra significantly improved predictive accuracy of asthma and COPD in AMR compared to PRS, with AUCs increasing from 0.509 to 0.630 (P<0.001) and 0.513 to 0.628 (P<0.001), respectively. PRS-xtra also better predicted COPD exacerbations compared to PRS, with AUC increasing from 0.572 to 0.600 (P<0.001). In conclusion, we conducted the most powerful multi-trait and multi-ancestry genetic analysis of respiratory diseases and auxiliary traits to date. We propose PRS-xtra as a method to model genetic correlation across traits and populations at the SNP level, demonstrating significantly better disease prediction—a critical step towards more equitable and generalizable preventative medicine.

Dissecting ancestry-aware molecular causal effects for type 2 diabetes

Authors: O. Bocher[1], A. Arruda[1,2], S. Yoshiji[3,4], C. Zhao[5], X. Yin[6,7], D. Cammann[8], H. Taylor[9,10], J. Chen[11], R. Mandla[3,12], A. Huerta[3], T-Y. Yang[4], K. Suzuki[13], A. Wood[14], F. Matsuda[4], J. Flannick[15,3], J. Mercader[3,12,16], C. Spracklen[5], J. Meigs[17,3,16], J. Rotter[18], M. Vujkovic[19,20], B. Voight[19,20], A. Morris[21], E. Zeggini[1,2]; [1]Helmholtz Zentrum Muenchen, Neuherberg, Germany, [2]Technical Univ. of Munich (TUM), Munich, Germany, [3]Broad Inst Harvard & MIT, Cambridge, MA, [4]Kyoto Univ. Graduate Sch. of Med., Kyoto, Japan, [5]Univ. of Massachusetts, Amherst, MA, [6]Nanjing Med. Univ., Nanjing, China, [7]Univ. of Michigan, Ann Arbor, MI, [8]Univ. of Nevada, Las Vegas, MD, [9]Natl. Human Genome Res. Inst., Bethesda, MD, [10]Univ. of Cambridge, Cambridge, United Kingdom, [11]UNLV-NIPM, Las Vegas, NV, [12]Massachusetts Gen. Hosp., Boston, MA, [13]Tokyo Univ, Tokyo, Japan, [14]Baylor Coll. of Med., Houston, TX, [15]Boston Children s Hosp., Boston, MA, [16]Harvard Med. Sch., Boston, MA, [17]Massachusetts Gen Hosp, Boston, MA, [18]Lundquist Inst. for BioMed. Innovation at Harbor-UCLA Med. Ctr., Torrance, CA, [19]Univ. of Pennsylvania, Philadelphia, PA, [20]Corporal Michael J. Crescenz VA Med. Ctr., Philadelphia, PA, [21]Univ. of Manchester, Manchester, United Kingdom

Abstract:

Multiple molecular mechanisms are involved in the pathogenesis of type 2 diabetes (T2D), with potentially different effects across ancestries. Recent large-scale efforts by the T2D Global Genomics Initiative (T2DGGI) have generated novel insights into the genetic architecture of T2D. These findings can be subsequently used to pinpoint molecular mechanisms leading to T2D in an ancestry-aware manner. In this work, we sought to explore the causal effects of gene and protein expression levels on T2D across ancestries by leveraging data from the latest T2DGGI multi-ancestry genome-wide association study (GWAS). We conducted two-sample Mendelian randomization (MR) analysis using blood

cis-expression and protein quantitative trait loci (eQTL/pQTL) derived from various datasets across four major ancestries. We then performed meta-analyses across ancestries and defined significance at a 5% FDR threshold. To corroborate our findings, we investigated evidence of colocalization using PWCoCo. We also performed MR analyses using eQTL data from other T2D relevant tissues, such as pancreatic islet or adipose tissue. We found evidence for causal effects of the genetically regulated levels of 78 genes and 2 proteins on T2D risk in the cross-ancestry meta-analysis. Both proteins, NELL1 and ALDH2, show a causal effect on increasing T2D risk. Additionally, we found that 249 genes and 11 proteins have a significant causal effect on T2D in ancestry-specific analyses but not in the meta-analysis. Among the ancestry-specific significant causal genes, 6 and 12 were specific to the African and Hispanic ancestry groups, respectively. Similarly, the causal effects of 7 proteins were specific to the East-Asian population. Finally, only two signals, SNUPN and PTPN9, were consistently found from eQTL and pQTL MR in the European population, both with increased levels being protective against T2D. Our findings highlight the power of large-scale GWAS and multi-omics MR analyses to identify causal pathways involved in T2D risk. Our results show how meta-analyzing ancestry-specific MR analyses can help to uncover ancestry-specific and shared causal pathways. This work emphasizes the need for expanding investigations into non-European ancestry populations to better understand T2D etiology.

Quantifying genetic effect heterogeneity across ancestral populations

Authors: Y. Wu, J. Miao, S. Dorn, Y. Wu, J. Fletcher, Q. Lu; Univ. of Wisconsin-Madison, Madison, WI

Abstract:

Understanding genotype-phenotype associations in diverse populations is one of the most important topics in complex trait genetics. As GWAS sample sizes of non-European populations grow, it has become increasingly urgent to quantify and understand the genetic effect heterogeneity across ancestral populations. While some studies have highlighted genetic effect differences at individual loci or imperfect genetic correlations for the same trait between populations, the underlying mechanisms behind such heterogeneity beyond allele frequency and linkage disequilibrium (LD) differences remain poorly understood. Here, we introduce a statistical framework named X-PIGEON that can produce robust estimates of genetic effect heterogeneity while accounting for a variety of technical confounders such as LD and allele frequency differences across ancestral populations. It takes multi-ancestry GWAS summary statistics and matched reference LD panels as input, and is highly computationally efficient to provide estimates for genetic

effect differences across populations. Applied to 192 complex traits with both European and East Asian GWAS summary statistics, X-PIGEON identified significant genetic effect heterogeneity for 68 traits, including anthropometric traits like height (p = 5.73E-81), body weight (p = 4.76E-142), and BMI (p = 5.56E-111), as well as disease outcomes such as prostate cancer (p = 1.27E-16), rheumatoid arthritis (p = 1.71E-5), Type 2 diabetes (p = 1.22E-38), colorectal cancer (p = 2.4E-13), breast cancer (p = 5.47E-12) , asthma (p = 1.22E-11), lung cancer (p = 2.81E-4), and stroke (p = 1.16E-11). In line with previous reports, we observed no significant heterogeneity for lipid traits such as HDL cholesterol, LDL cholesterol, total cholesterol, and triglycerides. X-PIGEON provides an urgently needed solution to quantify genetic effect heterogeneity across ancestries, and advances our understanding of population differences of genetic associations while adjusting for technical, non-biological differences. This framework is particularly timely given the increasing ancestral diversity in human genomic studies, and will become a vital tool in future follow-up analysis of multi-ancestry genetic association studies.

AI-STAAR: An ancestry-informed association analysis framework for large-scale multi-ancestry whole genome sequencing studies

Authors: W. Wang[1], L. Y. Zhou[2], D. Dutta[3], Y. Li[4], T. Sofer[5], N. Franceschini[6], Z. Li[7], J. G. Ibrahim[1], X. Li[1], NHLBI TOPMed Consortium Kidney Function Working Group; [1]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [2]Indiana Univ. Sch. of Med., Carmel, IN, [3]Natl. Cancer Inst., Rockville, MD, [4]Univ North Carolina, Chapel Hill, NC, [5]Beth Israel Deaconess Med. Ctr. / Harvard Med. Sch., Boston, MA, [6]Univ North Carolina at Chapel Hill, Chapel Hill, NC, [7]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA

Abstract:

Introduction

Large-scale whole genome sequencing (WGS) studies enable the detection of common and rare variants (RVs) associated with complex diseases or traits. With the increasing availability of WGS data representing participants from diverse populations, it is of interest to address heterogeneity in allelic effect sizes across ancestries to improve statistical power of association analyses and detect complex trait loci when the underlying causal variants are shared between ancestry groups with heterogeneous effects. Existing association analysis methods are limited in leveraging multi-ancestry variant effect heterogeneity, especially for under-represented ancestry populations.

Methods

We propose AI (Ancestry-Informed)-STAAR, a powerful and scalable association analysis

framework for ancestry- and functionally-informed genetic association analysis in biobank-scale multi-ancestry sequencing studies. AI-STAAR performs ancestry-informed association analysis to improve the power of single variant analysis for common variants and variant-set analysis for rare variants by modeling the potential heterogeneity through ensemble weighting informed by ancestry-specific variant allele frequencies and effect sizes, while accounting for population stratification and relatedness within and across ancestries. AI-STAAR further facilitates functionally-informed association analysis of both coding and noncoding RVs by incorporating multiple categorical and quantitative functional annotations for variant grouping and weighting.

Results

We applied AI-STAAR to perform WGS common and rare variant analysis of derived kidney function traits, estimate glomerular filtration rate (eGFR) and urine albumin-creatinine ratio (UACR), from the NHLBI TOPMed consortium. Among 45,090 and 18,869 participants with eGFR and UACR from diverse ancestries, AI-STAAR detected single variant 22-40220108-G-A for eGFR and 1-231196875-C-A for UACR, as well as RVs residing in *BAZ2A* enhancer regions and of *CIR1* UTR for UACR. These were missed by methods that do not account for heterogeneous ancestry effects. In addition to improved power for detecting associations accounting for effect size heterogeneity, AI-STAAR identifies the ancestry group(s) with strongest variant associations: 22-40220108-G-A for eGFR and 1-231196875-C-A for UACR were driven by East Asian and European ancestries, respectively; the RVs of *BAZ2A* and *CIR1* for UACR were African ancestry.

Summary

AI-STAAR is a powerful and computationally scalable framework that leverages allelic heterogeneity by ancestry for genetic association analysis in multi-ancestry sequencing studies.

A multi ethnic meta analysis of genome wide association studies identified additional novel genomic loci associated with cervical cancer ⭐

Authors: A. Kamiza[1], J-T. Brandenburg[2], M. Ramsay[2], C. Mathew[2]; [1]Kamuzu Univ. of Hlth.Sci., Blantyre, Malawi, [2]Univ. of Witwatersrand, Johannesburg, South Africa

Abstract:

Introduction: Cervical cancer is the second most common cause of cancer deaths among women worldwide. Genome-wide association studies (GWAS) have identified genetic variants associated with cervical cancer. However, these discoveries have been limited by the small sample sizes of individual cohorts. To identify additional novel loci associated with cervical cancer, we performed a multi-ethnic meta-analysis of GWASs. We also performed fine-mapping and polygenic risk scores. **Methods:** GWAS summary data were

obtained from women of European ancestry (UK Biobank, Finland, Estonia Biobank, Germany, and Kaiser), African ancestry (Johannesburg Cancer Study and Pan UK Biobank), and Asian ancestry (Biobank Japan). Meta-analyses were performed using fixed effect inverse variance weighted method implemented in GWAMA. To localize putative genomic loci associated with cervical cancer, we performed fine mapping using the Bayesian approach by calculating the marginal posterior probability of causality for each single nucleotide polymorphisms (SNPs) and 99% credible set size. We also developed and assessed the performance of polygenic scores using PRScs. **Results:** We identified 25 independent loci associated with cervical cancer. Of these loci, nine were novel. These include rs111611884 (OR=1.20, 95%CI=1.15-1.26, P=1.53E-13), rs41560220 (OR=0.70, 95%CI=0.65-0.75, P=7.97E-16), rs9266265 (OR=1.64, 95%CI=1.59-1.70, P=2.91E-34) , rs188481108 (OR=0.89, 95%CI=0.79-0.98, P=2.98E-23), rs1200371217 (OR=1.09, 95%CI=1.06-1.13, P=1.11E-08), rs763308335 (OR=1.59, 95%CI=1.55-1.62, P=2.29E-13) , rs12660769 (OR=1.16, 95%CI=1.12-1.20, P=6.22E-10), rs2854260 (OR=0.91, 95% CI=0.86-0.95, P=6.91E-14), and rs461807(OR=0.86, 95%CI=0.83-0.89, P=1.03E-09). These loci are implicated in various carcinogenic pathways. Our Bayesian fine mapping identified five loci with a marginal posterior probability of causality more than 0.99 and reduced the 99% credible set sizes for genomic loci. Moreover, our PRS derived from multi-ancestry summary data performed and predicted better than the PRSs derived from ancestry-specific data. **Conclusion:** We identified additional novel loci associated with cervical cancer and our fine mapping identified genomic loci with a high posterior probability of being causal.

A multi-ethnic reference panel to impute classical and non-classical *HLA* class II alleles: Enhancing HLA Imputation Accuracy in Admixed Populations

Authors: N. Silva[1,2], S. Bourguiba-Hachemi[1], S. H. Y. Knorst[3], R. T. Carmo[3], C. Masotti[3], D. Meyer[4], M. Naslavsky[4], Y. A. O. Duarte[4], M. Zatz[4], P-A. Gourraud[1], S. Limou[1], E. Castelli[2], N. Vince[1]; [1]Nantes Université, INSERM, Ecole Centrale Nantes, CR2TI, UMR 1064, Nantes, France, [2]UNESP - Molecular Genetics and Bioinformatics Lab., Botucatu, Brazil, [3]Dept. of Molecular Oncology, Hosp. Sírio- Libanes, São Paulo, Brazil, [4]Univ. of São Paulo, São Paulo, Brazil

Abstract:

HLA genes are vital for immune system modulation and activation. Some *HLA* variants are linked to diseases and drug response, and *HLA* polymorphisms significantly impact the outcome of transplants. Algorithms such as HIBAG can predict *HLA* alleles from SNPs

genotyped in array platforms in genome-wide association studies (GWAS), reducing the time and cost of *HLA* typing. However, the high *HLA* polymorphism and the lack of non-European reference panels make imputation challenging in diverse and admixed populations. The SNP-HLA Reference Consortium (SHLARC) aims to improve HLA imputation and enhance association studies in worldwide populations.
We obtained SNP genotypes and *HLA* alleles from the 1,000 Genomes Project (1KG, n=2,504) and the SABE cohort (Brazilians, n=1,170) using the hla-mapper workflow. For imputation, we used HIBAG to develop models and predict *HLA* alleles. Specifically, we evaluated the imputation of classical and non-classical *HLA* class II genes, including TAP1 and TAP2, using three reference panels, 1KG and SABE, and combined both in a single panel (full). We cross-validated each dataset by creating multiple reference panels and testing set pairs. This involved performing 10 random resampling of 200 samples from all biogeographic regions (Africa, America, East Asia, South Asia, Europe, and Brazil).
The full model achieved the highest overall imputation accuracy across all regions (95% to 100% in most cases). However, *HLA-DRB1* had the lowest accuracy, particularly among Americans (83%). Consequently, the F1 score, which balances the coverage of specific alleles and the proportion of correct calls, was notably low for *HLA-DRB1*, reflecting challenges in predicting rare alleles. We also conducted imputation for classical *HLA* class II genes in an independent Brazilian cohort (n=192) using our models and the Michigan Imputation Server. The SABE model outperformed the 1KG and Michigan Server in *HLA-DPB1* predictions, achieving 97% accuracy compared to 93% and 92.7%, respectively, and increased the full model's accuracy to 95%. When compared to the Michigan Imputation Server, which uses ~20,000 samples in the reference panel, our panels showed superior performance, particularly for *HLA-DQA1*, improving accuracy from 67.9% (Michigan) to 96.1% (full and 1KG).
Our findings underscore the importance of enhancing reference panels to reflect the genetic diversity found in different populations. This is crucial for improving HLA imputation accuracy and expanding our knowledge of immunogenetics across different populations. Moreover, it highlights the importance of a suitable method to call SNPs and *HLA* alleles for building suitable reference panels.

Session 23: Not Only Genetics: Integrating Other Omics Approaches

Location: Four Seasons Ballroom 4

Session Time: Wednesday, November 6, 2024, 10:15 am - 11:45 am

Single nucleus multiome optimizations for postmortem human brain and large scale multiome profiling of Alzheimer's disease progression reveal novel gene regulatory mechanisms and effects of APOE

Authors: A. Turner, S. Menon, S. Chang, A. Shah, A. W. Johnson, M. Chidrawar, M. R. Corces; Gladstone Inst.s, San Francisco, CA

Abstract:

Alzheimer’s Disease (AD) is a common neurodegenerative disease characterized by amyloid plaques and neurofibrillary tangles and this pathology correlates with cognitive decline. AD pathogenesis involves changes to cellular state across many cell types and brain regions. Genetic modifiers such as *APOE* exacerbate accumulation of amyloid plaques and accelerate cognitive decline. The aims of this project are twofold. First, we set out to generate a comprehensive dataset of matched single cell chromatin accessibility and gene expression from a large cohort of human AD patients in disease-relevant brain regions across differing pathologies, levels of cognition, and *APOE* genotype to explore these combinatorial effects and prioritize candidate functional noncoding AD risk variants. Second, due to the size of this dataset our goal was to perform methodological optimizations for both single nucleus data quality, experimental design, and computational analyses. We first performed pilot experiments to benchmark different nuclear isolation protocols and buffer compositions. In addition to standard ATAC-seq and single-nucleus (sn)RNA-seq quality control, we also considered often overlooked metrics such as levels of ambient RNA that often confound scRNA-seq analysis. We generated a large paired Multiomic single nucleus atlas from 144 individuals (80 AD, 64 control) and 2 brain regions (hippocampus, posterior parietal cortex) involved in AD pathogenesis. Our nuclei isolation protocol optimizations confirmed low ambient RNA and high quality snATAC-seq/snRNA-seq data for frozen human tissue. Our atlas of joint chromatin accessibility and gene expression (>1.5 million high-quality nuclei) has highlighted novel noncoding regulatory mechanisms across different *APOE* genotypes and pathologies. We have obtained over 500,000 hippocampal nuclei, a region not profiled in many human AD studies. To mitigate batch effects, we implemented an experimental strategy that leverages the genotype of each donor. In this strategy we pooled nuclei across individuals/brain regions before input

into the 10x Genomics protocol and have benchmarked computational tools to optimally link sequencing reads in the pool back to the original samples. We have also systematically observed data quality is substantially better using nuclei immediately isolated from frozen brain compared to nuclei cryopreserved and stored at -80°C. We have made key optimizations to improve single nucleus multi-omic data quality from postmortem frozen human brain and the corresponding large-scale atlas of AD human brain samples has identified novel cell-type specific noncoding DNA drivers of AD.

An Atlas of Protein Quantitative Trait Loci in Olink and Somascan Platforms Uncover Genetic Insights into Gastroenterological and Hepatological diseases

Authors: C. Khunsriraksakul[1], H. Markus[2], S. Chen[2], L. Wang[2], D. Chen[2], L. Carrel[2], B. Jiang[2], D. J. Liu[2]; [1]Johns Hopkins Hosp., Baltimore, MD, [2]Penn State Coll. of Med., Hershey, PA

Abstract:

Identifying protein quantitative trait loci (pQTL) is important for understanding the genetic basis of protein expression and its impact on various phenotypes. So far, there have been a few large-scale studies profiling both genetic information and plasma protein levels using Olink and Somascan platforms. Yet, little efforts exist to meta-analyze these datasets to maximize power. We present a comprehensive meta-analysis of pQTL data, encompassing up to 46,000 European individuals with 2,941 Olink targets from the UK Biobank cohort and up to 45,000 European individuals with 5,880 Somascan targets from the deCODE, AGES, INTERVAL, and KORA cohorts. Our Somascan meta-analysis reveals a 25% increase in the number of sentinel pQTLs, with 38,797 identified compared to 30,962 in the previous largest Somascan pQTL study by deCODE. The meta-analysis results are further replicated in the independent EPIC-Norfolk and Fenland cohorts. Summary-statistic-based inter-platform genetic correlations via LDSC demonstrate comparable estimates to previously calculated individual-level-based inter-platform correlations from the Iceland 1K cohort (Spearman correlation of 0.73, P value < 2.2E-16). To enhance the discovery of trans pQTLs, we perform meta-analysis of significantly genetically correlated (FDR P value < 0.05) Olink and Somascan targets (959 pairs with matched UniProt IDs) using MTAG. MTAG analyses result in a 35% increase in identified sentinel pQTLs for Olink (19,952 vs. 14,834) and a 101% increase for Somascan (23,531 vs. 11,733). Additionally, we conduct conditional analysis via GCTA and statistical fine-mapping via SuSiE to pinpoint causal genetic variants influencing plasma protein levels. We provide extensive insights into the genetic architecture and upstream regulators of plasma protein levels through the application of

transcriptome-wide association studies with PUMICE-derived gene expression prediction models from GTEx v8. Tissue enrichment analysis of the plasma proteome highlighted the liver and terminal ileum as the top enriched tissues, consistent with findings from the Human Protein Atlas regarding the tissue origin of the blood proteins. Finally, applying proteome-wide association studies and proteome-wide Mendelian randomization to 13 gastroenterological and hepatological diseases reveals several implicated proteins, including TNFR and ERBB2 in Crohn's disease, CHRDL2 and IL5RA in colorectal cancer, NSF in pancreatic cancer, and GGT1 in metabolic dysfunction-associated steatotic liver disease. These findings highlight potential pathogenic mechanisms, novel biomarkers, and therapeutic targets for these diseases.

Computational Analysis of Microbiome Genetics in Head and Neck Squamous Cell Carcinoma

Authors: K. Ansingkar[1], K. Yadav[1], S. Khader[2]; [1]Texas A&M Sch. of Engineering Med., Houston, TX, [2]Sch. of Publ. Hlth., Faculty of Med., Imperial Coll. London, London, United Kingdom

Abstract:

Squamous cell carcinoma of the head and neck (HNSCC) refers to cancers affecting the squamous cell mucous membranes of the nasal, oral, pharyngeal, and laryngeal cavities. This can be narrowed down into two main categories: human papillomavirus (HPV) induced and HNSCC not induced by HPV, otherwise known as HPV-positive and HPV-negative HNSCC that carry significant genomic differences. Interestingly, the population with HPV-positive HNSCC is often skewed towards younger individuals, those of Caucasian descent, and a higher socioeconomic status. Previous studies have indicated both HPV-positive or PHV-negative HNSCC conditions display a significant presence of Lactobacillus gasseri/johnsonii and Lactobacillus vaginalis in the saliva of these patients. Importantly, Lactobacillus derivatives are responsible for the healthy regulation of the vaginal microbiome.The predominant metabolic pathways in Homo sapiens, Lactobacillus gasseri/johnsonii, and Lactobacillus vaginalis were compared using the comparative analysis tool in the BioCyc database. Subsequently, gene pathway analysis was conducted using Enrichr, which identified relevant genes contributing to HNSCC formation. This gene pathway list was cross checked with the microorganism pathway results to identify which pathways overlap.
Common pathways shared by the microbiota and disease pathogenesis included cysteine and methionine metabolism, pyruvate metabolism, glycolysis, and portions of the DNA

mismatch repair pathway including purine and pyrimidine biosynthesis, phosphorylation, dephosphorylation, and salvage. Lactobacillus gasseri was identified to be primarily involved in the biochemical modification of macromolecules, such as post translational modification of lipoproteins. Pathways identified with Lactobacillus vaginalis were amino acid metabolism, co-factor and carrier protein biosynthesis, and carbohydrate degradation. Lactobacillus johnsonii was identified to be involved in the degradation of nucleosides and nucleotides, contributing significantly to the functions of Lactobacillus gasseri/vaginalis. While not all of these pathways are directly correlated with genetic contributions to HNSCC development, their functions can be appreciated in assisting the proliferation of tumor cells. The increased growth of these microbiota in the primary tumors of HNSCC strongly suggests their involvement in the pathogenesis process, presumably through altered tumor metabolism. Further studies need to be conducted to isolate these pathways and identify potential mechanisms of the progression of HNSCC through the microbiota.

Prioritizing Alzheimer's Disease genetic risk variants with massively parallel reporter assays and 3D chromatin structure

Authors: M. Bond[1], I. Quiroga[2], S. D'Costa[1], J. Bell[1], I. Sahasrabudhe[1], J. McAfee[1], K. Reed[3], N. Kramer[1], S. Lee[1], M. Patrucco[1], Y. Wu[1], H. Won[1], D. Phanstiel[4]; [1]UNC Chapel Hill, Chapel Hill, NC, [2]Univ. of North Carolina, Chapel Hill, NC, [3]Univ. of North Carolina, Chapel Hill, Chapel Hill, NC, [4]UNC, Chapel Hill, NC

Abstract:

Alzheimer's Disease (AD) is the leading cause of dementia world-wide, affecting over 55 million people. Genome-wide association studies (GWAS) have identified 75 loci associated with AD; however, interpreting these loci is difficult for many reasons. Multiple variants fall within linkage disequilibrium (LD) and are inherited together, making it challenging to identify the causal variant. Further, a majority of variants are located within the noncoding genome and it is unclear which genes they are affecting. Quantifying the regulatory impact of disease-associated variants using Massively Parallel Reporter Assays (MPRAs) can identify putative causal variants and mapping 3D chromatin interactions can help identify their target genes. However, due the context-specificity of regulatory features it is critical to map these features in disease-relevant cell types and conditions.
Since AD risk variants are enriched in the open chromatin regions of innate immune cells, we performed MPRAs for 3,576 AD-associated variants in resting and activated human macrophages. We identified 379 active regulatory elements including, 226 which exhibited

activity in only one of the two conditions, and 181 variants that induced a significant (mpralm, p-adj < 0.05) change in regulatory activity. Integrating these results revealed 53 enhancer-modulating variants (emVars) that were active in at least one condition and exhibited allelic differences in regulatory activity.
To further understand the regulatory impact of these active and allelic variants, we mapped transcriptional regulatory networks in resting and activated iPSC-derived microglia-like cells using Hi-C, ATAC-seq, H3K27ac CUT&RUN, and RNA-seq. We leveraged all genomic data into the ABC model to infer condition-specific enhancer-promoter pairs. Intersection of these enhancer promoter pairs existing microglia eQTLs, and our emVars allowed us to connect 43 of our emVars to 77 putative target genes. These genes were enriched for the disease-associated microglia (DAM) gene signature that is found in AD brains (fisher test, p = 0.0001). Among our putative target genes were prominent AD risk genes such as APOE and BIN1, as well as less established genes including ZYX and STAG3. We have used Electrophoretic Mobility Shift Assay (EMSA) to confirm that one emVar also induces a loss of transcription factor binding and are in the process of testing two more. By quantifying the change in regulatory activity of AD risk variants and mapping 3D chromatin structures in resting and activated macrophages/microglia, we have identified putative AD risk variants and genes for further research and therapeutic development.

Multi-omic Profiling in a 61 day Pig Kidney to Human Decedent Xenotransplant Reveals a Concerted Acute Rejection Immune Response

Authors: B. Keating[1], E. Schmauch[2], A. Dowdell[3], M. Mohebnasab[4], A. Stukalov[5], M. Kellis[6], J. Boeke[7], R. Montgomery[1], B. Piening[8]; [1]NYU Grossman Sch. of Med., New York, NY, [2]Broad Inst., Cambridge, MA, [3]Earle A. Chiles Res. Inst., Providence Hlth.,, Portland,, OR, [4]Univ. of Pittsburgh, Pittsburg, PA, [5]Seer, Redwood, CA, [6]MIT, Brookline, MA, [7]New York Univ. Langone, New York, NY, [8]Earle A Chiles Res. Inst., Portland, OR

Abstract:

Background: Only 1 in 4 people on solid-organ waitlists in the United States will receive an organ. Xenotransplantation is a promising approach to address human organ shortages. We previously pioneered brain-dead decedent human recipient models to test gene-edited pig kidney xenografts in 3 day studies. We recently extended this procedure to 61 days, facilitating assessment of adaptive immune responses and deep tissue and blood multi-omic profiling across dense longitudinal timepoints. **Methods**: Long-read whole genome sequencing, analyses of Xenium spatial transcriptomic data from xenograft biopsies (n=7 timepoints), both bulk and single-cell RNA-sequencing of PBMC (n=30), B and T cell

repertoire sequencing (n=27), and deep proteomic profiling (n=63) across the 61-day procedure were performed. **Results:** Blood NK and dendritic cell frequencies increased beginning postoperative day (POD)12, this was concordant with human immune cell infiltration into the xenograft tissue, B-cell receptor clonotype and T-cell receptor expansion, and higher levels of shared clonotypes present at the biopsy-confirmed antibody-mediated rejection (AbMR) episode at POD33. Kidney tissue injury involved pro-fibrotic epithelial and interstitial cells, with high *CXCL14* and *SPP1 (osteopontin)* expression at POD21. This continued into POD33, with hallmarks of a Type 1 immune response including *CCL19* expression and expansion of human immune cells in the xenograft, expressing *CXCL9, CXCL10,* and *CXCL11*. Blood protein signatures corresponding to both porcine and human complement activation indicated cross-species activation, which reduced following initiation of AbMR treatment (POD34). **Conclusions:** Integrative omics profiling enables extensive characterization of the human temporal immune response to a pig kidney xenograft, including an AbMR event and its subsequent successful treatment. We are actively applying these multiomic approaches to a living human in NYU who received a pig kidney xenograft in April 12th 2024.

Targeted CRISPRa/CRISPRi screen identifies functional variants and novel target genes at multiple renal cell carcinoma (RCC) susceptibility loci

Authors: T. Winter[1], T. Myers[1], L. Colli[2], L. Jessop[1], J. Choi[1], M. J. Machiela[1], M. P. Purdue[1], **K. Brown**[1], S. Chanock[1]; [1]Natl. Cancer Inst., Rockville, MD, [2]Ribeirao Preto Med. Sch. - Univ. of Sao Paulo, Sao Paulo, Brazil

Abstract:

In a previous genome-wide association study (GWAS) 13 renal cell carcinoma (RCC) risk regions were identified (Nat Comm 2017) and await functional characterization. Detailed investigation of how these regions function is required to reveal the underlying biological bases of disease susceptibility. While detailed study of four loci have implicated critical pathways in RCC, causal genes and pathways for most loci remain unidentified. To identify such causative variants and target genes, we evaluated a set of variants across the 13 loci plus 7 high interest loci (close by GWAS significance) with a Massively Parallel Reporter Assay (MPRA), eQTL analysis, capture Hi-C, and an arrayed CRISPR activation/inhibition (CRISPRa/CRISPRi) screen. The MPRA identified 196 variants across 19 regions with significant allele-specific effects, indicating cis-regulatory activity. Of these, 39 variants across 10 RCC loci displayed chromatin loops allowing physical interaction with the promoter of 24 nearby putative gene targets, as well as a significant cis-eQTL with the same

gene. To confirm the presence of a cis-regulatory relationship between these 39 candidate variants and 24 putative target genes, we performed an arrayed CRISPRa/CRISPRi screen in ACHN (RCC) and HEK293T (embryonic renal) cells covering each of the 10 nominated regions. Cells were stably transduced with inactive Cas9 fused to the ZIM3 transcriptional repressor or VP64 transcriptional activator domain and cis-regulatory relationships were assessed by TaqMan qPCR for the target genes. The screen identified multiple novel relevant target genes of RCC predisposition variants, including UCK2 at 1q24, ABL2 and TDRD5 at 1q25, ZEB2 and GTDC1 at 2q22, TFDP2 at 3q23, GPR37 at 7q31, AGBL3 and CALD1 at 7q33, and MAP2K1 at 15q22. At some regions, evidence indicates cis-regulatory relationships occur between a target gene and all variants predicted to be ‘causative’ at the region (e.g., UCK2 at 1q24 and ABL2 at 1q25). For other regions, cis-regulatory relationships occur between a gene and only one variant tested (e.g., GPR37 at 7q31 and TDRD5 at 1q25).This study identifies ten novel target genes across seven additional RCC susceptibility regions, adding to the four RCC susceptibility regions with strong published studies detailing the biological underpinnings of RCC susceptibility. This work represents a significant advance in understanding of the underlying biology of RCC susceptibility variants. Furthermore, ongoing investigation of the function of select genes in renal cells continues to provide further insight into the underlying biology at these regions.

Session 24: The Sex-Specific Landscape: Variation, Regulation, and Expression

Location: Room 505

Session Time: Wednesday, November 6, 2024, 10:15 am - 11:45 am

CHD1-deficiency shows sexual dimorphism mediated by androgen exposure

Authors: K. Anderson[1], S. T. Halldorsdottir[2], H. T. Bjornsson[2,3,1,4]; [1]Dept. of Genetics and Molecular Med., Landspitali Univ. Hosp., Reykjavik, Iceland, [2]Louma G. Lab. of Epigenetic Res., Faculty of Med., Univ. of Iceland, Reykjavik, Iceland, [3]McKusick-Nathans Dept. of Genetic Med., Johns Hopkins Univ. Sch. of Med., Baltimore, MD, [4]Dept. of Pediatrics, Johns Hopkins Univ., Baltimore, MD

Abstract:

Pilarowski-Bjornsson Syndrome (PILBOS, OMIM: 617682) is a neurodevelopmental disorder characterized by growth retardation, hypotonia, intellectual disability and autism spectrum disorder, caused by variants in the chromatin remodeler gene *CHD1*. To explore the mechanistic basis of this disorder we used CRISPR-Cas9 to generate a mouse model harboring a patient-specific variant (*Chd1*$^{R616Q/+}$) from a PILBOS patient. An *in vitro* nucleosome remodeling assay revealed a deficiency of chromatin remodeling of synthesized CHD1 harboring the variant compared to the wildtype protein, supporting loss of function as the disease mechanism. Interestingly, heterozygous mice displayed significant female-limited weight ($P<0.001$) and motor deficits ($P<0.01$), and increased anxiety-like behavior ($P<0.05$) compared to wildtype littermates. Mutant females also displayed elevated oxytocin levels in the hypothalamus ($P<0.05$). We hypothesized that androgens may protect against the phenotype in the *Chd1*$^{R616Q/+}$ males and to test this, we manipulated testosterone levels in mice of both genotypes. Consistent with a protective role of androgens, orchiectomy of male mice at postnatal day 15 unveiled a significant weight deficit in *Chd1*$^{R616Q/+}$ compared to WT littermates ($P<0.01$), which was previously masked in mutant males. Conversely, testosterone supplementation in female mice led to rescue of the growth phenotype. Using embryonic neural progenitor cells (NPCs) isolated from *Chd1*$^{R616Q/+}$ mice, we observed a significant proliferation defect compared to wild-type NPCs ($P<0.05$). We observed that testosterone treatment of pregnant dams was able to normalize proliferation rates in the embryonic mouse cortex ($P<0.001$). *In vitro* treatment of NPCs with dihydrotestosterone induced a widespread rescue of the transcriptional abnormalities observed in *Chd1*$^{R616Q/+}$ cells, suggesting that androgens may counteract CHD1 dysfunction by promoting the normal transcriptional program. Finally, using the

gnomAD database, a dataset mostly devoid of individuals with rare genetic disorders, we identified significant enrichment of rare missense alleles in *CHD1* within the male population compared to females (p = 8.13e-04), suggesting that males may be protected from the disease phenotype even at the population level. These findings unveil a novel mechanism underlying sexual dimorphism in PILBOS and pave the way for future investigations into sex-specific contributions in Mendelian causes of autism spectrum disorder.

Sex differences in brain cell-type specific chromatin accessibility in schizophrenia

Authors: Y. Ma, K. Girdhar, G. Hoffman, J. Fullard, J. Bendl, P. Roussos; Icahn Sch. of Med. at Mount Sinai, New York City, NY

Abstract:

Our understanding of the sex-specific role of the non-coding genome in serious mental illness remains largely incomplete. To address this gap, we explored sex differences in 1,393 chromatin accessibility profiles, derived from neuronal and non-neuronal nuclei of two distinct cortical regions from 234 cases with serious mental illness and 235 controls. We identified sex-specific enhancer-promoter interactions and showed that they regulate genes involved in X-chromosome inactivation (XCI). Examining chromosomal conformation allowed us to identify sex-specific cis- and trans-regulatory domains (CRDs and TRDs). Co-localization of sex-specific TRDs with schizophrenia common risk variants pinpointed male-specific regulatory regions controlling a number of metabolic pathways. Additionally, enhancers from female-specific TRDs were found to regulate two genes known to escape XCI, (XIST and JPX), underlying the importance of TRDs in deciphering sex differences in schizophrenia. Overall, we have created the largest resource to date and produced maps of sex-specific alterations in the epigenome of both neurons and non-neurons, focusing on changes in chromatin accessibility as well as other higher-order chromosomal conformations. These maps provide valuable insights into the mechanisms underlying sex-specific aspects of schizophrenia, particularly in relation to XCI. This enables the scientific community to further study sex-differentiated mechanisms underlying brain-related disorders.

Disentangling mechanisms underlying sex differences in gene regulation using population-scale multi-omics

Authors: J. Lake, E. Padhi, M. Degorter, P. Goddard, S. Montgomery; Stanford Univ., Stanford, CA

Abstract:

Many human diseases exhibit sex differences arising from complex interplays between genetic, hormonal, and environmental factors. For example, around 80% of autoimmune disease patients are female (46, XX), a disparity linked to the dosage of the X chromosome. At the molecular level, sex-biased gene expression is widespread and highly tissue-specific, with projects such as GTEx reporting sex-biased expression for 37% of genes in at least one tissue. Despite the prevalence of this effect, the gene regulatory mechanisms underlying these differences are not well understood. We hypothesize that the use of orthogonal modalities such as chromatin accessibility and transcriptional start site mapping can help to resolve the molecular mechanisms underlying sex-biased gene expression. Using a large RNA-seq (n = 600) and ATAC-seq (n = 100) resource of lymphoblastoid cell lines (LCLs, EBV-transformed B cells) derived from African individuals, we identified hundreds of autosomal genes with sex-biased expression without confounding by gendered environmental differences. To rule out potential false positives resulting from differences in the sampling distribution of reads, we examined the subset of sex-biased genes that replicate between RNA-seq and gene expression microarrays from HapMap (n=727) and found high replication ($\pi1 = 0.75$). We further identify 83 autosomal and 1,026 X-chromosome regions with sex-biased chromatin accessibility (FDR 10%). We are working to replicate regions with sex-biased chromatin accessibility in an independent cohort of LCLs from the GBR population (n = 100). To explore other mechanisms of gene regulation that may drive sex differences in RNA-seq data, we are investigating differences in transcription initiation using CAGE-seq data from LCLs of African origin (n = 108), which will reveal the sequence patterns that contribute to variability in promoter activity. Collectively, this work leverages the extensive functional genomics resources available in LCLs to explore the mechanisms that underlie sex-biased molecular phenotypes. We develop a putative map of sex-biased regulatory elements throughout the genome which will aid in the dissection of immune-related traits and diseases that exhibit sex differences.

Let's talk about sex: how biological sex affects functional variation across the genome to alter risk of human disease

Authors: A. Jones, T. Dalapati, G. Connelly, L. Wang, D. Ko; Duke Univ., Durham, NC

Abstract:

Background: Humans display sexual dimorphism across a variety of traits. Sex differences are broadly evident in the incidence, prevalence, severity, and treatment response in human disease. These differences are starkly prevalent in immunity where females are twice as likely to develop autoimmune disease but show greater resistance to most infectious agents than males. However, little is known about genetic mechanisms underlying sexual dimorphism and their impacts on disease susceptibility. **Methods and Results:** Here, we utilized single-cell RNA-seq of lymphoblastoid cell lines (240 females, 240 males) to discover mechanisms driving sex-biased differential gene expression and identify human genetic variants that regulate sex-biased gene expression and are associated with diverse sex-biased diseases. We found that the vast majority (79%) of sex-biased genes are targets of sex-biased transcription factors. Through linear modeling and RNAi knockdown datasets, we show that removing the sex effect of these transcription factors remediates sex differences in their downstream targets. In addition, we developed a unique two-step regression method to identify sex-biased expression quantitative trait loci (sb-eQTL) that affect expression of nearby genes across the genome (n=24,726; FDR 5%), including 94 genes on the X chromosome. We further discovered that these sb-eQTL share genetic architecture with over >500 phenotypes collected in the NHGRI-EBI GWAS catalog, many of which are highly sex-biased. We show that anthropometric traits such as body height are enriched for sb-eQTL and identify sex-specific effects at these loci through sex-stratified analyses of the UK Biobank. Lastly, we identify 19 sb-eQTL signals that colocalize with multiple sclerosis risk loci and affect sex-biased expression of disease-associated genes such as *VMP1*, *DDX6*, and *IKZF3*. **Conclusions:** Our results demonstrate that there are widespread genetic impacts on sexual dimorphism and identify possible mechanisms and potential clinical targets for sex differences in diverse diseases, including highly sex-biased autoimmune diseases such as multiple sclerosis. These results will support the development of personalized medicine that considers biological sex as a crucial variable in human disease.

Large-scale analyses of variants with sex-biased population allele frequencies, sex-biased association with phenotypes, and sex-biased allele penetrance in 43 human tissues

Authors: R. Soemedi[1,2], W. Hasuki[3], B. Spitzer[4], I. Chiquier[5], K. Hwa[5], S. Nakhawa[1], J. Rotter[6], S. Rich[7], D. Demeo[1], C. Lopes-Ramos[8], C. Kooperberg[9], A. Reiner[10], B. Mitchell[11], A. Morrison[12], M. Fornage[13], S. Redline[1], D. Levy[14], T. Sofer[15,1,16]; [1]Brigham and Women's Hosp., Boston, MA, [2]Brown Ctr. for BioMed. Informatics, Providence, RI, [3]Indonesia Intl. Inst. for

Life Sci., Jakarta, Indonesia, [4]Beth Israel Deaconess Med. Ctr., Boston, MA, [5]Brown Univ., Providence, RI, [6]Lundquist Inst., Harbor-UCLA Med Ctr, Torrance, CA, [7]Univ. of Virginia, Charlottesville, VA, [8]Harvard Med. Sch. Brigham and Womens Hosp., Boston, MA, [9]Fred Hutchinson Cancer Ctr., Seattle, WA, [10]Univ of Washington, Seattle, WA, [11]Univ Maryland, Baltimore, Baltimore, MD, [12]Univ. of Texas Hlth.Sci. Cntr Houston, Houston, TX, [13]Univ. of Texas Hlth.Sci. Ctr. at Houst, Houston, TX, [14]NHLBI/NIH, Framingham, MA, [15]Beth Israel Deaconess Med. Ctr. / Harvard Med. Sch., Boston, MA, [16]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA

Abstract:

Sex differences are fundamental elements of human diseases, but the genetic etiology and mechanisms remain elusive. Here, we present large-scale comprehensive analyses of variants with (1) sex-biased allele frequencies, termed “freqSBVs”, using data from 7 population groups in gnomAD, (2) sex-biased phenotypic associations, termed “pheSBVs”, using GWAS summaries of 43 phenotypes in UK Biobank, and (3) sex-biased allele expressions, termed “aeSBVs”, using allele-specific expression data of 43 tissue types in GTEx. The freqSBVs’ direction and magnitude of sex bias were highly correlated between gnomAD populations and were validated in TOPMed European and GTEx individuals. FreqSBVs appearing in ≥3 gnomAD populations were enriched in developmental genes. Autosomal freqSBVs in transcription factor binding sites were the most highly shared across populations, compared to freqSBVs in other functional regions. On the other hand, freqSBVs that were specific to one population were overrepresented in genes responsible for responses to environmental stimuli, including light stimulus, low oxygen levels, and temperature. PheSBVs were highly pleiotropic, particularly pheSBVs that were also freqSBVs (phe-freqSBVs). We observed profound positive or negative correlations of the magnitude of sex bias in phenotypic associations with pheSBVs and, to a higher degree, with phe-freqSBVs that were shared between phenotypes. The greatest enrichment for aeSBVs was in the splice-site regions, which increased with approaching distance to the splice-sites. Furthermore, the direction of sex bias for aeSBVs differed between male- and female-biased freqSBVs. Similarly, the direction of sex bias for aeSBVs differed between male- and female-biased pheSBVs in tissue-specific and phenotype-specific manner. Moreover, we found freqSBVs with opposite direction of sex bias between the younger (≤50 yo) and older (>50 yo) populations in TOPMed. These age-inversed freqSBVs showed age-inversed allele penetrance in GTEx tissues in tissue-specific manner, and they were enriched in genes associated with cellular aging and aging-associated diseases, including neurodegenerative disorders and many types of cancers. In conclusion, we showed that the three modes of discovery of sex-biased variants can be leveraged to understand the etiology and mechanisms of sex-biased disorders. Our study greatly aids the

understanding of the functional mechanisms of sex-divergent disease traits and facilitates the realization of sex-aware precision medicine.

GenESIS: enhancing transferability of polygenic scores with gene-by-sex interactions

Authors: Y. Tanigawa, M. Kellis; MIT, Cambridge, MA

Abstract:

Advancing precision medicine requires accurate prediction of disease liability and medically relevant traits from individuals' genetic, demographic, and environmental factors. A critical gap is the limited transferability of polygenic scores (PGS) across genetic ancestry groups.
We hypothesize that incorporating nonlinear and context-dependent effects, such as genetic dominance, gene-by-environment (GxE), and gene-by-sex (GxS) interaction effects, can better capture likely causal effects and improve the transferability of PGS.
Here, we present GenESIS (GENe, Environment, and Sex Interaction Score), the first unified predictive modeling framework integrating linear and nonlinear effects of millions of genetic variants, demography, environmental factors, and their interactions. As an initial application, we analyze n=406,659 individuals across the continuum of genetic ancestry in UK Biobank and develop predictive models for 99 traits, focusing on additive, dominance, and genome-wide GxS effects represented in 2,630,335 predictor variables.
The GenESIS models contain a median of 8.0% of predictors capturing GxS effects. We demonstrate that GenESIS's GxS effects are highly consistent with sex-stratified genome-wide associations, validating our approach.
We show that modeling GxS with GenESIS improves transferability. For predicting hip circumference in Africans, for example, GenESIS's predictive performance (R^2=.067) shows a statistically significant (p=8.0x10^{-7}) improvement over the additive-only inclusive PGS (R^2=.018) and outperforms all publicly available models for Africans in the PGS catalog. Similarly, 32 traits exhibit improved prediction, including body mass index for Africans and South Asians and total bilirubin for white British.
Lastly, we show GenESIS reveals biologically plausible hypotheses. Genetic variants with GxS effects have pleiotropic associations across sex-specific factors, such as associations between a missense variant in *GCKR* (rs1260326) and sex hormone-binding globulin levels and age at menopause. Genome-wide GxS effects for hip circumference are enriched for relevant biological processes in obesity, including negative regulation of interleukin-4 production (FDR=3x10^{-6}) and phosphofructokinase activity (FDR=1.5x10^{-3}), nominating

attractive targets for context-dependent interventions.
Overall, our results underscore the importance of mapping nonlinear and context-dependent effects, highlight the critical benefits of integrating such effects in improving PGS transferability, and, more broadly, pave the way for designing more inclusive and nuanced health interventions.

Session 25: Decoding Gene Expression Cis and Trans

Location: Room 401

Session Time: Wednesday, November 6, 2024, 1:15 pm - 2:15 pm

Colocalization of >1,200 skeletal muscle genes with GWAS loci for musculoskeletal and cardiometabolic traits: a muscle eQTL study of 1,002 individuals

Authors: E. Wilson[1], K. A. Broadaway[1], N. Narisu[2], S. M. Brotman[1], H. M. Stringham[3], M. R. Erdos[2], T. A. Lakka[4], M. Laakso[4], J. Tuomilehto[5,6,7], M. Boehnke[3], H. A. Koistinen[5,8,9], F. S. Collins[2], S. C. Parker[3,10,11], L. J. Scott[3], K. L. Mohlke[1]; [1]Dept. of Genetics, Univ. of North Carolina, Chapel Hill, NC, [2]Natl. Human Genome Res. Inst., NIH, Bethesda, MD, [3]Dept. of Biostatistics, Ctr. for Statistical Genetics, Univ. of Michigan, Ann Arbor, MI, [4]Inst. of Clinical Med., Univ. of Eastern Finland, Kuopio, Finland, [5]Dept. of Publ. Hlth.and Welfare, Finnish Inst. for Hlth.and Welfare, Helsinki, Finland, [6]Dept. of Publ. Hlth., Univ. of Helsinki, Helsinki, Finland, [7]Diabetes Res. Group, King Abdulaziz Univ., Jeddah, Saudi Arabia, [8]Dept. of Med., Univ. of Helsinki and Helsinki Univ. Hosp., Helsinki, Finland, [9]Minerva Fndn. Inst. for Med. Res., Helsinki, Finland, [10]Dept. of Computational Med. and Bioinformatics, Univ. of Michigan, Ann Arbor, MI, [11]Dept. of Human Genetics, Univ. of Michigan, Ann Arbor, MI

Abstract:

eQTL can identify candidate genes linked to GWAS traits via colocalization of shared variants. However, the modest sample sizes and typical focus of many eQTL studies on only primary eQTL signals obscures detection of eQTL and colocalizations. We conducted a meta-analysis of jointly identified conditionally distinct eQTL signals detected in 1,002 bulk skeletal muscle samples from the GTEx and FUSION studies. We identified 16,188 eQTL signals corresponding to 11,311 eGenes, 30% of which contained two or more signals ($p<1e-6$). The meta-analysis detected >35% more signals and >28% more eGenes than either study on its own. Primary signals were on average closer to the gene transcription start site, had larger effect sizes, and had higher minor allele frequencies than non-primary signals (all pdiff<3e-16). We isolated these muscle eQTL signals by conditioning each signal on all other signals and colocalized them with isolated conditionally distinct GWAS signals for 28 musculoskeletal and cardiometabolic traits. We identified 2,791 colocalized GWAS-eQTL signal pairs for 1,217 eGenes (lead variant LD $r2\geq0.5$, coloc $PPH4\geq0.5$). Non-primary eQTL signals accounted for 18% of the overall colocalizations, suggesting that they may play a key role in explaining GWAS associations. The colocalized GWAS-muscle eQTL signals included 110 BMI, 109 triglyceride, and 89 creatinine signals. To explore the effect

of tissue source on eQTL colocalizations for type 2 diabetes (T2D), we compared colocalizations using eQTL detected in three tissues analyzed using the same pipeline: muscle (n=1,002 samples; 68 genes colocalized with 46 T2D signals), subcutaneous adipose (n=2,256; 151 genes, 81 T2D signals), and liver (n=1,183; 44 genes, 37 T2D signals). Across tissues, 103/403 T2D signals (26%) colocalized with at least one eQTL signal. Only thirteen T2D signals colocalized with at least one eQTL signal in all three tissues, while seven T2D signals colocalized only with muscle eQTL, 38 only with adipose eQTL, and ten only with liver eQTL. Among the T2D-eQTL colocalizations only detected in muscle was a signal for *FBXL22* (eQTL p=2e-31, colocalization PPH4=0.85), which has been shown to play a role in muscle atrophy, suggesting a potentially muscle-specific mechanism underlying this T2D association. Together, we identified >16k distinct muscle eQTL signals, including signals for 1,217 genes that colocalized with relevant GWAS traits, and showed that T2D colocalizations with eQTL varied across tissues, suggesting that tissue remains a key factor in understanding the mechanisms of disease loci.

Identifying the molecular mechanisms of complex disease through a genome-wide *trans*-eQTL meta-analysis in 43,301 individuals

Authors: R. Warmerdam[1], A. van der Graaf[2], D. Zhernakova[1], T. van Lieshout[1], eQTLGen Consortium, T. Esko[3], B. Strober[4], Z. Kutalik[2], A. Price[4], H-J. Westra[1], L. H. Franke[1], U. Võsa[3]; [1]Univ. Med. Ctr. Groningen, Groningen, Netherlands, [2]Univ. of Lausanne, Lausanne, Switzerland, [3]Univ. of Tartu, Tartu, Tartumaa, Estonia, [4]Harvard Sch. of Publ. Hlth., Boston, MA

Abstract:

Gene expression quantitative trait loci (eQTLs) are frequently used to identify the mechanisms of disease-associated variants. However, previous studies have mainly reported on *cis*-eQTLs due to limitations in statistical power. While these *cis*-eQTLs are generally stronger and easier to detect than *trans*-eQTLs, they explain only 11% of disease heritability (Yao et al., 2020, *NG*). Therefore, the eQTLGen Consortium previously identified 59,786 *trans*-eQTLs in 31,000 blood samples, while testing for only 10,000 disease associated variants (Võsa et al., 2021, *NG*). Now, eQTLGen phase 2 aims to identify *trans*-eQTLs on a genome-wide scale and uncover the downstream effects of GWAS variants. Robust pipelines allowed us to perform automated, comprehensive data quality control and harmonized genetic imputation. An efficient pipeline adapted from the HASE framework allowed us to perform genome-wide meta-analyses for all blood-expressed genes, without having to share terabytes of data. Instead, the cohorts had to share just 50

to 500Gb of data, all the while ensuring participant privacy. We included 52 blood gene expression datasets, representing 43,301 individuals (including 4,790 of non-European ancestry). Colocalizations were performed using HyprColoc. ToppGene and cisTarget databases were used to test support for transcription factor (TF) and target gene links. We identified a significant ($p<2.5x10^{-12}$) *trans*-eQTL for 65% of the genes expressed in blood, reflecting a 2-fold increase over eQTLGen phase 1. We found that 1,235 *trans*-eQTLs colocalize with *cis*-eQTLs. Not only are the linked *cis*-eGenes enriched for TFs (1.94-fold increase), the motifs associated with these TFs are also strongly enriched around the linked *trans*-eGenes ($p=5.4x10^{-3}$). This indicates that large-scale *trans*-eQTL analyses allow us to identify biologically interpretable directed links among gene-pairs. We also observed that *trans*-eQTLs frequently colocalize with complex traits. For instance, two loci associated to IBD colocalize with both a *cis*- and *trans*-eQTL for *ETS2*, a gene causally linked to IBD. Replication in single-cell data and Mendelian Randomization analyses will ultimately validate our gene-disease links, helping to prioritize these genes as potential drug targets.
To our knowledge, this work presents the most complete blood eQTL study to date, yielding *trans*-eQTLs for twice as many blood-expressed genes compared to eQTLGen phase 1. Moreover, these *trans*-eQTLs allow us to break down the molecular mechanisms through which disease-associated variants exert their effect, and thus, provide a valuable resource for the genetics research community.

Polymorphic short tandem repeats shape single-cell gene expression across the immune landscape

Authors: H. Tanudisastro[1,2,3,4], A. S. E. Cuomo[1,2,3,5,6], B. Weisburd[7,8], M. Welland[1,2,3], M. Franklin[1,2,3], A. Xue[3,5,6], B. Bowen[3,5,6], E. Spenceley[3,5,6], M. Harper[1,2,3], K. Bobowik[1,2,3], B. Swapna Madala[1,2,3], C. Uren[1,2,3], H. Nicholas[1,2,3], E. Dolzhenko[9], C. Wallace[10], M. Gymrek[11,12], A. W. Hewitt[13,14,15], G. A. Figtree[16,17], K. M. de Lange[1,2,3], J. E. Powell[3,5,6], D. G. MacArthur[1,2,3]; [1]Ctr. for Population Genomics, Garvan Inst. of Med. Res., Sydney, NSW, Australia, [2]Ctr. for Population Genomics, Murdoch Children's Res. Inst., Melbourne, VIC, Australia, [3]Faculty of Med. and Hlth., Univ. of New South Wales, Sydney, NSW, Australia, [4]Faculty of Med. and Hlth., Univ. of Sydney, Sydney, NSW, Australia, [5]Garvan-Weizmann Ctr. for Cellular Genomics, Garvan Inst. of Med. Res., Sydney, NSW, Australia, [6]UNSW Cellular Genomics Futures Inst., Univ. of New South Wales, Sydney, NSW, Australia, [7]Program in Med. and Population Genetics, Broad Inst. of MIT and Harvard, Cambridge, MA, [8]Analytic and Translational Genetics Unit, Massachusetts Gen. Hosp., Boston, MA, [9]Pacific BioSci.s of California, Menlo Park, CA, [10]Univ. of Cambridge,

Cambridge, United Kingdom, [11]Dept. of Med., Univ. of California San Diego, La Jolla, CA, [12]Dept. of Computer Sci. and Engineering, Univ. of California San Diego, La Jolla, CA, [13]Menzies Inst. for Med. Res., Univ. of Tasmania, Hobart,TAS, Australia, [14]Dept. of Ophthalmology, Royal Hobart Hosp., Hobart, TAS, Australia, [15]Ctr. for Eye Res. Australia, Univ. of Melbourne, Melbourne, VIC, Australia, [16]Charles Perkins Ctr., The Univ. of Sydney, Sydney, NSW, Australia, [17]Kolling Inst. of Med. Res., Royal North Shore Hosp., Sydney, NSW, Australia

Abstract:

Short tandem repeats (STRs) are polymorphic genomic loci composed of repeating copies of a 1-6bp motif. There are over 1 million STR loci, comprising 3-5% of the human genome. STRs play a key role in regulating gene expression and biological function. STR variants cause over 60 Mendelian disorders, including Huntington's disease and Fragile X Syndrome, and influence complex traits such as height.

Studies of how STRs regulate gene expression have mainly relied on bulk tissue data, but advances in single-cell RNA sequencing (scRNA-seq) enable the study of cell type-specific effects of STRs at a finer resolution at population scale. The TenK10K project is a new initiative that will perform whole genome sequencing (WGS) and scRNA and ATAC-seq on peripheral blood mononuclear cells for 10,000 individuals, generating the largest set of paired human WGS and single-cell data to date.

We present preliminary single-cell expression quantitative trait STR loci (sc-eSTR) analysis from Phase 1 of TenK10K, comprising 5,084,027 immune cells from 1,790 individuals. Using WGS, we genotyped 2.6 million STR loci, selected using existing catalogs and polymorphism in samples from 1KGP and the Human Pangenome Reference Consortium. Across 28 immune cell types, we identified 51,832 unique sc-eSTRs associated with 14,325 genes (FDR < 5%), of which 70% were cell type-specific. Type 1 error was well-controlled, assessed by permutation. More cell type-specific sc-eSTRs were found in abundant cell types, but we detected cell type-specific sc-eSTRs in every cell type. We validated known eSTRs, including the association between (CGGGG)* and *CSTB* expression in whole blood, and further show that this association is significant in CD4+ CTL, CD14+ monocytes, pDC, and ILC cell types.

We performed comparisons with STR GWAS loci of 21 blood traits in UK Biobank, revealing 1,087 colocalized loci (posterior probability [PP] > 0.95), of which 21% were cell type-specific. For example, a sc-eSTR associated with *CARD9* expression, a regulator of innate immune activation, colocalized with higher levels of C reactive protein in CD16+ monocytes only.

Comparing the IIBDGC Immunochip GWAS for inflammatory bowel disease and sc-eSNPs from our dataset, we found 7 colocalizing loci (PP > 0.8) with an overlapping sc-eSTR signal

that was more significant than the top sc-eSNP. In all but one case, conditioning on the sc-eSTR removed all the signal in the region, suggesting that the STR is a strong candidate as the causal variant.
The discovery of immune cell type-specific sc-eSTRs opens new avenues for understanding the genetic drivers of immune function and disease, with further fine-mapping promising deeper insights.

Variation and regulatory mechanisms of the small RNA transcriptome across human tissues

Authors: P. Stojanov[1], T. Coorens[1], J. Fernandez del Castillo[1], S. Steelman[2,1], S. Young[1,2], C. Nussbaum[2,1], K. Ardlie[3], G. Getz[4], F. Aguet[5]; [1]Broad Inst. of MIT and Harvard, Cambridge, MA, [2]Cellarity, Somerville, MA, [3]Broad Inst., Cambridge, MA, [4]Broad Inst MIT & Harvard, Cambridge, MA, [5]Illumina, Inc., Foster City, CA

Abstract:

Standard RNA-sequencing protocols exclude small RNAs and thus preclude the study of thousands of small noncoding RNAs with essential roles in the post-transcriptional regulation of gene expression. Notably, the contribution of small RNAs to human complex traits and diseases remains poorly understood, and may elucidate trait associations for which underlying regulatory and transcriptional mechanisms remain unknown.
Here, we present the characterization of small RNAs across 16,814 samples, 47 tissue sites and 978 donors in the GTEx Project. We quantified the expression of a total of 41,458 small RNAs, including microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), transfer RNAs, small nuclear RNAs, small nucleolar RNAs, Y RNAs, and others. We used supervised classification to identify putative novel RNAs not present in references, and detected 57 novel high-confidence miRNAs. We mapped QTLs in cis and trans, identifying 100s to 1000s of cis-eQTLs for each small RNA species. Among them, we discovered two trans-QTLs for tRNAs, corresponding to splice QTLs in TRMT1 and DTWD1, which alter the base editing activity of their respective target tRNAs. To investigate the propagation of genetic effects on coding genes through miRNAs, we fine-mapped and co-localized SNPs that affect miRNA expression in cis and mRNA expression in trans. Using the miRNA eQTLs as instrumental variables, we perform mediation analysis to confirm that miRNA-mRNA pairs are causally related, which we further corroborate through seed-pairings and predicted binding affinity scores from a well-established model. This resulted in a set of such pairs whose interaction can elucidate the putative mechanism behind complex traits. For example, we identify an interaction between miR-5683 and ODAD1 in the cerebellum, a

gene involved in the motility of glial cells. Furthermore, we found that this shared causal variant colocalized with the GWAS trait of tau levels. We also found that the tissue specificity of miRNA expression is reflected in the tissue specificity of complex traits co-localizing with miRNA eQTLs.
In summary, we demonstrate the importance of characterizing the full spectrum of small RNAs, which play critical roles in the regulation of gene expression, including in development and disease.

Session 26: Genetic Approaches Informing Drug Targets and Mechanism

Location: Room 501

Session Time: Wednesday, November 6, 2024, 1:15 pm - 2:15 pm

Replication of genetic associations across diverse ancestry groups is indicative of drug target success in clinical trials ⭐

Authors: C. Eijsbouts, 23andMe Research Team, A. Auton, S. Pitts, X. Wang, W. Wang, Y. Jiang; 23andMe, Sunnyvale, CA

Abstract:

To identify novel candidate drug targets, genome-wide association studies to date have relied disproportionately on European-ancestry genetic data. The discovery of genetic associations missed in the European-ancestry population alone (e.g. rs59039403 for atopic dermatitis or rs900776 for lipid metabolism, identified in Japanese and African populations, respectively) demonstrates the value of diverse genetic data empirically, but systematic data showing how cross-ancestry analyses aid target discovery are lacking. Here, we systematically interrogate fine mapping results from six discretized genetic ancestry groups (European, Ashkenazi, East Asian, South Asian, African American, Latine) across 443 traits reported on by direct-to-consumer genetics research participants (23andMe, Sunnyvale, CA). We find that even variants which appear to be confidently finemapped in one ancestry group (PIP>0.9) are regularly not supported by fine-mapping evidence (PIP>0.1) in other ancestry groups, even if the variant is common (MAF>5%), we are powered to detect it, and the variant is genome-wide significant (p<5e-8) in the replication group (fine mapping replication rate ~50% across 244 binary traits). Importantly, our analysis avoids inter-population heterogeneity arising from differences in phenotyping, genotyping and imputation methodology found in recent cross-biobank analyses. We also find that replication, or a lack thereof, has a strong influence on the performance of targets in clinical trials. Across 162 traits matched to PharmaProjects data, targets supported by fine mapped genetic associations (p<5e-8, PIP>0.9) that replicate (p<5e-8) in a secondary sample, regardless of ancestry, are 1.5-3 times more likely to proceed from Phase I clinical trials to approval than targets without replicating support (e.g. RR=2.71, 95%CI=[2.15 - 3.26] for European:East Asian replication). This effect persists after accounting for effect size differences between replicating and non-replicating loci, which, in contrast to recent literature, we observe to correlate with target viability in clinical trials. We consider that

replication acts as a filter to eliminate false positives arising from fine mapping error as well as non-causal associations.

Prioritising New Antihypertensive Drug Targets and Unravelling Disease Modulation by Antihypertensive Drugs Using Mendelian Randomisation

Authors: N. Le[1], Q. Tran[1], D. Gill[2], S. Padmanabhan[1]; [1]BHF Cardiovascular Res. Ctr., Sch. of Cardiovascular and Metabolic Hlth., Univ. of Glasgow, Glasgow, United Kingdom, [2]Dept. of Epidemiology and Biostatistics, Sch. of Publ. Hlth., Imperial Coll. London, London, United Kingdom

Abstract:

Background: Understanding the effects of antihypertensive drugs on cardiovascular, diabetic, and renal outcomes is crucial for optimising treatment strategies. This study utilises Mendelian Randomisation (MR) to explore the causal impacts of antihypertensive drugs and genetically predicted systolic blood pressure (SBP) reduction on these outcomes. **Methods:** Two-sample MR was used to investigate the causal effects of genetically predicted SBP reduction and BP-lowering drug classes on coronary artery disease (CAD), myocardial infarction (MI), atrial fibrillation (AF), heart failure (HF), ischemic stroke, chronic kidney disease (CKD), and type 2 diabetes (T2D). This is followed by a summary-based MR to investigate if the observed associations were mediated through the transcription of genes encoding proteins targeted by the corresponding drug classes. Summary data were obtained from the largest European ancestry GWAS and GTEx v8. Data from eQTLGen consortium was used to perform replication analysis for the observed associations. HEIDI test and colocalisation analysis were conducted to examine if the observed associations were due to a shared causal variant or linkage scenario. **Results:** Genetically predicted lower SBP was associated with a reduced risk of all outcomes. Consistent with findings from RCTs, our study found established associations of calcium channel blockers, beta-blockers, and angiotensin-converting enzyme inhibitors with a range of cardiovascular diseases. Novel associations were observed between angiotensinogen inhibition and a decreased risk of CAD and ischemic stroke; endothelin receptor antagonists and a decreased risk of CAD and ischemic stroke; PDE5 inhibition and a lower risk of CAD, ischemic stroke, and CKD; sGC stimulation and a decreased risk of CAD, MI, and CKD. Genetically increased *GUCY1A3* expression in tibial artery was associated with lower SBP and decreased risk of CAD ($p_{SMR} = 1.74\times10^{-06}$; $p_{HEIDI} = 0.664$, $H_4 = 0.99$). Genetically increased *PDE5A* expression in aorta was associated with higher SBP and increased risk of CAD ($p_{SMR} = 9.12\times10^{-06}$, $p_{HEIDI} = 0.338$, $H_4 = 0.48$).

Genetically increased *KCNH2* expression in the brain cerebellum was associated with higher SBP and increased risk of AF ($p_{SMR} = 6.02\times10^{-05}$, $p_{HEIDI} = 0.195$, $H_3 = 0.99$). **Conclusion:** Our findings underscore the potential of integrating genetic and pharmacological data to identify novel antihypertensive drug targets and unravel the basis disease modulation by antihypertensive drugs. Our findings of the protective effect of *GUCY1A3* expression against CAD and the risk elevation associated with *PDE5A* expression point to potential therapies.

Identification of plasma proteins as promising therapeutic targets to treat hypertension

Authors: A. Chignon[1,2], **G. Lettre**[1,2]; [1]Montreal Heart Inst., Montreal, QC, Canada, [2]Université de Montréal, Montreal, QC, Canada

Abstract:

Hypertension (HTN) is a strong risk factor for cardiovascular diseases, affecting about 30% of adults worldwide. Despite the availability of medications, about 10% of patients have resistant hypertension. The generation of large-scale plasma proteomic datasets and its integration with human genetic results offer an opportunity to identify novel therapeutic targets for HTN. For our analyses, we used publicly available genome-wide association studies (GWAS) results for systolic (SBP) and diastolic blood pressure (DBP) measured in more than one million individuals, and GWAS results for the expression of 4,717 plasma proteins quantified in 35,559 individuals. First, we used Mendelian randomization (MR) to identify candidate causal proteins associated with SBP and DBP. We used this information in a Bayesian GWAS analysis to recalibrate association results with SBP and DBP, allowing us to define a causal plasma proteome for blood pressure (CP) that includes 12 proteins. In the UK Biobank data, we assessed the role of the CP on HTN and its pleiotropic effects by using a polygenic score-based phenome-wide association study (PS-PheWAS) framework on 811 phenotypes collected in about 500,000 individuals. The PS-PheWAS demonstrated a strong association of the CP with HTN (P=10e-55) and HTN-related diseases, including ischemic heart disease, coronary artery disease (P=10e-25), chronic kidney failure (P=10e-6), and hypothyroidism (P=10e-40). MR analyses revealed many concordant and causal associations between the CP and HTN-related diseases, such as myocardial infarction or chronic kidney disease. To assess the impact of the plasma levels of the CP on the incidence of HTN and HTN-related diseases, we used prospectively collected clinical data from the UK Biobank to perform multivariate Cox regression models (adjusting for age and sex). This analysis identified six proteins in the CP, such as PLXNB2, NCAN and SPON1,

that are significantly associated with the incidence of HTN, atherosclerosis, or the death (P<0.001 to P<10e-16). With our algorithm, we first used the genetically-determined expression of plasma proteins to identify and characterize a CP associated with blood pressure, and then we conducted a prospective analysis using the plasma levels of the CP to prioritize six new candidate plasma targets to treat HTN and HTN-related diseases.

Deep learning modeling of rare noncoding genetic variants in human motor neurons defines CCDC146 as a therapeutic target for ALS

Authors: J. Cooper-Knock[1], S. Zhang[2], T. Moll[3], J. Rubin-Sigler[4], S. Tu[4], S. Li[5], E. Yuan[6], M. Liu[2], A. Butt[1], C. Harvey[3], S. Gornall[1], E. Alhathli[7], A. Shaw[1], C. dos Santos Souza[1], L. Ferraiuolo[1], E. Hornstein[8], T. Shelkovnikova[1], C. van Dijk[9], I. Timpanaro[9], K. Kenna[10], J. Zeng[5], P. Tsao[11], P. Shaw[3], J. Ichida[4], M. Snyder[12]; [1]Univ. Of Sheffield, Sheffield, United Kingdom, [2]Univ. of Florida, Gainesville, FL, [3]Univ. of Sheffield, Sheffield, United Kingdom, [4]Univ. of Southern California, Los Angeles, CA, [5]Westlake Univ., Hangzhou, China, [6]Tsinghua Univ., Beijing, China, [7]The Univ. of Sheffield, Sheffield, United Kingdom, [8]Weizmann Inst. of Sci., Rehovot, Israel, [9]UMC Utrecht, Utrecht, Netherlands, [10]Univ. Med. Ctr. Utrecht, Utrecht, Netherlands, [11]VAPAHCS/Stanford Univ., Palo Alto, CA, [12]Stanford Dept. of Genetics, Palo Alto, CA

Abstract:

Amyotrophic lateral sclerosis (ALS) is a fatal and incurable neurodegenerative disease caused by the selective and progressive death of motor neurons (MNs). Understanding the genetic and molecular factors influencing ALS survival is crucial for disease management and therapeutics. In this study, we introduce a deep learning-powered genetic analysis framework to systematically uncover the genetic basis of ALS survival. Using functional genomics data from human induced pluripotent stem cell (iPSC)-derived MNs, this method prioritizes functional noncoding variants using deep learning, links cis-regulatory elements (CREs) to target genes using epigenomics data, and integrates these data with genetic association analysis to identify survival-modifying variants, CREs, and genes. We apply this approach to analyze 6,715 ALS genomes, and pinpoint four novel rare noncoding variants associated with survival, including chr7:76,009,472:C>T linked to CCDC146. CRISPR-Cas9 editing of this variant increases CCDC146 expression in iPSC-derived MNs and exacerbates ALS-specific phenotypes, including TDP-43 mislocalization. Suppressing CCDC146 with an antisense oligonucleotide (ASO), showing no toxicity, completely rescues ALS-associated survival defects in neurons derived from sporadic ALS patients and from carriers of the ALS-associated G4C2-repeat expansion within C9ORF72. ASO

targeting of CCDC146 may be a broadly effective therapeutic approach for ALS. Our framework provides a generic and powerful approach for studying noncoding genetics of complex human diseases.

Session 27: Interrogating Variant Function at Scale

Location: Four Seasons Ballroom 2&3

Session Time: Wednesday, November 6, 2024, 1:15 pm - 2:15 pm

High-throughput Deep Mutational Scanning to determine pathogenicity of Variants of Uncertain Significance in genes in the Sonic Hedgehog Pathway

Authors: D. Baldridge, J. Flury, M. Grupe, A. Sokolic, E. Orr, C. Chitwood, J. Shepherdson, B. Cohen; Washington Univ. in St. Louis, Saint Louis, MO

Abstract:

One major limitation for the molecular diagnosis of individuals with suspected monogenic disorders is the abundance of Variants of Uncertain Significance (VUS) due to insufficient evidence for determining if variants are pathogenic or benign. More than 1.5 million VUS are now present in the ClinVar database. High-throughput functional assessment of variant effects, such as Deep Mutational Scanning (DMS), offers a scalable solution to this critical problem in genomic medicine. Our aim is to demonstrate the clinical value of DMS to simultaneously experimentally assess thousands of variants in genes of interest, using reliable, carefully calibrated, cell-based assays.
We have successfully completed a parallelized mutational scan for ~3,000 missense variants, including ~200 clinically observed variants (i.e., all variants present in ClinVar at the start of the project) in the gene, *GLI2*, part of the sonic hedgehog signaling pathway. We have achieved highly reproducible and near perfect discrimination of known pathogenic and benign variants in this gene, pathogenic variants in which cause Culler-Jones syndrome, involving endocrine and skeletal patterning abnormalities. Using cells engineered with a synthetic GFP reporter and transduced with lentivirus expressing *GLI2* variants, we conducted a SortSeq experiment, coupling fluorescence-activated cell sorting (FACS) to sequencing, to demonstrate the functional effects of these variants. The *GLI2* assay performed better than any previously published cell-based assay, with an OddsPath of 38.8, allowing these results to be used in clinical variant reclassification with a strength of “Very Strong.” In addition to enabling the potential reclassification of ~100 clinically observed missense VUS (plus pre-emptive functional evaluation of missense variants that will be observed in the future), we also identified previously mis-classified variants, emphasizing limitations of the current “gold standard” ACMG variant classification approach. Further, we provide evidence to support the expansion of this approach to include at least three other clinically relevant genes in the sonic hedgehog pathway, including *PTCH1*, *SUFU*, and *SMO*. Pathogenic variants

in *SMO* are associated with Pallister-Hall-Like (PHLS) and Curry-Jones (CJS) syndromes, caused by loss-of-function and gain-of-function variants, respectively; our assay reliably detected both increased and decreased pathogenic activity for this gene.
Our results demonstrate the feasibility for a pathway-based approach for mutational scanning to study many genes in the genome, providing a potentially scalable solution to the VUS problem.

PerturbVI: A scalable latent factor model to infer regulatory modules from large-scale CRISPR perturbation data

Authors: N. Mancuso[1], D. Yuan[1], H. Pimentel[2]; [1]Univ. of Southern California, Los Angeles, CA, [2]Univ. of California Los Angeles, Los Angeles, CA

Abstract:

Large-scale CRISPR screens (e.g., Perturb-seq) provide a promising avenue towards inferring gene regulation through direct interventions, i.e., CRISPRg. However, due to the sheer scale of these data, their analysis is challenging, and standard differential expression testing is underpowered due to the massive number of tests required. While matrix-decomposition-based approaches can mitigate the testing burden, these methods either fail to model the target information explicitly or cannot scale computationally beyond a few dozen targets with genome-wide measurements, precluding their application to large-scale data.Here, we introduce PerturbVI, a target-informed, ultra-scalable latent factor model designed to infer regulatory modules from large-scale experimental perturbation data: First, PerturbVI models sparse latent factors for each cell as a function of the perturbation design matrix; Second, PerturbVI leverages our previous work SuSiE-PCA to model ultra-sparse loadings that represent downstream gene effects; Lastly, PerturbVI uses the variational approach implemented in JAX for ultra-fast inference. In simulations, we demonstrate that PerturbVI outperforms existing methods in identifying regulatory modules while exhibiting dramatically reduced runtimes (~60x) and memory usage (~90% less).We analyzed CROP-seq data targeting 14 Autism spectrum disorder-relevant genes in 8708 human LUHMES cells with measurements across 6000 genes, and compared results with recent Guide Sparse Factor Analysis (GSFA). Overall, PerturbVI identified more enriched pathways relevant to all targets than GSFA (1016 vs. 678; FDR < 0.05) while retaining functionally relevant neuron-related pathways. Further, PerturbVI identified a new target gene, POGZ, which suppresses the expression level of genes that negatively regulate neuron differentiation (e.g. ITM2C, DRAXIN).Next, we analyzed the genome-wide Perturb-seq dataset targeting 2057 perturbations sequenced in 310,385 K562 cells across 8563

genes. Overall, PerturbVI identified numerous cellular process-related pathways and corresponding perturbation-gene pairs in the cell cycle, ribosome biogenesis, etc., consistent with previous findings. Notably, we identified that the repressing of PHB2, GSPT1, and CHMP3 inhibits the cell cycle transition; and the repressing of the snRNP complex disrupts rRNA processing but upregulates EIFs and RPLs involved in ribosome assembly.Overall, PerturbVI is an accurate method for inferring gene regulatory modules from large-scale CRISPR perturbation data, providing robust insights and highly improved efficiency than comparable methods.

Changes in Kv11.1 (*hERG*/*KCNH2*) protein interactomes from hiPSC-derived cardiomyocytes of individuals with extreme QT interval polygenic scores and CRISPR edited rare variants

Authors: S. Woo[1], M. Ku[1], C. Egly[1], L. Barny[2,3], L. Vanags[1], D. Mitchell[1], N. Patel[1], A. Shen[1], B. Knollmann[1], L. Plate[2], B. Kroncke[1]; [1]Vanderbilt Univ. Med. Ctr., Nashville, TN, [2]Vanderbilt Univ. Dept.s of Chemistry and Biological Sci., Nashville, TN, [3]Vanderbilt Univ. Sch. of Med., Nashville, TN

Abstract:

Long QT Syndrome (LQTS) is characterized by an abnormally long QT interval on an electrocardiogram with approximately 30% of cases coming from rare variants in the cardiac repolarization potassium channel K_v11.1, encoded by *KCNH2*. Individuals with LQTS are at a higher risk of developing life-threatening arrhythmia torsades de pointes. Incomplete penetrance of LQTS complicates making accurate diagnoses in patients as some with pathogenic variants may not consistently exhibit symptoms. Specifically, the majority of individuals with variant R148W do not exhibit LQTS, while those with variant R823W have increased susceptibility to LQTS. Variability in disease presentation could possibly be due to differences in protein interactions. Better understanding the effects of an individual's genetic background on interactive proteomics may enhance our understanding of phenotype variability in K_v11.1-mediated LQTS. Our aim is to analyze K_v11.1-protein interactions that are differentially affected by varying genetic backgrounds using polygenic scores and how these backgrounds influence the phenotypic effect of rare variants. We also seek to characterize the K_v11.1 interactome in CRISPR-edited variants, R148W and R823W, with these different backgrounds to potentially explain variability in clinical presentations. We selected patient cell lines with low (<0.5%), medium (~50%), and high (>99%) polygenic risk scores for LQTS. Their peripheral blood mononuclear cells (PBMCs) were reprogrammed into human-induced pluripotent stem cells (hiPSCs),

followed by cardiac induction for 30 days. To isolate K_v11.1 and its interactors, co-immunoprecipitation (Co-IP) was conducted on cardiomyocyte samples using anti-K_v11.1 antibody beads. Mass spectroscopy was performed on 10 replicates of isolated protein samples to collect quantitative and qualitative data on channel protein complexes to elucidate the identity of proteins, relative abundance of proteins, and protein interaction dynamics. We identified 3824 protein interactors along K_v11.1 in at least one of five samples using mass spectroscopy, including known interacting proteins encoded by *CANX*, *SYNPO2*, *CAV3*, and *CAVIN1*. When K_v11.1 protein interactions were studied between loss of trafficking pathogenic phenotypes and wildtype, we discovered consistent variability in the abundance of interactive proteins among different genotypes linked to LQTS. By investigating K_v11.1 protein interactions that lead to LQTS, we can provide more reliable knowledge on the fundamental mechanisms that affect K_v11.1-related deficiencies, informing future endeavors to diagnose and treat such diseases.

A shared autophagy pathway dysregulated in multiple neurodegenerative diseases revealed by phenotypic CRISPR screens of iPSC-derived neurons with familial mutations

Authors: L. Evans[1], E. Tuck[1], S. Horswell[1], C. Dominicus[1], S. Tamaddon-Jahromi[1], P. Zalmas[1], C. Kayatekin[2], A. Bassett[3], G. Trynka[1]; [1]Wellcome Sanger Inst., Hinxton, United Kingdom, [2]Sanofi, Worcester, MA, [3]Wellcome Sanger Inst., Cambridge, United Kingdom

Abstract:

GWAS of Alzheimer's (AD) and Parkinson's disease (PD) have identified autophagy dysfunction as a key pathogenic feature of disease. Dysfunction of autophagy pathways is thought to contribute to the deposition of protein aggregates associated with neurodegeneration. Thus, modulating autophagy to clear protein aggregates is emerging as a priority for disease treatment.

To identify genetic modulators of autophagy dysfunction relevant to neurodegenerative disease, we performed genome-wide CRISPR FACS-based screens in human iPSC-derived neurons with familial mutations for neurodegenerative conditions (AD, PD and lysosomal storage disorders). These disease models recapitulate neurodegenerative autophagy and lysosomal dysfunction. In 60 parallel screens, we phenotypically evaluated billions of neurons; this has enabled us to identify genetic modifiers of autophagy and lysosomal function specific to disease-relevant genetic backgrounds and isogenic control neurons. We developed a novel analytical framework: Phenotypes Integrated in Relation to Weighted Outcomes of Recognised Knockouts (PhIReWORK) that combines CRISPR

effects across multiple phenotypic readouts for a single unified analysis. PhIReWORK assessment of control neurons identified a core set of 199 genes ($p<0.001$) that modulate autophagy function, forming extensive networks of protein interactions (STRING: 278 edges; $p=0.009$). Central to the largest network is ubiquitin fusion protein S27a, encoded by *RPS27A* (perturbation increased lysosomal signal). Patients with late onset AD express reduced levels of *RPS27A*, presenting links between autophagy and neurodegeneration. Building on these findings, we established subsets of genes that modify autophagy and lysosomal phenotypes in monogenic disease backgrounds associated with AD and PD, when contrasted against isogenic control. Including Amyotrophic Lateral Sclerosis (ALS) associated gene: *SIGMAR1*, in which perturbation aberrantly reduces autophagic compartment signal ($p=0.004$ and $p=0.0004$, respectively; ns in control neurons). These data provide evidence of a common neurodegenerative pathway involving autophagosome formation. Performing screens in familial mutation backgrounds has allowed us to identify specific key modifiers that enhance or suppress disease phenotypes, providing networks of genes that relate to the causal mutations.

Session 28: Liver, Laugh, Love: New Insights into Liver Disease

Location: Room 405

Session Time: Wednesday, November 6, 2024, 1:15 pm - 2:15 pm

Genetic Determinants of Liver Function Markers in African Ancestry Populations

Authors: R. Mogire[1], C. Rotimi[2], A. Adeyemo[3]; [1]NHGRI, Bethesda, Maryland, MD, [2]NIH, Woodstock, MD, [3]NIH, Bethesda, MD

Abstract:

Background: Genetic variation influences liver function enzyme levels as well as the severity and progression of liver diseases. However, the genetic determinants of liver enzyme levels in African ancestry populations remain poorly characterized. Objectives: This study aims to describe the genetic architecture of markers of liver function in populations of African ancestry. Methods: We analyzed data from over 7,200 individuals encompassing African adults from Nigeria, Ghana, and Kenya enrolled in the Africa America Diabetes Mellitus (AADM) Study and African American adults enrolled in the Howard University Family Study (HUFS). Genome wide association (GWAS) analyses were conducted using a linear mixed model in GCTA, adjusting for age, sex, alcohol consumption, type 2 diabetes status and principal components of the genotypes. Results: Our GWAS analysis identified multiple loci that are significantly associated with the liver function biomarkers, including nine loci for albumin, nineteen for total protein, six for alanine aminotransferase, and eight for alkaline phosphatase, with several novel discoveries (including *RHEX*, *MYO1E,* and *PPP4R2*). Many of these loci are located in genes integral to liver physiology, including *RHEX,* a gene involved with hemoglobinization and erythroid cell expansion; *MYO1E* which plays a role in intracellular transport; *LYPD3,* involved in cell migration; and *PTPRD* a gene with functions in cell growth and differentiation. Functional annotation and replication in populations of African ancestry in the *All of Us* Research Programme and dbGaP is currently ongoing. Conclusion: The strong associations identified highlights the significant genetic effects on liver enzyme levels. Additionally, the discovery of novel loci offers new insights into the complex genetic architecture of liver function in African ancestry populations.

Investigating the Role of LYPLAL1 Loss-of-Function in Metabolic Dysfunction-Associated Steatotic Liver Disease

Authors: R. Hussain, C. Raut, P. Ponnandy, A. Irshaid, B. Halligan, E. Speliotes; Univ. of Michigan, Ann Arbor, MI

Abstract:

Introduction: Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD) is becoming the most common liver disorder worldwide. The etiology of this disease is complex and strongly influenced by genetic factors. Human genetic studies have identified single-nucleotide polymorphisms (SNPs) near *LYPLAL1* that associate with distribution of subcutaneous and visceral adipose tissue in humans. The precise biological mechanisms through which LYPLAL1 causes MASLD remain unclear. To better understand how LYPLAL1 might influence MASLD, we analyzed rare loss-of-function LYPLAL1 mutations using UK BioBank (UKBB) Whole Exome Sequencing (WES) data and knocked out LYPLAL1 in human liver cells.

Methods: We used CRISPR-Cas9 to knock out LYPLAL1 in a human liver cell line (HuH-7). Intracellular lipid droplet levels were quantified using fluorescent imaging after staining with a neutral lipid stain. We measured triglyceride (TG) and cholesterol accumulation with biochemical assays. We assessed mitochondrial oxidative stress by measuring reactive oxygen species (ROS) levels and determined the expression levels of major beta-oxidation enzymes using antibody staining and FACS. From WES data we identified 100 LYPLAL1 variants that were missense or stop-gain variants in the coding region, had an alphamissense or gpn-msa score in the top 80% of LYPLAL1 variants, and had an allele frequency below 5e-5. We carried out burden testing in European males and females for effects on inverse normally transformed UKBB MRI Proton Density Fat Fraction (PDFF) (N=30,050 males and 32,135 females) and UKBB ICD-based NAFLD (N=172,664 males and 208,795 females) with Regenie. We meta-analyzed results using fixed effects meta-analysis in RAREMETAL using sample size and direction of effect.

Results: Rare loss-of-function LYPLAL1 variants were protective against fatty liver in European individuals ($Z = -2.071$, $p = 0.03837$) across PDFF and ICD diagnoses and sexes. LYPLAL1 KO resulted in reduced lipid content compared to wild-type (WT) HuH-7 cells when treated with 100-1000 μM oleic acid (all $p < 0.001$). LYPLAL1 KO cells showed a trending decrease in TG levels but no significant difference in intracellular cholesterol compared to controls. Expression of major beta-oxidation enzymes ACADVL, ACADM, and HADHA was higher in LYPLAL1 KO cells ($p < 0.001$). There were no differences in reactive oxygen species (ROS) levels between WT and LYPLAL1 KO cells.

Conclusion: Our data suggest that loss of LYPLAL1 function protects against MASLD. Loss

of LYPLAL1 led to increased levels of beta-oxidation enzymes following LYPLAL1 reduction which may contribute to the lower triglyceride levels we observed.

Characterizing 99 candidate genes for a role in MASLD and MASH using CRISPR/Cas9, *in vivo* imaging and deep learning in zebrafish larvae

Authors: E. Mujica[1], A. Emmanouilidou[1], H. Zhang[1], E. Mazzaferro[1], C. Metzendorf[1], M. Bandaru[1], N. Cook[1], J. Costa[1], G. Alavioon[1], B. Andersen[2], S. Vienberg[2], J. Flannick[3], A. Larsson[1], A. Allalou[1], M. den Hoed[1]; [1]Uppsala Univ., Uppsala, Sweden, [2]Novo Nordisk A/S, Måløv, Denmark, [3]Boston Children s Hosp., Boston, MA

Abstract:

INTRODUCTION: Metabolic dysfunction-associated steatotic liver disease (MASLD) is characterized by lipid accumulation in the liver that can progress to steatohepatitis (MASH), cirrhosis and hepatocellular carcinoma. GWAS have identified at least 29 loci that are associated with MASLD and several clinical trials are exploring therapeutic agents using these genetically validated targets. Still, only one drug has been FDA-approved for the treatment of MASH. Here, we validate image-based models in zebrafish larvae for MASLD, and characterize the role of 99 cardiometabolic candidate genes on liver fat. **METHODS:** Liver fat was visualized in 10-day-old transparent zebrafish larvae using a dye and live fluorescence microscopy, followed by image analysis using deep learning-based neural networks. For validation, we compared liver fat in larvae with/without a metabolic challenge (n=721), treatment with rosiglitazone (n=865), or CRISPR/Cas9-induced mutations in proof-of-principle genes as *marc1* (n=383) or *gpam* (n=307). Additionally, larvae with fluorescently labelled hepatic stellate cells (HSC) (Tg[hand2:EGFP]) were imaged with 3% ethanol and/or 4% extra dietary cholesterol (n=54). Finally, we jointly targeted all zebrafish orthologues of 99 human cardiometabolic candidate genes (one-by-one) using CRISPR/Cas9 (n=16,141) and examined effects on liver fat. **RESULTS:** On average, 4% extra dietary cholesterol results in 1.4-fold more liver fat (0.61±0.10 SD units); 3% glucose in 2.2-fold more liver fat (1.35±0.10); treatment with rosiglitazone in 40% less liver fat (-0.25±0.09); and CRISPR/Cas9-induced mutations in *marc1* and *gpam* in 6% (-0.28±0.14) and 9.5% less (-0.56±0.24) liver fat. These findings are directionally consistent with effects of human loss-of-function (LoF) variants and are in line with these genes being the culprits in loci identified by GWAS for MASLD. A challenge with 3% ethanol and 4% extra cholesterol results in more HSC activation (1.41±0.50). Fifteen of 99 genes affect liver fat when perturbed in zebrafish larvae. For 12 of these genes, LoF variants in human and/or mice have previously implicated them in liver fat but only 3

had been identified through GWAS for MASLD traits. Ten genes (83%) show directionally consistent effects across species. Genes affecting liver fat in zebrafish larvae are enriched for common variant associations with type-2 diabetes in humans. Lastly, we identify 3 novel genes not previously implicated in liver fat accumulation. **CONCLUSION:** Systematically characterizing cardiometabolic candidate genes for a role in MASLD using zebrafish larvae can meaningfully pinpoint putative causal genes.

Machine learning-based subtyping and validation with longitudinal patient data in metabolic dysfunction-associated steatotic liver disease

Authors: S. Tian[1], T. Sultana Priya1[1], H. Yan[1], D. Yao[2], E. Klee[1]; [1]Mayo Clinic, Rochester, MN, [2]Virginia Tech, Blackburg, VA

Abstract:

Background: MASLD is a prevalent liver disease affecting more than 25% of the global population. MASLD poses challenges for pharmacotherapy due to the broad variability in clinical manifestations and natural history. Disease subtyping might offer an effective approach for tailored care and for predicting patients' long-term outcomes. **Methods**: Two patient cohorts (A and B) were selected from two population genomics studies, Mayo Clinic Project Generation and Tapestry Study. Cohort A (5,176 cases, 11,080 controls) was split into a development/validation set (75%:25%), and cohort B (3,534 cases, 8,158 controls) was used as an independent validation cohort. The clinical variables and 10-year longitudinal data were extracted from the EHRs, and polygenic risk score was calculated to measure genetic disposition to MASLD. Sixteen significant ($p<0.001$) indicator variables from the development set were selected for clustering using stepwise forward and backward regression approach. Latent Class Analysis (LCA) was used to construct MASLD subgroups in the development set. Three approaches, distance-, density-, and probability-based models, were implemented for assigning new patients to the above subgroups. To showcase the utility of our subtyping strategy, logistic regression was used to analyze 10-year follow-up data for disease progression among subgroups. **Results:** We identified five latent subgroups, i.e., C1 (Non-Obese Metabolic NAFLD, 33%), C2 (Male Dominant Cardiorenal NAFLD, 23.5%), C3 (Metabolic-Multi-Morbid NAFLD with Psychoneurological Burden, 24%), C4 (Non-Metabolic NAFLD, 13%) and C5 (Elevated NAFLD with High Genetic Risk, 6.5%). Clinical variables and PRSs were used to characterize the signatures of the five subgroups. The top-performing patient assignment approach is density-based for validation set, with 97% accuracy, and distance-based for independent Cohort B (90%

accuracy). Over the 10-year follow-up, C4 showed increasing BMI (from 26 to 30) and potentially elevating hepatocellular carcinoma risk. C5, with a high genetic risk, displayed elevated liver tests and increased risk of chronic kidney disease and fibrosis. **Conclusions**: The analysis revealed 5 latent subgroups with distinct clinical and genetic signatures. Three new models were implemented for assigning patients to pre-built groups, in particular the density-based approach that exhibits the highest prediction accuracy. The analysis of longitudinal data for these groups demonstrates the clinical value of disease subtyping in MASLD.

Session 29: Modeling Rare Neurodevelopmental Disorders in Human iPSCs and Mice

Location: Room 505

Session Time: Wednesday, November 6, 2024, 1:15 pm - 2:15 pm

Rapid generation of mouse model mimicking VUS uncovers novel pleiotropy in neurodevelopmental disorders

Authors: S. Hayashi[1], Y. Suzuki[1], D. Fukushi[1], K. Yamada[1], N. Nakamura[2], S. Otsuji[2], T. Uehara[2], M. Inaba[2], S. Mizuno[2], H. Miyahara[3]; [1]Inst. for Dev.al Res., Aichi Dev.al Disability Ctr., Kasugai, Japan, [2]Central Hosp., Aichi Dev.al Disability Ctr., Kasugai, Japan, [3]Dept. of Neuropathology, Inst. for Med. Sci. of Aging, Aichi Med. Univ., Nagakute, Japan

Abstract:

Background: Genetic pleiotropy is a condition in which a single gene or genetic variant influences two or more phenotypes. Although the concept of pleiotropy was proposed over a century ago, our understanding of genuine pleiotropy in neurodevelopmental disorders (NDD) remains limited. In Recent years, genetic analysis for NDD has identified a large number of both pathogenic variants and variants of uncertain significance (VUS). When VUS is identified in a gene that has already been established to cause a phenotype different from the patient's, it is challenging to determine the pathogenicity of the VUS. **Methods:** To address this issue, we have attempted to rapidly generate a mouse model that precisely mimics VUS using a novel genome editing method, *i*-GONAD. The method directly induces genome-editing reagents into fertilized ova to produce transgenic mice only in 19 days. The mice model can be used for both in vitro analyses, such as protein analysis, and in vivo analyses, such as behavior analysis. This provides robust evidences of genotype-phenotype correlation. **Result:** To date, we have successfully generated mouse models of 16 VUSs of 12 genes and analyzed their phenotypes. For instance, we generated a mouse reproducing a VUS at exon-intron junction of *TENM4* resulting in in-frame exon skipping. This variant was identified in a familial case of intellectual disability (ID) and epilepsy. Although missense variants of *TENM4* had previously been reported as a cause of essential tremor or schizophrenia, the phenotype of our case significantly differed from those observed in the previous studies. The mRNA and protein extracted from the mice's brains replicated the in-frame exon skipping, and the significant increase in susceptibility to epilepsy induced by pentylenetetrazol demonstrated the exon skipping caused epilepsy. Additionally, the mice exhibited thinner corpus callosum. Consequently, the exon skipping resulted in three notable alterations: a

protein change, a behavior change, and a brain structure change. In other instances, we identified a frameshift variant of *CTNND2* in a familial case of ID and epilepsy. Previous studies had reported that missense variant and in-frame deletions caused familial cortical myoclonic tremor with epilepsy and autistic spectrum disorder, respectively. The mice model mimicking the frameshift variant exhibited significant hypoactivity which was recovered by a GABA antagonist. This may be indicative of epilepsy. **Conclusion:** These findings successfully identified the pleiotropic condition of each gene and suggest that the generation of the mouse model reproducing VUS is an effective methodology to prove genetic pleiotropy.

Loss of SZT2 leads to an increase in outer radial glia by hyperactivation of mTORC1 in human brain organoids

Authors: E. Sato[1], Y. Nakamura[1], M. Fujimoto[1], I. Shimada[2], T. Iwaki[1], D. Ieda[1], Y. Negishi[1], A. Hattori[1], Y. Kato[2], S. Saitoh[1]; [1]Dept. of Pediatrics and Neonatology, Nagoya City Univ. Graduate Sch. of Med. Sci., Nagoya, Japan, [2]Dept. of Cell Biology, Nagoya City Univ. Graduate Sch. of Med. Sci., Nagoya, Japan

Abstract:

The seizure threshold 2 (SZT2) gene is located on chromosome 1p34.2 and contains 71 exons. The gene is expressed predominantly in the parietal lobe, frontal cortex, hippocampus, cerebellum and dorsal root ganglia. In recent years, there have been an increasing number of literatures indicating that biallelic loss-of-function pathogenic variants of *SZT2* cause neurodevelopmental disorders with the main symptoms being epilepsy, developmental delay, macrocephaly, and abnormalities in the morphology of the corpus callosum. SZT2 forms the KICSTOR complex downstream of the amino acid-sensitive pathway in the mTOR cascade and represses mTORC1. Regulation of mTORC1 has been reported to play an important role in neurological processes, including neural development and circuit formation. We previously reported that loss-of-function mutations in *SZT2* caused constitutive activation of mTORC1 in patient-derived lymphoblastoid cell lines. Nevertheless, the impact of loss of SZT2 function on human brain development remains unclear. Here, we examined the effects of loss of SZT2 on brain development using brain organoids. We generated *SZT2* knockout (KO) induced pluripotent stem cell (iPSC)-derived brain organoids and compared them with control iPSC-derived brain organoids. *SZT2*-KO brain organoids showed higher mTORC1 activity than control brain organoids in the subventricular zone (SVZ), where neural progenitor cells amplify for cortical expansion in response to mTORC1 activity. Furthermore, in the SVZ, outer radial

glial (oRG) cells were significantly increased in *SZT2*-KO organoids and upper layer neurons differentiating from oRG cells also tended to increase in *SZT2*-KO brain organoids compared to control brain organoids. In conclusion, we showed that loss of SZT2 caused mTORC1 hyperactivation during brain development and increased the number of oRG cells. Our results suggest that dysregulation of mTORC1 in early neural development may underlie macrocephaly and developmental delay in SZT2-related diseases.

Investigating NuRDopathies with GATAD2B-associated Neurodevelopmental Disorder (GAND): clinical evaluations and modeling with patient-derived IPSCs and mice

Authors: T. Pierson[1,2,3], C. Abad[4], M. Otero[5], M. Robayo[6], M. del Mar Muñiz-Moreno[7], C. Freeman[5], S. Y. Zhao[5], M. T. Bernardi[8], K. Walz[4], J. Young[4]; [1]Div. of Pediatric Neurology, Dept. of Pediatrics, Guerin Children's, Cedars Sinai Med. Ctr., Los Angeles, CA, [2]Dept. of Neurology, Cedars Sinai Med. Ctr., Los Angeles, CA, [3]Board of Governors Regenerative Med. Inst., Los Angeles, CA, [4]John P. Hussman Inst. for Human Genomics, Miller Sch. of Med., Univ. of Miami, FL, USA., Miami, FL, [5]Cedars Sinai Med. Ctr., Los Angeles, CA, [6]Univ. of Miami Sch. of Med., Miami, FL, [7]4John P. Hussman Inst. for Human Genomics, Miller Sch. of Med., Univ. of Miami, FL, USA, Miami, FL, [8]John P. Hussman Inst. for Human Genomics, Miller Sch. of Med., Univ. of Miami, FL, USA, Miami, FL

Abstract:

NuRDopathies are a group of neurodevelopment disorders associated with the Nucleosome Remodeling and Deacetylase (NuRD) Complex. NuRD complex subtypes are composed of seven sets of paralogous subunits organized into a Histone Deacetylase (HDAC) core (HDAC1/2:MTA1/2/3:RBBP4/7) and a Chromatin Remodeling Subcomplex (CRS) (MBD2/3:GATAD2A/B:CHD3/4/5). To date, NuRDopathies only involve CRS subunit paralogs and possess overlapping clinical phenotypes. *GATAD2B*-associated Neurodevelopmental Disorder (GAND) acts as a "pan-NuRDopathy" because its neurodevelopment phenotype is an amalgamation of the other NuRDopathies. The GATAD2B protein is the predominant GATAD2 paralog during neurodevelopment and so links each chromatin remodeling subunit (CHD3/4/5) to NuRD during this process.
We have previously published a 50-subject GAND cohort with individuals possessing loss-of-function (LoF), splice-site, and missense variants and found a consistent phenotype with some variant-type specific clinical phenotypes (increased epilepsy with missense variants). The majority of this cohort consisted of LoF variants and so our present retrospective clinical study focuses on evaluating phenotypes of subjects possessing

splice-site and missense variants. We have also modeled GAND using mouse and patient-derived IPSC-based models. GAND-IPSCs were capable of being differentiated into neuro-progenitor cells (NPCs) and cortical neurons (CNs) in the context of 2-D cultures and cerebral organoids. GAND cells with loss-of-function variants had haploinsufficient expression of *GATAD2B* mRNA and protein, which is consistent with the Gatad2b$^{+/-}$ (Gand) mouse model. Bulk-transcriptomic analysis of GAND-IPSCs and -NPCs indicated GATAD2B played a major role in NPC gene expression, while GATAD2A was more likely playing the more dominant role in IPSCs. GAND neurons had altered temporal and quantitative expression of cortical laminar markers (TBR1, CTIP2, SATB2) . Adult GAND mice were shown to have significant cognitive and behavioral deficiencies. While Day E16.5 Gand mouse embryos had altered cortical morphology and abnormal patterning of cortical laminar markers. Single-nucleus transcriptomics of E16.5 mouse cortices had similar changes as the human IPSC and also noted that GATAD2B was more highly expressed throughout the cortex during this time that its paralog, GATAD2A.
We are currently pursuing cellular, molecular, and snTranscriptomic evaluations of IPSC-derived Dorsal and Ventral Forebrain organoids, as well as E14.5 and E18.5 embryonic Gatad2b mice to better understand how neurodevelopmental programs are altered in GAND and other NuRDopathies.

Variants in cohesin release factors define a novel class of cohesin balance disorders

Authors: P. Boone[1], K. N. W. Faour[2], R. Harripaul[3], A. Mohamed[4], G. Hallstrom[5], S. Haghshenas[6], R. Yadav[7], B. Jana[8], D. Springer[9], E. Kao[10], E. Denhoff[11], K. Mohajeri[12], J. Lemanski[13], J. Kerkhof[14], H. McConkey[15], J. Rsaza[15], M. Larson[1], W. Zsabre[1], D. Lucente[1], D. S. Westphal[16], K. M. Riedhammer[16], J. Gusella[17], E. Joyce[5], B. Sadikovic[18], K. E. Pfeifer[19], D. Tai[1], M. Talkowski[17]; [1]MGH, Boston, MA, [2]Div. of Genetics and Genomics, Boston Children's Hosp., Boston, MA, [3]Massachusetts Gen. Hosp., Cambridge, MA, [4]Section of Epigenetics, Natl. Inst. of Child Hlth.and Dev., Bethesda, MD, [5]Epigenetics Inst., Univ. of Pennsylvania Sch. of Med., Philadelphia, PA, [6]London Hlth.Sci. Ctr., London, ON, Canada, [7]MGH, HMS and Broad Inst., Boston, MA, [8]MGH, Medford, MA, [9]Murine Phenotyping Core Facility, Natl. Heart Lung and Blood Inst., Bethesda, MD, [10]Inst.al Ctr.s for Clinical and Translational Res., Boston Children’s Hosp., Boston, MA, [11]9 Inst.al Ctr.s for Clinical and Translational Res., Boston Children’s Hosp., Boston, MA, [12]Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [13]Massachusetts Gen. Hosp., Somerville, MA, [14]London Hlth.Sci. Ctr., London, ON, Canada, [15]6 Molecular Diagnostics Program and Verspeeten Clinical Genome Ctr., LHSC, London, CA, [16]Inst. of Human Genetics, MRI, TUM, Sch. of Med., Munich,

DE, [17]Massachusetts Gen. Hosp., Boston, MA, [18]LHSC, London, CA, Canada, [19]Section on Epigenetics, NICHD, Bethesda, MD

Abstract:

Cohesin orchestrates 3D genome organization and regulates gene expression via DNA loop extrusion to form topologically associating domains (TADs) and specific pairwise loops. Loss of cohesin or its positive regulators, such as the cohesin loader NIPBL, causes prominent neurodevelopmental phenotypes including Cornelia de Lange syndrome (CdLS). No directed therapy exists for these conditions, nor are their molecular signatures - presumed to be loss of loops/TADs and gene misexpression - fully established. Furthermore, mutation of *WAPL*, *PDS5A*, and *PDS5B,* which negatively regulate cohesin by removing it from DNA, is of unknown phenotypic consequence. We identified heterozygous, predicted damaging variants in these three mutationally-constrained genes (*WAPL* (n=24), *PDS5A* (n=6*)*, and *PDS5B* (n=7)) in individuals with abnormal neurodevelopment. Phenotypic comparisons of individuals with *WAPL* point mutations to those with recurrent 10q22q23 deletions containing *WAPL* and 15 other genes indicated that *WAPL* is a driver gene within this genomic disorder region. Specifically, neurodevelopmental delays, facial dysmorphisms, cardiovascular defects, and musculoskeletal anomalies occur at similar rates among *WAPL* point mutation and 10q22q23 deletion patients. Furthermore, *WAPL* is one of only two genes in 10q22q23 in which predicted damaging variants are enriched in a large exome-sequenced cohort of individuals with developmental delay and autism spectrum disorder. *Wapl*$^{+/-}$ mice had normal motor skills but exhibited deficits in learning/memory. We modeled loss of *NIPBL*, *WAPL*, and the recurrent 10q22q23 deletion or duplication containing *WAPL* via CRISPR engineering in human iPSCs and differentiated them into induced neurons. RNA sequencing identified robust molecular signatures for *NIPBL* loss, demonstrating aberrations at hundreds of neuron-expressed loci. *WAPL* and 10q22q23 deletion signatures, which show partial overlap, were milder than that of *NIPBL*, consistent with the milder phenotypic manifestation in human and mouse. Finally, we repressed *WAPL* in *NIPBL*-deficient cells using antisense oligonucleotides with the goal of restoring cohesin balance and correcting transcriptome network disturbance, of potential interest as a treatment approach for CdLS and other cohesinopathies. In summary, we discovered and defined a novel class of genetic conditions caused by cohesin release factor deficiency, illuminating the bidirectional dosage sensitivity of human cohesin as a pathogenetic mechanism and a potential therapeutic vulnerability.

Session 30: Novel Aspects of Modeling Genetic Architectures of Complex Traits

Location: Mile High Ballroom 2&3

Session Time: Wednesday, November 6, 2024, 1:15 pm - 2:15 pm

Selection, pleiotropy, and chance: why rare and common variant association studies often implicate different genes

Authors: H. Mostafavi[1], J. Spence[2], M. Ota[2], N. Milind[2], T. Gjorgjieva[2], Y. Simons[3], G. Sella[4], J. Pritchard[2]; [1]New York Univ. Sch. of Med., New York, NY, [2]Stanford Univ., Stanford, CA, [3]Univ. of Chicago, Chicago, IL, [4]Columbia Univ., New York, NY

Abstract:

Genome-wide association studies (GWAS) and rare variant burden tests both aim to identify trait-associated genetic variation but often implicate different genes. To understand why these similar methodologies nominate dissimilar genes, we developed a population genetics model identifying three key determinants of gene discovery. First, natural selection acts on multiple phenotypes simultaneously, leading to smaller allele frequencies for pleiotropic variants and lower statistical power to detect them. This affects the broadly-acting coding variation used in burden tests and the highly context-specific non-coding variation in GWAS differently. Consequently, burden tests prioritize trait-specific genes and downweight pleiotropic ones, while GWAS prioritize trait-relevant genes regardless of pleiotropy. Second, random genetic drift significantly influences variant frequencies. GWAS considers single variants making it sensitive to the role of chance, but burden tests average over many variants minimizing this effect. Finally, this averaging over multiple variants causes burden tests to prioritize long genes. We validated our model using GWAS and burden tests from the UK Biobank for 210 quantitative traits. We found that the burden association signal is strongest for genes expressed specifically in the most relevant cell types and diminishes for more broadly expressed genes. Many large-effect pleiotropic genes have weak or no burden signal due to the very low frequency of coding variants. Conversely, in GWAS, heritability is enriched at such genes and is higher for variants in cell type-specific open chromatin regions. We also characterized the role of genetic drift on these top GWAS hits and found a strong confounding effect of gene length on the observed burden signals. In summary, our results show that burden tests and GWAS reveal interesting yet distinct aspects of trait biology, with major implications for gene prioritization efforts based on association studies.

Determining the driving factors shaping genetic architecture of complex traits in recently admixed populations

Authors: M. Kim, X. Zhang; Univ. of Michigan, Ann Arbor, MI

Abstract:

The complexity of diverse genetic backgrounds in admixed populations often challenges our understanding of the genetic underpinnings of complex traits, including complex diseases that exhibit unusual high prevalence in admixed populations. A lack of proper understanding of the influence of admixture in complex disease studies can lead to consequences in biomedical applications. Here, we comprehensively investigate the effect of admixture on complex trait genetic architecture and determine the cause of changes in GWAS performance and accuracy through the lens of population genetics. Specifically, we use SLiM to simulate the evolution of complex traits under a human demographic model with admixture, capturing the difference in genetic architectures by using varying relationship between the effect sizes of causal variants and selection coefficients. We account for varying degrees of population size changes and migration rates by modeling five different admixture scenarios in humans. We then simulate phenotypes based on these genotypes and introduce environmental variance. This approach allows us to perform GWAS to assess their power and fine-mapping ability under different genetic architectures.Our findings reveal a striking difference in GWAS power associated with the relationship between genetic architecture and population history. Specifically, when there is little correlation between effect sizes and selection coefficients (eg. a neutral anthropometric trait), we observe a notable increase in GWAS power than in traits where there is moderate relationship between the trait and fitness (eg. a genetic disease). Furthermore, the GWAS power is substantially higher in populations that experienced recent bottleneck events than populations that are recently expanded in size, suggesting that rare variants play a more prominent role in explaining GWAS performance and missing heritability in disease causing traits. Surprisingly, neither changes in genetic architecture nor variation in population history affected the fine-mapping ability of GWAS, suggesting a confined precision of GWAS across different complex traits. We are currently testing our hypotheses based on simulation predictions in empirical data from diverse human populations including the All of Us database. Our approach provides insights into the relationship between demographic history, the genetic basis of complex traits, and their heritability. Our research aims to enhance the accuracy and reliability of genetic studies in diverse populations, contributing to better-informed biomedical applications and personalized medicine strategies.

Genomic and ethnolinguistic diversity in >40,000 eastern and southern Africans highlights the ongoing impact of cultural affiliation shaping genetic variation

Authors: M. Yohannes[1,2], Y. Lyu[3], L. Majara[1,2,4], T. Boltz[1,2], G. Genovese[1], D. Akena[5], L. Atwoli[6,7], B. Gelaye[1,3], R. Stroud[1,3], S. Kariuki[8], C. Newton[8,9], D. Stein[4,10], S. Teferra[11], Z. Zingela[12], E. Atkinson[1,2,13], K. Koenen[1,3], A. R. Martin[1,2], NeuroGAP-Psychosis Study Team; [1]The Broad Inst. of MIT and Harvard, Cambridge, MA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]Harvard T. H. Chan Sch. of Publ. Hlth., Boston, MA, [4]Univ. of Cape Town, Cape Town, South Africa, [5]Makerere Univ., Kampala, Uganda, [6]Moi Univ. Coll. of Hlth.Sci., Eldoret, Kenya, [7]The Aga Khan Univ., Nairobi, Kenya, [8]KEMRI-Wellcome Trust Res. Programme-Coast, Kilifi, Kenya, [9]Univ. of Oxford, Oxford, United Kingdom, [10]Univ. of Cape Town and NeuroSci. Inst., Cape Town, South Africa, [11]Addis Ababa Univ., Addis Ababa, Ethiopia, [12]Nelson Mandela Univ., Gqebera, South Africa, [13]Baylor Coll. of Med., Houston, TX

Abstract:

African ancestry populations-mostly African Americans-comprise only ~1.1% of genome-wide association studies. The Neuropsychiatric Genetics of African Populations (NeuroGAP)-Psychosis study has enrolled >43k participants from Ethiopia, Kenya, South Africa, and Uganda. The cohort includes extensive demographic surveys, including up to 5 ethnic affiliations as well as languages that they, their parents, and their grandparents speak (up to 3 per individual). Participants reported 126 ethnic groups and 130 languages, totaling 2 ethnicities and 3 unique languages per individual on average. We deployed a blended genome-exome (BGE) strategy across participants, which sequences the whole genome at ~2-3X and the exome at ~30X. We have conducted quality control and imputation of genomes separately for autosomal, PAR, and non-PAR regions of the X chromosome to facilitate analysis of sex-biased demography. To evaluate the relationships between genes, geography, ethnicity, and language, we mapped all variables to geographic coordinates, then compared their relationships. Specifically, we used autosomal and sex-based principal components analysis to place each individual onto geographical points on a sphere with geographical distance from recruitment latitude and longitude informed by genetic divergence. We inferred linguistic coordinates using phoneme inventories that quantify speech sounds, weighting multiple languages spoken. We find higher correlations between language and geography versus genetics and geography, with evidence for sex-biases in genetic-linguistic relationships in East Africa. We also find consolidation of languages in current versus previous generations, consistent with globalization as well as the urban nature of participant enrollment. Given limitations in recorded histories among

study participants, we are also investigating historical patterns of divergence by reconstructing ancestral recombination graphs using genetic data to learn about demographic histories within and among NeuroGAP sites. While NeuroGAP only scratches the surface of cultural and genetic variation across Africa and study participants are not necessarily representative of local populations from which they come from, this study is of unprecedented scale within Africa and allows us to better understand the vast and complex spectrum of genetic and ethnolinguistic diversity in eastern and southern Africa. We highlight the immense impact that individuals' cultural affiliations and ongoing linguistic changes have on genetic variation, which is critical to understand to calibrate and contextualize statistical genetics analyses optimally.

Detecting ongoing natural selection affecting allele frequencies across generations to uncover genetic variants contributing to disease susceptibilities

Authors: J-L. Chen[1], M-L. Kang[1], J-H. Loo[2], M. Lin[2], Y. Satta[3], W-Y. Ko[1]; [1]Natl. Yang Ming Chiao Tung Univ., Taipei, Taiwan, [2]Mackay Mem. Hosp., New Taipei City, Taiwan, [3]Res. Ctr. for Integrative Evolutionary Sci., SOKENDAI, Hayama, Japan

Abstract:

Genetic variants that affect a complex trait such as a common disease could also impact fitness and, consequently, are suppressed by purifying selection. Hence, genetic variance of a common disease could be largely contributed by mutations at low frequency in the population and are likely to be population specific. Here, we analyzed 509,817 whole-genome genotyped variants in 72,635 Han Taiwanese individuals to detect candidate variants experiencing ongoing selection by comparing differences in allele frequency across generations. We detected 168 variants significantly departing from the neutral expectation of allele frequencies in different age groups after controlling for the possible age-dependent genetic structure. Most candidate variants (160) appear to show decreases in allele frequency in younger generations which are consistent with the action of purifying selection, suggesting that these candidate variants could reduce fitness of their carriers and likely contribute to disease susceptibilities. Among them, 86 candidate variants (53.8%) are indeed reported previously to be associated with a wide spectrum of diseases including both early onset (e.g., Marfan syndrome, Matthew-Wood syndrome, etc.) and late onset diseases (many candidates (35) were reported to increase cancer susceptibilities). In particular, we identified a number of pathogenic variants that are in strong linkage disequilibrium (LD) in *BRCA1* and *BRCA2*, separately. Analyses of the 1487 individuals whose whole-genome sequencing data are available further revealed a strong signal of

positive selection favoring the alternative haplotype in *BRCA1*, providing evidence of Darwinian selection on the gene in the Han Taiwanese people. We also performed phenome-wide association analyses across 30 physiological, hematological and cardiometabolic measurements to further detect any possible functional consequences for each of these candidate variants. We found that a candidate variant (rs2072114) in the intron of *FADS2* appears to be associated with multiple traits (i.e., total cholesterol, triglyceride, fasting glucose level in blood, hemoglobin level, red blood cell and platelet count, and heartbeat). *FADS2* has been shown strongly linked to cardio-metabolic diseases. In addition, we also found several cancer-related pathogenic variants associated with the size of red blood cell and hemoglobin level. Our findings are expected to unveil the roles of natural selection in contributing to disease susceptibility, as well as to facilitate disease prevention and management.

Session 31: Therapies for Genetic Disorders

Location: Four Seasons Ballroom 1

Session Time: Wednesday, November 6, 2024, 1:15 pm - 2:15 pm

A drug repurposing screen identifies NSAIDs and COX1/2 enzyme inhibition as potential therapies for MAN1B1-CDG, a rare congenital disorder of glycosylation.

Authors: C. Chow[1], K. Beebe[1], K. A. Hope[1], E. Coelho[1], H. D. Evans[1], C. Massey[1], C. Fast[2], E. O. Perlstein[3]; [1]Univ. of Utah, Salt Lake City, UT, [2]Univ. of British Columbia, Vancouver, BC, Canada, [3]Perlstein Lab PBC, San Francisco, CA

Abstract:

MAN1B1-CDG is a congenital disorder of glycosylation caused by autosomal recessive mutations in the *MAN1B1* gene. MAN1B1-CDG is characterized by intellectual and developmental delay, hypotonia, truncal obesity, verbal and physical aggression, and facial dysmorphisms. Clinical symptoms common to other CDGs, like seizures are rare. MAN1B1-CDG is a rare disorder with fewer than 100 patients reported. The MAN1B1 gene encodes a protein localized to the endoplasmic reticulum and Golgi and plays several different roles in protein quality control. Like most rare disorders, MAN1B1- CDG lacks any therapeutic treatments. Most treatments are focused on the management of symptoms. Drug repurposing provides a rapid way forward for identifying drugs that might treat rare diseases. Drug repurposing involves the reuse of FDA approved drugs in new indications. In particular, there are likely approved drugs that may be applied to rare diseases. This approach bypasses the years or even decades of drug development needed to bring a new molecule to the clinic. To find new potential treatments for MAN1B1-CDG, we performed a drug repurposing screen in Drosophila for MAN1B1-CDG. We developed a rough-eye model of MAN1B1 loss of function and subjected the model to 1,520 FDA approved drugs. We found 50 drugs that provided rescue and 47 drugs that enhanced or worsened the eye phenotype. Nearly 20% of the drugs that provided rescue are non-steroidal anti-inflammatory drugs (NSAIDs). Nearly all the NSAIDs showed rescue in an independent dose curve validation experiment. Because NSAIDs primarily block COX1/2 enzyme activity, we tested whether genetic reduction of COX enzyme activity could mimic NSAID rescue. RNAi knockdown of *Drosophila* COX enzyme provided strong rescue of the MAN1B1-CDG eye model. Finally, we tested whether the strongest NSAID from the screen, Ibuprofen, could rescue nervous system dysfunction associated with loss of MAN1B1 in *Drosophila*. Ibuprofen was able to provide strong rescue of seizure behavior in a neuron-specific knockdown of MAN1B1. N=1 studies with ibuprofen were conducted in three

children living with MAN1B1-CDG. Together, this study indicates that inhibition of COX activity by NSAIDs might be a viable therapeutic approach for MAN1B1-CDG.

NAGLU co-expressed with a modified phosphotransferase has increased mannose-6-phosphorylation and shows preclinical efficacy as a treatment for mucopolysaccharidosis IIIB (Sanfilipo B Syndrome)

Authors: G. Austin[1], S. Le[2], B. Doray[2], A. Sorensen[2], J. Srnak[3], L. Liu[3], S. Kornfeld[2], P. Dickson[2]; [1]St. Louis Children's Hosp., St. Louis, MO, [2]Washington Univ. Sch. of Med., St. Louis, MO, [3]M6P Therapeutics, St. Louis, MO

Abstract:

Mucopolysaccharidosis IIIB (MPS IIIB) is an autosomal recessive lysosomal storage disorder caused by Alpha-N-acetylglucosaminidase (NAGLU) deficiency and characterized by multisystem progressive symptoms including developmental delays, seizures, gait disorders, coarse facies, hearing loss, and death. As of now, there are no approved therapies for MPS IIIB. Intravenous NAGLU without modifications as enzyme replacement therapy (ERT) did not produce significant benefit in a phase 1/2 open-label clinical trial, likely due to poor endogenous mannose-6-phosphate (M6P) levels on NAGLU. IGF2 has cross reactivity with the M6P receptor (MPR) and a hybrid NAGLU-IGF2 protein delivered intracerebroventricularly (ICV) has been evaluated preclinically and clinically in Europe. Although evidence suggests that NAGLU-IGF provides clinically significant benefits, there are concerns that the IGF2 arm will lead to many off target effects, something seen in a GAA-IGF2 hybrid protein for Pompe disease. We propose a method for increasing lysosomal delivery of NAGLU by co-expression of a modified GlcNAc-1-phosphotransferase (S1S3) with NAGLU to produce NAGLU with adequate M6P levels (NAGLU-M6P). We show that NAGLU-M6P is stable, binds more efficiently than NAGLU in vitro to the MPR, is better taken up into MPS IIIB fibroblasts in culture, and that the uptake is inhibited by M6P. Intracellular heparan sulfate (HS) glycosaminoglycans (pathologically accumulated in MPS IIIB) were reduced 77% in MPS IIIB fibroblasts treated with NAGLU-M6P, but not NAGLU. NAGLU-M6P distributed better in the brain than NAGLU or NAGLU-IGF2 when delivered ICV to MPS IIIB mice. Treatment of MPS IIIB mice with NAGLU-M6P significantly reduced HS levels in the brain and was as effective as NAGLU and NAGLU-IGF2. In addition to ERT, gene therapy via AAV9 vectors to increase NAGLU or NAGLU-M6P was investigated. MPS IIIB mice were administrated with AAV9-NAGLU with or without AAV9-S1S3 ICV. NAGLU expression in MPS IIIB mice increased 50-fold over baseline NAGLU+/- mouse (carrier) levels in brain; and was returned to carrier levels in heart in both

groups. Beta-Hexosaminidase (also pathologically accumulated in MPS IIIB) level was reduced to carrier levels in brains and hearts of the NAGLU/S1S3 co-expression group, but not in hearts treated with AAV9-NAGLU alone. These data show NAGLU-M6P is at least as efficacious as NAGLU and NAGLU-IGF2 when delivered or artificially expressed in vivo in mice while performing better than NAGLU at entering cells and having less theoretical side effects than NAGLU-IGF. Given these results, NAGLU-M6P as ERT or gene therapy shows promise as a possible treatment for MPS IIIB.

Antisense oligonucleotide therapy in an individual with KIF1A-associated neurological disorder

Authors: A. Ziegler[1], J. Carroll[1,2], J. Bain[1], T. T. Sands[1], R. J. Fee[1], D. Uher[1], C. Leduc[3], D. E. Miller[4], C. H. Kanner[1], J. Montes[1], S. Glass[5], J. Douville[5], L. Mignon[5], J. G. Gleeson[6,7], S. T. Crooke[5], W. K. Chung[1,2]; [1]Columbia Univ., New York, NY, [2]Harvard Med. Sch., Boston, MA, [3]Columbia Univ, New York, NY, [4]Univ. of Washington and Seattle Children's Hosp., Seattle, WA, [5]n-Lorem Fndn., Durham, NC, [6]Univ. of California, San Diego, La Jolla, CA, [7]Rady Children's Inst. for Genomic Med., San Diego, CA

Abstract:

Nano-rare diseases affecting less than 30 patients worldwide are not financially viable in traditional drug development programs even though these diseases collectively affect millions of people. Meeting the needs of these affected individuals requires innovative solutions to address a wide range of challenges including but not limited to pre-clinical drug development, clinical outcome measures, manufacturing, and funding. Antisense oligonucleotide (ASO) therapies are one approach to delivering targeted therapies for ultra-rare diseases. *KIF1A* associated neurological disorder (KAND) is a neurodegenerative and often lethal ultra-rare disease with a wide phenotypic spectrum that is typically associated with heterozygous de novo missense variants in *KIF1A*. Here we report the case of one patient with a severe form of KAND characterized by refractory spells of behavioral arrest and heterozygous for a p.Pro305Leu variant in *KIF1A* who was treated with intrathecal injections of an allele specific ASO designed to degrade the mRNA from the pathogenic allele by targeting a noncoding polymorphism (rs7578279) in cis with the pathogenic variant. The ASO was safe and well tolerated over a 17-month treatment period. The maximal dose was 100 mg per intrathecal injection. Most outcome measures including severity of the spells of behavioral arrest, number of falls, and quality of life improved. The EEG showed improvement from the baseline EEG and remained stable. There was little change in the 6-minute walk test distance, but qualitative changes in gait resulted in

meaningful reductions in falls and increased ambulation and independence. Cognitive performance on the DAS-II was stable while qualitatively speech fluency and complexity improved. Our results support the use of an allele specific ASO as a possible treatment for KAND. We used long-read sequencing to identify 14 additional KAND patients with different pathogenic variants potentially eligible for treatment using the same ASO. Clinical outcome measures for assessment of other KAND patients will require tailoring given the wide range of severity associated with a large allelic spectrum in *KIF1A* and the clinical course in children that includes periods of acquisition of new skills, periods of neurodegeneration, and limitations due to visual impairment associated with KAND.

Rescue of Proteus syndrome lethality in mice with prenatal miransertib treatment

Authors: S. Raji Abdul Rahiman Sirajuddeen[1], M. J. Lindhurst[2], J. Johnston[1], L. Brinster[3], L. G. Biesecker[1]; [1]Natl. Human Genome Res. Inst., NIH, Bethesda, MD, [2]Natl. Inst. of Arthritis, Musculoskeletal and Skin Diseases, Bethesda, MD, [3]Div. of Vet. Resources, Office of Res. Services, NIH, Bethesda, MD

Abstract:

Proteus syndrome (PS) is a rare disorder characterized by disproportionate, asymmetric overgrowth affecting various organs and tissues. The disorder arises from a heterozygous, mosaic c.49G>A, p.(E17K) variant in the AKT1 gene, a critical component of the phosphoinositide 3-kinase (PI3K)/AKT signaling pathway. The pathophysiology of PS is linked to the continuous AKT1 activation, resulting in reduced apoptosis and increased growth. The disease fits the Happle model, which posits that the syndrome is caused by somatic mosaicism for a variant that is 100% lethal in the non-mosaic state. Due to its low incidence of 1 case per 1-10 million, conducting therapeutic trials for PS has been challenging. To address this, our lab developed a mouse model, featuring endogenously regulated, mosaic expression of the *Akt1* c.49G>A, p.(E17K) variant. This mouse model has the capability to produce variant-positive cells with random distribution, making it a suitable representation of symptoms like those observed in human patients. Because ubiquitous expression of $Akt1^{E17K}$ resulted in embryonic lethality, treatment for the mice was administered *in utero*. Our hypothesis is that a modest therapeutic effect of a small molecule could potentially rescue the 100% embryonic lethality and serve as a screen for therapeutic potential of an agent. We started by testing miransertib, an allosteric, pan-AKT inhibitor that our group used in a Phase 0/1 study and are currently evaluating in a Phase 2 study. We first confirmed the ability of miransertib to cross the placental barrier. We then

tested for embryonic lethality rescue. Dams from timed matings between *ACTB-Cre* and *Akt1*$^{WT/flx}$ mice were dosed at specific gestational time points. A startling observation occurred during this study. Several apparently non-mosaic (near-constitutive) *Akt1*^{E17K} mice were seen to not only survive intrauterine demise but were apparently healthy until nearly one year of age. Genotyping of the surviving offspring confirmed their mutant status in nearly all tissues, demonstrating that inhibition of AKT1 signaling by miransertib extended the survival of embryos expressing *Akt1*^{E17K}. Tissues from these animals exhibited a high frequency of recombined alleles. We hypothesize that *in utero* administration of miransertib reduced the frequency or severity of vascular anomalies, effectively rescuing the murine lethality compared to those that did not receive the treatment. This observation has significant implications for the pathophysiology of the lethality of PS and embryonic rescue of (near) constitutive mutants can serve as a rapid platform for *in vivo* assessment of the effectiveness of AKT1 inhibitors.

Session 32: Unifying Multimodalities: Insights from Single Cell Analyses

Location: Four Seasons Ballroom 4

Session Time: Wednesday, November 6, 2024, 1:15 pm - 2:15 pm

Leveraging single-cell multi-omic profiling to investigate non-coding variants in Parkinson's disease ⭐

Authors: S. Menon, A. Turner, R. Corces; Gladstone Inst.s, San Francisco, CA

Abstract:

A complex interplay of genetic and environmental factors influences Parkinson's disease (PD). To understand these contributions, we have created an atlas of single-cell resolution multi-omic data from a cohort of 80 individuals with PD and 21 age-matched cognitively healthy controls. This atlas provides unprecedented insight into the mechanisms driving sporadic PD. Our downstream analyses focused on two main objectives: (i) understanding the impact of rare and common coding and noncoding genetic variants on the disease and (ii) identifying transcriptional and gene regulatory differences characteristic of PD. We generated single-nucleus multi-omic data, including chromatin accessibility and RNA-seq, from 80 PD cases and 21 controls across five brain regions that are variously involved in PD: substantia nigra, temporal gyrus, cingulate gyrus, putamen, and cerebellum. We analyzed over 3.3 million nuclei from these 505 samples, with 96% of samples contributing more than 4,000 high-quality nuclei. Additionally, we performed 30x whole genome sequencing for all 101 individuals. Our joint analysis of chromatin accessibility and gene expression data has revealed novel regulatory mechanisms in neurons, glial cells, and vascular cells. Precisely, we captured over 7,000 midbrain dopaminergic (DA) neuron nuclei, allowing us to detail gene regulatory mechanisms in diverse DA neuron subtypes. The comprehensive resolution of this atlas also enables us to explore regional differences across multiple brain cell types. Using this dataset, we mapped common quantitative trait loci (QTLs) for cell type-specific chromatin accessibility (caQTLs) and gene expression (eQTLs). We trained cell type-specific convolutional neural network models to investigate rare noncoding variants to predict the effects of over 50 million PD-relevant rare noncoding variants. We identified a subset of 40,000 high-impact variants for functional validation through massively parallel reporter assays. This approach uncovered novel putative noncoding drivers of the disease in genes already associated with PD and implicated dozens of new genes with potential clinical and therapeutic relevance. Furthermore, we leveraged transcriptomic and chromatin accessibility data from our large-scale atlas to

pinpoint cell type-specific molecular changes linked to PD. In summary, we have developed a single-nucleus multi-omic atlas from 101 individuals, identifying new genetic and molecular contributors to PD pathogenesis. This work lays a foundation for discovering novel therapeutic targets and devising effective strategies for the clinical stratification of PD patients.

Single-cell eQTL analysis in >2,000 individuals in conjunction with single-cell multiomics analysis in 271 individuals infers causal disease mechanisms

Authors: R. Oelen[1], M. W. van der Werf[1], M. V. Korshevniuk[1], J. Niewold[1], D. Kaptijn[1], C. Losert[2], sc-eQTLgen consortium, H-J. Westra[1], M. Heinig[2], L. H. Franke[1], M. Bonder[1], M. G. P. van der Wijst[1]; [1]UMCG, Groningen, Netherlands, [2]Helmholtz Inst., Munich, Germany

Abstract:

To better understand disease-associated variants, they can be linked to downstream molecular effects through quantitative trait locus (QTL) mapping. However, these effects are often cell type- and context-dependent, thus requiring single-cell data across multiple modalities to find them. Therefore, to identify how genetic variants affect both gene expression (eQTL) and co-expression (co-eQTL), we meta-analyzed single-cell data from peripheral blood mononuclear cells (PBMCs) across 11 cohorts in >2,000 donors (sc-eQTLGen consortium). This identified 13,291 fine-mapped *cis*-eQTLs in 7,310 unique genes across 6 major PBMC cell types. For up to 11% of these eQTLs, we observed that the eQTL SNP (eSNP) also affected co-expression of the eQTL gene with other genes. Interestingly, these co-eQTL SNPs were up to 5.3x more often associated with disease, indicating these effects may be most informative to understand disease. To investigate the underlying regulatory mechanism of such QTLs and pinpoint likely causal variants, we generated sc-multiomics (scATAC + scRNA) data in 271 individuals. We could then determine which SNPs overlap cis-regulatory elements (CRE, i.e. chromatin accessibility peaks that correlate with nearby gene expression levels) and which transcription factors bind to such regions. Overlaying these data layers with eQTLs and co-eQTLs from sc-eQTLGen in monocytes, we identified that 40% of the fine-mapped eQTLs overlap with accessible chromatin. For 33% of those eQTLs, the peak chromatin accessibility also correlated with the expression of the eQTL gene, suggesting that such fine-mapped eSNPs are likely causal. Here we highlight the example of rs1108577, whose regulatory mechanism can be fully explained by integrating all our data layers. This variant affects the expression level of *TMEM176A/B* (encoding a cation channel promoting antigen presentation) in myeloid cells and resides within a CRE that is correlated with *TMEM176A/B* expression. To identify

the transcription factor whose binding is affected by rs1108577, we leveraged our sc-multiomics data, co-eQTL data and TF binding motif predictions. This revealed that STAT1, ELF1 and SP3 transcription factors could bind to this CRE and that their expression levels correlate with *TMEM176A/B*. However, only the co-expression of *STAT1* and *ELF1* with *TMEM176A/B* was affected by rs1108577, with only STAT1 mapping in the vicinity (5 bp). Altogether, this suggests STAT1 binds to rs1108577, and thereby regulates *TMEM176A/B*. We envision this approach can be widely applied to systematically disentangle how genetic variants lead to disease, and thereby helps to identify new personalized drug targets.

A Multiomics Single Cell Atlas Redefining the Human Maternal-Fetal Interface by Spatial Cellular Mapping

Authors: C. Wang[1], Y. Zhou[1], Y. Wang[1], T. Guha[2], L. Zhida[1], R. Wong[2], S. England[3], L. Giudice[1], D. Stevenson[2], G. Shaw[2], M. Snyder[4], S. Fisher[1], V. Winn[2], J. Li[1]; [1]UCSF, San Francisco, CA, [2]Stanford Univ., Palo Alto, CA, [3]Washington Univ. in St. Louis, St. Louis, MO, [4]Stanford Dept. of Genetics, Palo Alto, CA

Abstract:

The placenta is a transitioning fetal organ that undergoes rapid development with a compressed lifespan, playing a pivotal role in influencing pregnancy outcomes and programming the health of the offspring. Despite its significance, molecular investigations into these complications have faced significant challenges due to the intricate cellular heterogeneity of the placenta, especially at the maternal-fetal interface where maternal and fetal cells intermingle. In our study, we captured paired single-nucleus epigenomes (ATAC-seq) and transcriptomes (RNA-seq) profiles for approximately 200,000 cells at the human maternal-fetal interface, spanning from early pregnancy to term. Our data revealed cell-type-specific transcriptional regulatory programs that drive the lineage differentiation of placental cytotrophoblasts, and fine-mapped the developmental trajectories of cytotrophoblasts towards multiple terminal states. By integrating spatial single-cell proteomics profiling, we localized these cell types *in situ* and characterized the dynamic stages of endothelial cells in maternal spiral arteries remodeled by these specialized extravillous cytotrophoblasts. From the single cell multiomics data across gestation ages, we were able to recapitulate the decidualization process of decidual stromal cells with their signature molecular profiles and functions of both known and novel cell subtypes. Our integrative analyses with large-scale population genomes identified the most vulnerable maternal and fetal cell types to pregnancy complications such as

preeclampsia, preterm birth, and miscarriage. In summary, our study presents the most comprehensive placental and decidual single-cell resource across gestation to date, offering new insights into the drivers of normal human placentation and uncovering the cellular basis of dysfunction associated with common pregnancy complications.

Identifying Noncoding Regulatory Variants by Multiome Single-Cell Sequencing in Prostate Cells

Authors: Y. Tian[1], L. Wu[2], C-C. Huang[3], L. Wang[1]; [1]Moffitt Cancer Ctr., Tampa, FL, [2]Univ. of Hawaii Cancer Ctr., Honolulu, HI, [3]UW Milwaukee, Milwaukee, WI

Abstract:

While genome-wide association studies and expression quantitative trait loci (eQTL) analysis have made significant progress in identifying noncoding variants associated with prostate cancer risk and bulk tissue transcriptome changes, the regulatory effect of these genetic elements on gene expression remains largely unknown. Recent developments in single-cell sequencing have made it possible to perform ATAC-seq and RNA-seq profiling simultaneously to capture functional associations between chromatin accessibility peak and gene expression. We hypothesize that this single-cell approach allows for mapping regulatory elements and their target genes at prostate cancer risk loci. In this study, we applied a 10X Multiome ATAC + Gene Expression platform to encapsulate Tn5 transposase-tagged nuclei from multiple prostate cell lines, including a total of 65,501 high quality single cells from RWPE1, RWPE2, PrEC, BPH1, DU145, PC3, 22Rv1 and LNCaP cell lines. To address data sparsity commonly seen in the single-cell sequencing, we performed targeted sequencing to enrich sequencing data at 273 prostate cancer risk loci and 2,730 associated genes. Although not increasing the number of captured cells, the enriched multiome data did improve eQTL gene expression abundance by about 20% and chromatin accessibility abundance by about 5%. Based on this multiomic profiling, we further developed an analytical method to associate RNA expression alterations with chromatin accessibility of germline variants at single cell levels. Cross validation analysis with GTEx prostate cohort showed high overlaps between the significant multiome associations (p-value ≤ 0.05, gene abundance percentage ≥ 5%) and the bulk eQTL findings. We found that about 20% of GTEx eQTLs were covered within the significant multiome associations, and roughly 10% of the multiome associations could be identified by significant GTEx eQTLs. We also analyzed accessible regions with available heterozygous SNP reads and observed more frequent association at genomic regions with allelically accessible variants (p = 0.0055). Among these findings were previously reported regulatory variants including

rs60464856-*RUVBL1 (*multiome *p*-value in BPH1: 0.0099*)* and rs7247241-*SPINT2 (*multiome *p*-value in 22Rv1 methanol: 0.0002; in 22Rv1 DHT: 0.0004*)*. We also functionally validated a new regulatory SNP and its target gene rs2474694-*VPS53 (*multiome *p*-value in BPH1: 0.00956; in DU145: 0.00625) by reporter assay and SILAC proteomics sequencing. Taken together, our data demonstrated the feasibility of the multiome single-cell approach for identifying regulatory SNPs and their regulated genes.

Session 43: All about Implementation

Location: Room 405

Session Time: Thursday, November 7, 2024, 10:15 am - 11:45 am

Implementation of hereditary cancer risk assessment in primary care settings: Strategies and proximal outcomes

Authors: A. Gurram, S. Lahiri, Y. Ma, E. Villa, M. Bowen, S. Pirzadeh-Miller, S. Leach, S. Makhnoon; UT Southwestern Med. Ctr., Dallas, TX

Abstract:

Background: Most individuals at greatest risk of hereditary cancers are unaware of their elevated risk status. Adherence to hereditary cancer screening guidelines is suboptimal; therefore, systematic implementation of guidelines is vital to cancer prevention and clinical management. We used intervention mapping for adaptation (IM-Adapt) to guide the steps and tasks for modifying and implementing an evidence-based program to screen for hereditary cancer risk. To inform the scale-up of this program at other primary care settings and identify implementation strategies, we conducted qualitative interviews with primary care stakeholders. **Methods:** The 7-question Family History Survey (FHS-7) was adapted for electronic patient portal administration prior to, or during the primary care provider (PCP) visit. Between February 2023 and March 2024, patients at a primary care clinic at an academic medical center were evaluated for hereditary cancer risk using the adapted FHS-7. Patients who screened positive were offered follow-up genetic counseling and testing. Eleven primary care stakeholders were interviewed to identify implementation strategies and barriers to guide future rollout. We used rapid qualitative methods and content analysis to integrate data and characterize recommendations. **Results:** Of the 4,540 patients offered screening, 3,497 (77%) responded to all questions and 1,267 screened patients (36.2%) screened positive for elevated hereditary cancer risk. 1,117 high-risk patients (88.2%) were eligible for follow-up via navigation. Those not eligible either did not complete the PCP visit (n=13) or previously underwent genetic counseling within 5 years (n=137). Most patients responded via the patient portal (84.2%) before the visit. In qualitative interviews, stakeholders expressed interest in integrating hereditary cancer screenings within their patient portal system but identified known and novel barriers including lack of insurance coverage, low priority for hereditary cancer, and concern of low patient compliance with survey completion. **Conclusion:** The implementation strategies and the resulting proximal outcomes may be incorporated into future implementation design efforts for maximal impact. Strategies that facilitated cancer risk assessment

include patient-driven report of family history data, data collection prior to PCP visits, and navigation to genetic counseling. These strategies can address certain stakeholder-reported barriers, indicating their usefulness and sustainability for routine clinical practice in other primary care settings.

The Million Veteran Program Return Of Actionable Results (MVP-ROAR) Study: Preliminary outcomes from participants receiving clinical genetic confirmation testing for familial hypercholesterolemia

Authors: M. Danowski[1], H. Leishman[1], C. Brunette[1,2], T. Yi[1], J. Vassy[1,2], for the Million Veteran Program; [1]VA Boston Hlth.care System, Boston, MA, [2]Harvard Med. Sch., Boston, MA

Abstract:

Background: The Million Veteran Program (MVP) is a mega-biobank linked to a national healthcare system. To determine the feasibility and outcomes of returning medically actionable genetic results to MVP participants, the program launched the MVP Return Of Actionable Results (MVP-ROAR) Study, with familial hypercholesterolemia (FH) as an exemplar actionable condition.
Methods: MVP participants with a possible variant associated with familial hypercholesterolemia (FH) in their research genotype data are recontacted and invited to receive confirmatory clinical genetic testing. The MVP-ROAR Study includes a nonrandomized pilot trial and randomized controlled trial (RCT) comparing immediate vs. delayed disclosure of results. Enrollees complete baseline surveys and biospecimen collection for low-density lipoprotein cholesterol (LDL-C) testing and confirmatory clinical gene panel sequencing. A genetic counselor (GC) discloses the results and provides FH-specific counseling. After 6 months, participants complete end-of-study surveys and repeat LDL-C testing. We have described the pilot outcomes (PMID 38762090). We now report the 6-month outcomes of participants randomized to immediate disclosure of clinical genetic results who have completed end-of-study data collection to date.
Results: Data from 49 RCT participants in 23 states were analyzed. Mean age was 66 years (range 36-89). Ten of 49 (20%) were women; 30 (61%), 12 (24%), and 2 (4%), 1 (2%) reported White, Black, Asian and Multiracial/Other race, respectively; and 3 (6%) reported Hispanic/Latino ethnicity. Clinical sequencing confirmed the MVP research result in 45 (92%) participants; all negative results occurred prior to improved rare variant calling. At baseline, mean LDL-C was 108.2 (SD 53.0) mg/dL; 25 (51%) participants had LDL-C<100 mg/dL; and 9 (18%) had LDL-C<70 mg/dL. After 6 months, mean LDL-C was 95.7 (SD 50.0)

mg/dL; 27 (55%) participants had LDL-C<100 mg/dL; and 12 (24%) had LDL-C<70 mg/dL. Mean 6-month ΔLDL-C was -12.6 mg/dL (95% CI -25.7 mg/dL, 0.6 mg/dL; paired t-test p=0.06). Among the 45 enrollees with positive clinical confirmation, mean 6-month change in LDL-C was -13.3 mg/dL (95% CI -27.2 mg/dL, 0.6 mg/dL; p=0.07). Of the 49 participants, 29 (59%) reported having shared the result with at least one relative.
Conclusion: Preliminary MVP-ROAR Study outcomes suggest that clinical confirmation and return of FH-associated genetic research results might modestly lower LDL-C after 6 months and promote family sharing. Final between-arm analysis of the full RCT cohort will further characterize the clinical value of returning actionable genetic results to biobank participants.

Introducing an efficient framework to evaluate oncology and cardiology gene-disease validity leveraging clinicogenomic biobank data

Authors: S. Candille, **D. Iacoboni**, I. Thibodeau, M. Ferber, E. Cirulli, K. Schiabor Barrett; Helix, San Mateo, CA

Abstract:

Introduction: Establishing gene-disease validity (GDV) is difficult and time-consuming. It is aided by ClinGen's semi-quantitative rubric that scores evidence to arrive at a certainty level including refuted, limited, moderate, strong, or definitive GDV. Many common conditions are categorized as "limited-evidence" (LE), indicating insufficient data to firmly establish a gene-disease relationship. It is unclear whether to include LE associations in diagnostic reports, which can complicate counseling. Systematic analysis of clinicogenomic datasets offers an efficient way to study candidate GDVs. Here, we test how the UK Biobank (UKB) can help both confirm known associations and elevate or refute LE associations for common conditions encountered in oncology and cardiology genetic testing.
Methods: We used UKB data to analyze internally-curated associations. A gene-level autosomal dominant (AD) model tested the association between predicted loss-of-function variants and disease represented by phecodes (groups of ICD codes). Power calculations determined the odds ratios (ORs) we were powered to detect for each association.
Results: Validating this approach, we reproduced well-established associations such as *ATM* and breast, pancreatic, and prostate cancer (OR=2.7 p=6.7E-19, OR=5.3 p=9.6E-12, OR=2.3 p=2.3E-8, respectively), and *TTN* and dilated cardiomyopathy (CM) (OR=6.0, p=1.5E-69). In addition, we did not find evidence of association

between *RINT1* or *XRCC2* and breast cancer, supporting ClinGen curations refuting these associations. This approach also produced clarifying data points for associations with less firm assertions. In cancer, we found evidence for the association of *ATM* with colon cancer (OR=2.0, p=1.5E-4), supporting ClinGen's moderate GDV. Conversely, we did not detect an association between BRCA1 or BRCA2 and melanoma or *PALB2* and *BRIP1* and prostate cancer, despite having 80% power to detect an ORs above 3. In cardiology, there was enrichment for *ALPK3* and AD hypertrophic CM (OR=6.6, p=8.1E-5), curated by ClinGen as definitive but for autosomal recessive disease, and *MYBPC3* and AD arrhythmias (OR=1.9, p=6.5E-4), curated as limited by ClinGen. There was also evidence supporting the "surgical complications" phecode and *CPT2*, suggestive of AD malignant hyperthermia (OR=2.6, p=8.6E-3).
Conclusion: Analysis of the UKB supported some gene-disease candidates while showing lack of evidence for others. These findings can be validated in additional biobanks such as the Helix Research Network™, All of Us, and MyCode cohorts. Further, this framework offers diagnostic labs a systematic and efficient method to prioritize LE GDVs.

Provider acceptance of patient-facing digital genetics service delivery tools: a qualitative study

Authors: D. Assamad[1,2,3], **S. Grewal**[2,3], M. Clausen[3], D. Hirjikaka[3], E. Reble[3], S. Luca[1], R. Hayeems[1,2], Y. Bombard[3,2], on behalf of Genetics Navigator study team; [1]Program in Child Hlth.Evaluative Sci., The Hosp. for Sick Children, Toronto, ON, Canada, [2]Inst. of Hlth.Policy, Management & Evaluation, Univ. of Toronto, Toronto, ON, Canada, [3]Genomics Hlth.Services & Policy Res. Program, Li Ka Shing Knowledge Inst., St Michael's Hosp., Unity Hlth.Toronto, Toronto, ON, Canada

Abstract:

Introduction: Digital tools are increasingly used to support the delivery of genetics services. Yet, little is known about the elements that influence provider acceptance and uptake of patient-facing digital tools in clinical genetics practice.
Methods: Semi-structured interviews were conducted with genetics providers across Canada to understand their perspectives on the elements shaping their acceptance of digital tools in clinical genetics services. Analysis followed an interpretive description approach, applying qualitative techniques including coding and thematic analysis.
Results: We interviewed 33 genetics providers across five provinces (22 genetic counselors and 11 medical geneticists ranging from 3-25+ years of practice). Four themes emerged as influencing providers' acceptance of digital genetics tools: (1) information

credibility, (2) reputable institutions, (3) workflow integration, and (4) inclusivity. (1) Providers were accepting of digital tools if the content was created from reliable sources and involved appropriate stakeholders, including genetics professionals and patient representatives. (2) Participants also discussed that their acceptance was influenced by whether the digital tool was developed and supported by reputable and trustworthy institutions. They noted that digital genetics tools developed with a commercial interest were less appealing. (3) Participants also expressed the need to consider the practical aspects of integrating a digital tool into their genetics practice. Specifically, they described considerations regarding the tool's flexibility to incorporate emerging evidence in the field of genomics, its ease of integration into clinical workflow, and the importance of institutional readiness to adopt new tools. (4) Lastly, providers noted the importance of inclusivity, suggesting digital tools consider representative imagery, language level and tone, ease of use, and health literacy level.
Conclusions: Our findings contribute to the limited body of knowledge around the drivers influencing provider acceptance of patient-facing digital genetics service delivery tools. System planners and providers looking to integrate digital tools into their clinic can use these key considerations to evaluate which tool is the best fit for their practice and optimize the success of its integration.

Does universal testing under payer medical policy equate with genetic testing coverage for patients with ovarian, pancreatic, male breast, and early-onset colorectal cancer?

Authors: E. Vaccari, T. Williams, D. Riethmaier, E. D. Esplin; Invitae Corp., San Francisco, CA

Abstract:

Background: Universal germline genetic testing (GGT) for patients with ovarian (OV), pancreatic (PANC), male breast (MB), and colorectal cancer under the age of 50 (CRC <50) is the medical necessary standard of care per clinical guidelines and many payer medical policies. We report actual coverage of GGT under these medical policies.
Methods: Patients with GGT between 6/1-12/31/2023 from a commercial laboratory were stratified by cancer type (using ICD-10s). We assessed differences in coverage rates, frequency and types of denial codes, and appeal success across cancer types. Reported p-values are from G-Tests of independence.
Results: We reviewed 8355 patients with cancer, 46% with PANC, 30% with OV, 4% with MB, and 20% with CRC <50. Of all insurance claims, 42% had no denials while 58% had at

least one denial code. Denial rates differed by the cancer types (p<2.2x10-16). CRC <50 had the most denied patient claims (74%), followed by OV (57%), PANC (53%) and MB (51%). Appeals were successful in 24% of cases across cancer types with denials. Appeal success rates were 38% for CRC <50 cases, compared to OV, PANC, and MB (18%, 18%, and 21%, respectively). Cases with denials with Medical necessity, Non-covered services, and/or Experimental denial codes made up 15% of successful appeals. Of all patients, GGT for was not covered for 34%, including 40% of patients with CRC <50, compared to 34%, 31%, and 31% for OV, PANC, and MB, respectively. Of all cases with no coverage, 28% of all patients across cancer types had clinically indicated GGT with Medical necessity, Non-covered services, and/or Experimental denial codes.
Conclusion: Despite payer medical policy affirming the medical necessity of universal GGT in these cancer types, these data demonstrate that payers did not cover almost 30% of patients with these cancers. Those denials overturned on appeal suggest those cases should not have been denied initially and emphasize the importance of ordering clinicians and laboratories collaborating on behalf of patients to ensure payers responsibly cover patient services stated in their medical policies. It is critical for providers to understand the complexities of the insurance system to be able to best advocate for patients and equitable coverage for GGT.

Costs and outcomes of opportunistic genomic screening: Findings from the Incidental Genomics randomized controlled trial

Authors: C. Mighton[1], R. Kodida[2], S. Shickh[3], E. Reble[4], J. Sam[4], M. Clausen[5], D. Hirjikaka[2], S. Grewal[2], S. Panchal[6], M. Aronson[6], S. R. Armel[7], T. Graham[8], N. Forster[9], J-M. Capo-Chichi[10], E. Greenfeld[6], A. Noor[6], I. Cohn[11], C. Morel[12], C. Elser[6], A. Eisen[8], J. Carroll[13], E. Glogowski[14], K. Schrader[15], J. Lerner-Ellis[16], R. Kim[10], K. E. Thorpe[17], K. K. Chan[17], Y. Bombard[17]; [1]St. Michaels Hosp. & Univ. of Toronto, Toronto, ON, Canada, [2]Unity Hlth.Toronto, Toronto, ON, Canada, [3]BC Cancer, Whitby, ON, Canada, [4]St. Michael s Hosp., Toronto, ON, Canada, [5]Genetics Adviser, Toronto, ON, Canada, [6]Sinai Hlth., Toronto, ON, Canada, [7]Princess Margaret Hosp., Toronto, ON, Canada, [8]Sunnybrook Hlth.Sci. Ctr., Toronto, ON, Canada, [9]Univ. Hlth.Network, University Health Network, ON, Canada, [10]Univ. Hlth.Network, Toronto, ON, Canada, [11]The Hosp. for Sick Children, Toronto, ON, Canada, [12]Univ Hlth.Network, Toronto, ON, Canada, [13]Univ of Toronto, Toronto, ON, Canada, [14]Sanofi, North Bergen, NJ, [15]BC Cancer / Univ British Columbia, Vancouver, BC, Canada, [16]Mount Sinai Hosp., Sinai Hlth., Toronto, ON, Canada, [17]Univ. of Toronto, Toronto, ON, Canada

Abstract:

Background: Concern about overwhelming the healthcare system with costly follow-up care has posed a barrier to the delivery of secondary findings (SFs) from genomic sequencing (GS). However, there is scarce evidence on costs and outcomes of SFs to inform decision-making. Through a randomized controlled trial (RCT) we aimed to evaluate the costs and effects of opportunistic screening for a broad range of SFs, encompassing risks for medically actionable and non-medically actionable monogenic disorders, carrier status for recessive disorders, pharmacogenomic variants, and risk variants for common/multifactorial disease. **Methods:** Adults with a personal and/or family history of cancer and uninformative results from standard-of-care genetic testing were recruited to the RCT from familial cancer clinics, and randomized. Participants in both arms had GS with primary cancer results returned, and the intervention arm had the choice of learning SFs in addition to primary cancer findings. Quality of life was measured at multiple timepoints up to 1 year after return of results using the 12-item short-form survey (SF-12), converted to utilities, and used to calculate quality-adjusted life years (QALYs). QALYs were compared between arms controlling for baseline utility using linear regression. Analysis followed the intention to treat approach. Trial data were linked to healthcare administrative databases held at ICES (formerly the Institute for Clinical Evaluative Sciences), which holds data on all healthcare encounters. We evaluated all healthcare system costs from return of GS results to 1 year after return of GS results. Due to the randomized design, a difference between arms is attributable to the intervention (SFs). Costs were expressed in 2023 Canadian dollars (CAD). **Results:** Participants (n=287 at baseline) were 87.1% female, 57.5% White/European, and on average 57.2 years old. Most (99.3%, 139/140) participants chose to learn SFs. All who elected to learn SFs had at least one category of SF returned. Healthcare costs after return of GS results were on average $533CAD higher per participant in the intervention arm compared to the control arm ($11,423CAD [SD $20,046CAD] vs.$10,890CAD [SD $21,777]). QALYs were higher for participants in the intervention arm than in the control arm (β=0.04, 95% CI 0.01-0.06, p=0.005). Cost-utility and cost-effectiveness analyses incorporating GS-related costs are underway and will be presented at the conference. **Conclusions:** Opportunistic screening for SFs among adult cancer patients was associated with a modest increase in post-test healthcare costs and improved quality of life in the year following return of GS results.

Session 44: Alzheimer's Disease from Gene Discovery to Multi-omics Integration

Location: Four Seasons Ballroom 2&3

Session Time: Thursday, November 7, 2024, 10:15 am - 11:45 am

Discovering Genes Associated with Alzheimer's Disease via multi-tissue and cell type Transcriptome-Wide Association Study

Authors: C. Liu[1], S. Qian[2], H. Sun[3], A. Wang[4], X. He[5], The FunGen-AD Consortium, G. Wang[6], F. Morgante[1]; [1]Clemson Univ., Greenwood, SC, [2]Univ. of Chicago, Westmont, IL, [3]Icahn Sch. of Med. at Mount Sinai, New York, NY, [4]Univ. of Hong Kong, Hong Kong, China, [5]Univ Chicago, Chicago, IL, [6]Columbia Univ., New York, NY

Abstract:

Late-onset Alzheimer's disease (AD) is a complex neurodegenerative disorder with a substantial genetic component unveiled through Genome-Wide Association Study (GWAS). Identifying genetic risk factors and elucidating their mechanisms remain challenging. Transcriptome-Wide Association Study (TWAS) was proposed to improve the statistical power and biological interpretability of GWAS by leveraging expression Quantitative Trait Loci (eQTL).

In this work, we introduce a novel TWAS resource trained on tissue, cell-type, and omics specific imputation models from multi-brain regions, harmonized across several cohorts by the FunGen-xQTL Consortium, such as eQTLs, protein quantitative trait loci (pQTLs), methylation QTLs (mQTLs) across 6 cell-types and 4 tissues from study cohorts of Religious Orders Study/Memory and Aging Project (ROSMAP), Knight Alzheimer's Disease Research Center (Knight-ADRC), Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task Study (STARNET), and Microglia Genomic Atlas (MiGA). We employ a suite of models, including single-context methods (*SuSiE, mr.ash, lasso, elastic net, bayesC, and bayesR*), and multi-context methods (*mr.mash, mvSuSiE*) that exploit shared effects across multiple contexts to improve the accuracy of gene expression imputation, especially in rare cell-types or brain regions with small sample size. We implement a multi-context causal TWAS (cTWAS) model to fine-map causal genes and variants for AD jointly while mitigating false positives from horizontal pleiotropy, using the most accurate imputation model for each gene-context pair and the latest GWASs.

By incorporating both context-specific and shared effects across various modalities in a causal framework, our approach captures a broader spectrum of genetic influences. For example, our cTWAS analysis identified Astrocytes (Ast) as the causal context for *CLU* as

opposed to Inhibitory Neurons and Oligodendrocytes, which are reported significant in conventional TWAS. Similarly, among all contexts identified by conventional TWAS, *PICALM* appears putative causal only in Ast and Microglia. The utilization of multiple methods increases the number of genes that can be imputed for TWAS analysis than using a single method, thereby increasing the power to identify associations. The use of cTWAS further limits false positives and clarifies the mechanism underlying the association. These improvements not only facilitate a more precise understanding of the genetic architecture of AD, but also yield a rigorously designed TWAS protocol and comprehensive TWAS data resources available to the community.

Integration of GWAS, 3D genomics, and CRISPRi screens in microglia implicates causal variants and genes at Alzheimer's disease loci, including at *TSPAN14*

Authors: S. Laub[1], N. Tulina[1], S. Ramachandran[2], S. Murali[1], L. Boateng[1], J. Faryean[1], M. Hoffman[1], K. Cook[3], J. Pippin[3], M. Conery[1,3], E. Burton[1,3], Y. Leung[1], A. D. Wells[3,1], L-S. Wang[1], G. D. Schellenberg[1], S. A. Anderson[3], S. F. A. Grant[3,1], A. Chesi[1]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Univ. of Wisconsin-Madison, Madison, WI, [3]Children's Hosp. of Philadelphia, Philadelphia, PA

Abstract:

GWAS of Alzheimer's disease (AD) have identified 75 risk loci. Most GWAS variants reside in non-coding regions of the genome, suggesting a regulatory role in distal gene expression. The pivotal challenge of post-GWAS follow-up is translating these genetic associations to functional mechanisms and actionable targets. To this end, we developed an integrated genomic approach based on high-resolution promoter Capture-C, ATAC-seq, and RNA-seq to identify 3D contacts between putative GWAS-implicated causal variants and their corresponding effector genes, leveraging our collection of datasets from 14 different neuronal, astrocyte, and microglial cell lines. Candidate variants were physically fine-mapped by requiring them to both reside in open chromatin and contact an open gene promoter. We identified 77 candidate regions contacting 85 putative effector genes. These findings were validated in high-throughput via a pooled enhancer CRISPRi screen with scRNA-seq readout of the human microglia cell line, HMC3. We designed 269 sgRNA guides targeting the 77 candidate enhancer regions (3 guides per region), 6 gene promoters as positive controls, and 22 non-targeting controls. Differential gene expression analyses using SCEPTRE and MAST in ~97,000 cells yielded 11 hits. We further validated the regulatory activity of several leads via bulk CRISPRi (2 regions) and luciferase assays (4 regions), including an enhancer region harboring 3 AD-associated SNPs located in an intron

of *TSPAN14*. Mutagenesis of these SNPs showed that carrying one or more AD risk alleles resulted in increased enhancer activity. Importantly, we showed (via bulk CRISPRi experiments) that this enhancer is microglia-specific and is not functional in a neuronal cell line (ReNcell VM). Precise deletion of this enhancer via CRISPR/Cas9 in 6 different HMC3 clonal lines resulted in consistent downregulation of *TSPAN14* expression, supporting its role as an AD effector gene. Further functional investigations are warranted to clarify the role of these variants and genes in AD pathogenesis. We plan to expand our screens to other cell types and anticipate this methodology will be portable to other neurodegenerative and neurodevelopmental disease contexts.

Deciphering single-cell genomic landscape of brain somatic mutations in Alzheimer's disease

Authors: S. Tichkule, P. Dong, C. Casey, E. Hennigan, A. Hong, M. Alvia, Z. Shao, J. Fullard, G. Hoffman, P. Roussos; Icahn Sch. of Med. at Mount Sinai, New York, NY

Abstract:

Introduction: Alzheimer's disease (AD) is a progressive, age-related neurodegenerative disorder that significantly impacts millions worldwide. It is characterized by the abnormal accumulation of amyloid-beta and tau proteins, leading to neuronal dysfunction and eventual death. Although the precise causes of protein misfolding and accumulation remain unclear, genetic mutations—driven by a combination of genetic, environmental, and lifestyle factors—are likely to contribute to the cascade of events that lead to AD progression. Somatic mutations (SMs), which occur post-zygotically, are believed to accumulate in aging brain cells. However, the factors that drive the accumulation of these mutations and their importance in AD-affected brain cells remain unknown. **Objective:** This study aims to determine whether the rate of SMs increases in AD compared to normal aging, identify the mutagenic processes involved, and elucidate the specific genomic locations and their association with AD pathology. **Methods:** We utilized single-cell data combining single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) from 135 AD cases and 111 unaffected controls, encompassing six brain cell types (Ex, In, Astro, Micro, Oligo and OPCs) across three brain regions (Brodmann area 22, 36 and 46). We performed de novo detection of somatic mutations in these datasets using the SComatic pipeline to determine the SM burden. We then functionally annotated these SMs, followed by a mutational signature analysis to identify the mutagenic processes involved. SM burden across cell types, brain regions and AD phenotypes was assessed using negative binomial models of the observed SM

counts. **Results:** Our findings reveal a higher burden of somatic mutations across all studied brain regions and cell types in AD cases compared to controls, which is attributed to increased defects in DNA mismatch repair and age-related changes specific to AD. Notably, the dorsolateral prefrontal cortex (DLPFC) and neuronal cell types exhibited a significantly higher SM burden compared to other regions and glial cells, respectively. Moreover, we detected somatic mutations in AD-specific genes, such as *APP* and *PSEN1*, suggesting a potential role in amyloid plaque formation during AD progression. **Conclusion:** The diverse distribution of somatic mutation burdens across various brain regions and cell types sheds light on intricate biological processes, potentially guiding more targeted approaches to understand and mitigate the impact of these mutations on Alzheimer's disease progression.

Large-scale proteomic and genomic analysis identify plasma proteins influencing human brain structure and Alzheimer's disease risk

Authors: C-Y. Chen[1], **C. Ayubcha**[2,3], B. Sun[1], T. Ge[4,5]; [1]Biogen, Cambridge, MA, [2]Harvard Med. Sch., Boston, MA, [3]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA, [4]Massachusetts Gen. Hosp., Boston, MA, [5]Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

Recent advances in proteomics technology enable us to comprehensively examine the impact of the proteome on Alzheimer's disease (AD) and intermediate phenotypes, such as brain imaging measures. Here, by leveraging large-scale proteomics, brain imaging, and AD genome-wide association studies (GWAS), we performed Mendelian randomization (MR) and colocalization analysis to investigate the impact of plasma protein levels on brain imaging measures and the link between plasma protein levels, brain imaging, and AD. We leverage the protein quantitative trait loci (pQTL) datasets from the UK Biobank Pharma Proteomics Project (UKB-PPP; 2,625 proteins by Olink assay) and deCODE genetics (Ferkingstad et al. 2021; 4,472 proteins by SomaLogic assay). UKB-PPP and the deCODE study represent the largest plasma proteomics GWAS to date based on two different proteomics platforms, where the total sample sizes are over 30,000 for each. For the brain imaging and AD genetic datasets, we used the brain imaging GWAS from UKB (453 T1 MRI and DTI imaging; Smith et al. 2021), and AD GWAS from Bellenguez et al. 2022. After extensive data processing and quality control, we performed two-sample MR and identified significant associations in 170 protein-DTI imaging pairs and 38 protein-MRI pairs for 60 unique proteins from UKB-PPP and 41 unique proteins from deCODE. Colocalization

analysis further prioritized proteins related to AD implicated in previous studies. We further extended the association between plasma protein and brain imaging to AD by performing MR and colocalization analyses using pQTL and AD GWAS. As an example, we showed that plasma progranulin (P28799) level showed a significant effect on AD risk ($P=6.79x10^{-12}$), while supported by associations between plasma progranulin level and caudate volumes (right and left T1 DTI, $P=8.41x10^{-6}$). In summary, we presented the largest to date study for the impact of plasma protein levels on brain imaging measures and a novel approach to identify biomarkers for AD. Overall, the integration of large-scale MRI, proteomics, AD phenotype, and genome-wide data holds great promise for enhancing our understanding of AD, and ultimately facilitates biomarker discovery and therapeutic development.

Unraveling the Propagation of Functional Genetic Effects in Alzheimer's Disease on a Population Scale

Authors: R. Feng[1], G. Wang[1], J. TCW[2], D. Nachun[3], A. Pelletier[2], H. Sun[4], The Alzheimer's Disease Functional Genomics Consortium, S. Montgomery[3], A. Renton[4], E. Marcora[4], X. Zhang[2], A. Goate[4], C. Cruchaga[5], P. De Jager[6]; [1]Columbia Univ., New York, NY, [2]Boston Univ., Boston, MA, [3]Stanford Univ., Stanford, CA, [4]Icahn Sch. of Med. at Mount Sinai, New York, NY, [5]Washington Univ., Saint Louis, MO, [6]Columbia Univ Med Ctr, New York, NY

Abstract:

The Alzheimer's Disease (AD) Sequencing Project Functional Genomics Consortium, consisting of over a dozen institutes, aims to elucidate the functional impact of AD-linked genetic loci. Our first collaborative project, FunGen-xQTL, integrates QTL data for histone modification, chromatin accessibility, DNA methylation, gene and protein expression and metabolomics from 62 brain, blood and CSF datasets across multiple cohorts, including in particular new cell type-specific brain snRNA and snATAC data. Through reprocessing, harmonization, QC and fine-mapping, we created a comprehensive genome-wide resource detailing the effects of genetic variation on molecular traits. Through integration of multi-context QTL and AD GWAS using methods developed within the consortium including novel molecular phenotype quantification, new xQTL discovery approaches, integrative and causal inference of cis/trans, interaction, quantile, and rare xQTL, we identified key AD susceptibility variants, loci, genes, gene sets, and cellular programs, and consequently disease risk modulation sequences.

This presentation focuses on insights from integrating FunGen-xQTL with 8 AD GWAS datasets, revealing potentially deleterious or protective common and rare xQTL conferring AD risk across 246 genes, with 95% xQTL credible sets accounting for 120 fine-mapped

GWAS loci. For established AD genes and loci, we identified new contexts providing insights into potential mechanisms of xQTL, such as CR1 with effects propagating through histone accessibility and cell specific expression. We also discovered new loci for known AD genes like BLNK in microglia, and promising new genes in AD-associated regions previously lacking strong evidence connecting to AD. Moreover, we identified 9 novel candidate genes and regions, as well as new gene programs accounting for AD association signals through cis and trans data integration. Our findings highlight the potential to elucidate complex diseases etiology from multi-context QTL perspectives, particularly underscoring the role of microglia in AD pathology. Various companion work in FunGen-xQTL, including novel approaches and discoveries from multi-context colocalization, TWAS, trans-xQTL, rare-xQTL and Mendelian Randomization will be presented throughout ASHG 2024.

Astrocytes from diverse ancestries reveal key differences in APOE expression and other AD risk genes across populations

Authors: A. Ramirez Angel[1], L. Bertholim-Nasciben[1], S. Moura[1], L. Coombs[1], F. Rajabli[1], B. DeRosa[2], P. Whitehead[1], L. Adams[1], T. Starks[3], M. Cuccaro[4], K. McInerney[1], P. Mena[5], M. Illannes-Manrique[6], S. Tejada[7,1], G. Byrd[8], M. Cornejo-Olivas[6], B. Feliciano-Astacio[9], L. Wang[1], W. Xu[10], F. Jin[11], M. A. Pericak-Vance[12], A. Griswold[1], D. Dykxhoorn[1], J. Young[1], J. M. Vance[1]; [1]Univ. of Miami, Miami, FL, [2]Univ. of Miami Miller Sch. of Med., Miami, FL, [3]Wake Forest Sch. of Med., Winston-Salem, NC, [4]John P. Hussman Inst. for Human Genomics, Miami, FL, [5]Univ. of Miami Hussman Inst., Miami, FL, [6]Inst. Natl. de Ciencias Neurológicas, Lima, Peru, [78]NC A&T State Univ, Greensboro, NC, [9]Univ. Central del Caribe, Puerto Rico, Puerto Rico, [10]Case Western Resreve Univ, Cleveland, OH, [11]Case Western Reserve Univ., Cleveland, OH, [12]Univ. of Miami Miller Sch. of Med, Miami, FL

Abstract:

In previous studies in frontal cortex, we found that individuals with European local ancestry (ELA) and two copies of the *APOE4* gene express significantly higher levels of *APOE* compared to those with African local ancestry (ALA). We also observed higher chromatin accessibility for *APOE4* carriers with ELA than for those with ALA. To explore whether these patterns are unique to *APOE4* or extend to *APOE3* carriers, we investigated the expression and regulation of *APOE* in induced pluripotent cell (iPSC) derived astrocytes developed within neural spheroids representing African (AF), Amerindian (AI), and European (EU) ancestral backgrounds. We obtained PBMCs from individuals with >95% global ancestry for EU and AI, and >85% global ancestry for AF, including both Alzheimer's

disease (AD) and cognitively normal, then these were reprogrammed into iPSCs and differentiated into neural spheroids for 76 days. Then, we characterized these spheroids using multiomic profiling, including Single Cell (sc)ATAC-seq for chromatin accessibility, scRNA-seq for transcriptome, and Hi-C analyses for chromatin interactions. We focused on the astrocytic population, as *APOE* is primarily expressed in astrocytes. Our results revealed that, in APOE 3/3 carriers, astrocytes with AF ancestry expressed more APOE than their EU counterparts, while in APOE 4/4 carriers, it was the opposite, with EU astrocytes expressing more APOE than their EU and AI counterparts, consistent with our findings in the frontal cortex. Furthermore, we observed higher APOE expression in astrocytes derived from AD patients compared to non-demented controls. Additionally, we found ancestry-specific differential expression of other 59 AD GWAS genes, including *ABCA1, CLU, SORL1*, and *IGFR1*. These results demonstrate that iPSC-derived astrocytes exhibit similar ancestral relationships for *APOE4* as observed in the frontal cortex, and also reveal regulatory differences for A*POE3* carriers, which interestingly oppose those of *APOE4*. These findings offer potential targets for manipulating *APOE4* expression and provide valuable multi-omic data for various studies of the central nervous system.

Session 45: Disease Insights from Omic-Wide Approaches

Location: Four Seasons Ballroom 1

Session Time: Thursday, November 7, 2024, 10:15 am - 11:45 am

Large-scale genome-wide association study meta-analysis across 1,962,069 individuals reveals insights into the genetic mechanisms of osteoarthritis

Authors: K. Hatzikotoulas, on behalf of the Genetics of Osteoarthritis Consortium; Inst. of Translational Genomics, Helmholtz Zentrum München, German Res. Ctr. for Environmental Hlth., Neuherberg, Munich, Germany

Abstract:

Osteoarthritis is the most prevalent musculoskeletal disease with over 600 million people affected worldwide.Here, we performed a GWAS meta-analysis for osteoarthritis in 489,975 patients and 1,472,094 controls from five major ancestry groups, across 11 joint sites. We conducted Bayesian fine-mapping of arising signals and biological pathway enrichment analysis. We highlighted effector genes by integrating information across 24 orthogonal lines of evidence. We performed a functional GWAS using single cell multiome data and transcription factor enrichment analysis to obtain an indication of important cell types and biological pathways. Finally, we analysed whole genome sequencing data, to assess association of loss of function (LOF) variants in the high-confidence effector genes with osteoarthritis using gene burden tests. We identified 962 independent variants ($P \leq 1.3 \times 10^{-8}$), of which 513 are newly reported. We established that 339 are unique and independent across all osteoarthritis phenotypes (236 new). We mapped the 339 variants to 286 loci comprising 176 newly reported loci, with 44 of the known loci also having at least one newly-reported variant. We found that 75/856 of fine-mapping credible sets included a single causal variant. For all genes within 1Mb of all credible set variants, we combined 24 orthogonal lines of evidence and identified 700 unique high-confidence effector genes with at least 3 lines of evidence in support of their involvement in osteoarthritis. The highest scoring effector gene was aldehyde dehydrogenase 1 family member A2 (*ALDH1A2*) with 11 supporting lines of evidence. We carried out overrepresentation analysis on high-confidence effector genes and found 20 significantly enriched pathways, including TGFβ signaling and skeletal system morphogenesis. We observed enrichment in transcriptional or chromatin accessibility signatures in osteoblasts, for total hip replacement, hip osteoarthritis, and finger osteoarthritis genetic association signals. Furthermore, we identified 1585 credible set variants that both reside within gene regulatory regions and affect a transcription factor binding motif in osteoblast

or chondrogenic cells. Finally, we found that the risk of disease was increased for LOF variants in *ADAMTSL3*, *VIT, COL27A1*, *IL11*, and *PMVK*. We present the largest GWAS meta-analysis of osteoarthritis, expanding the number of patients studied by 2.8-fold and identifying 513 new risk variants and 176 new risk loci. By combining functional evidence, we highlight high-confidence effector genes and distinct biological pathways, providing a solid basis for developing or repurposing targets for drug discovery.

Multi-ancestry proteome-wide Mendelian randomization offers a comprehensive protein-disease atlas and potential therapeutic targets

Authors: C-Y. Su[1], A. van der Graaf[2], W. Zhang[3], S. Selber-Hnatiw[1], T-Y. Yang[1], Y. Chen[4], K. Liang[1], G. Butler-Laporte[1], F. Matsuda[5], B. Richards[6], V. Mooser[1], J. Flannick[7], S. Zhou[1], T. Lu[8], S. Yoshiji[9]; [1]McGill Univ., Montreal, QC, Canada, [2]Univ. of Lausanne, Lausanne, Switzerland, [3]Montreal Heart Inst., Montreal, QC, Canada, [4]Lady Davis Inst., Montreal, QC, Canada, [5]Kyoto Univ., Kyoto, Japan, [6]McGill Univ. | 5 Prime Sci., Montreal, QC, Canada, [7]Harvard Med. Sch., Boston, MA, [8]Univ. of Toronto, Toronto, ON, Canada, [9]Broad Inst., Cambridge, MA

Abstract:

Background: Circulating proteins influence disease risk and are valuable drug targets, yet existing studies mainly focused on European ancestry populations. To increase the discovery of protein-phenotype associations and identify potential therapeutic targets for diverse populations, we conducted multi-ancestry proteome-phenome-wide Mendelian randomization (MR), followed by comprehensive sensitivity analyses and druggability assessment.
Methods: We performed genetic ancestry-stratified two-sample MR scanning a total of 2,265 unique proteins from two platforms (SomaScan v4 and Olink Explore 3072)—2,110 proteins from four European cohorts (n = up to 35,559 individuals), 1,144 from two African cohorts (n = up to 1,871), and 581 from a novel East Asian cohort (n = 1,823). We curated the largest GWAS for 179 diseases or traits in Europeans, 26 in Africans, and 206 in East Asians. To minimize risk of horizontal pleiotropy, we used stringent filtering by ensuring instruments were *cis*-pQTL for the proteins of interest only and had the strongest link to the corresponding protein-coding gene based on multiple sources of evidence (highest Open Targets V2G score). We then performed multiple sensitivity analyses including colocalization with PwCoCo and SharePro, a novel method we recently developed. We evaluated shared causal effects of prioritized proteins across ancestries and assessed overlap with the druggable genome, Open Targets, DrugBank, and DGIdb.

Results: We tested 726,035 protein-phenotype pairs in Europeans, 33,078 in Africans, and 115,352 in East Asians. Notably, due to minor allele frequency differences (MAF < 0.01 in Europeans), 119 proteins were instrumentable only in Africans, and 17 only in East Asians, highlighting the value of multi-ancestry inclusion. We found causal effects for 4,028 unique protein-phenotype pairs in Europeans, 55 in Africans, and 325 in East Asians (FDR < 5%). We confirmed proteins known to influence multiple phenotypes, such as GCKR in Europeans and ALDH2 in East Asians. Of 62 protein-phenotype pairs present in multiple ancestries, 51 pairs (82.3%) involving 35 proteins had concordant effects, including ANGPTL4 on triglycerides, SWAP70 on HbA1c, INHBB on HDL, and IL1RL1 on eczema. Notably, 25 of the 35 proteins (71.4%) are targeted by licensed drugs and drug candidates in clinical trials or under development.
Conclusion: This study provides the largest comprehensive atlas of protein-disease associations across three ancestries, expanding insights into disease etiology and opportunities for prioritizing therapeutic targets. Results will be publicly available at the Common Metabolic Diseases Knowledge Portal.

Transcriptome-wide association study of early substance use reveals associations between tobacco use and predicted gene expression in adolescents

Authors: J. Yu, S. Jones, J. Barth, M. Benton; Baylor Univ., Waco, TX

Abstract:

Alcohol and tobacco use in adolescents can increase lifetime risk for dependence and substance use disorders. Furthermore, adolescents engaged in early substance use are more likely to experience negative physical and mental health outcomes, making adolescent substance use a serious public health issue. Previous work has shown significant heritability for substance initiation, dependence, and age at first use; however, little is known about genetic risks contributing to early initiation of substance use (<= 14 years old for alcohol, <= 16 years old for tobacco). Additionally, most genetics studies of substance use are performed in adults, where physiological changes caused by the substance itself can confound the results, and phenotypes such as ‘age of initiation’ rely on participant recall.
Here, we perform a transcriptome-wide association study (TWAS) to identify genetic factors associated with early adolescent substance use. We use a cohort of 9,523 individuals from the Adolescent Brain Cognitive Development (ABCD) Study and examine initiation of use for individuals between ages 9-10 years old (baseline) and ages 13-14 years (4-years follow-up). TWAS is a powerful method that can detect associations between

genetic variants and phenotypes mediated by changes in gene expression. We use Joint-Tissue Imputation models to predict transcriptome levels across 43 human tissues, including 11 brain regions, in the ABCD participants. We find significant associations between early use of nicotine-containing products and the predicted expression of several genes, including *CIB1*, *MYO5B*, and *CFAP53*. Among these associations, increases in predicted expression of *CIB1* and *CFAP53* and decreases in predicted expression of *MYO5B* are associated with tobacco use. Our most significant hit was *CIB1*, a calcium binding protein implicated in cell signaling and a variety of human diseases, including Alzheimer's disease. Predicted expression of this gene was associated with tobacco use in the amygdala, putamen, spinal cord, and substantia nigra. We did not observe any significant associations between predicted expression and early initiation of alcohol use. Overall, we identify several genes whose predicted expression is associated with early initiation of tobacco use in adolescents in the ABCD cohort. Our work provides a unique opportunity to gain a deeper understanding of the genetics of adolescent substance use and contributes to the development of new approaches to prevent early use.

All by All of Us: common and rare variant association testing in 245,000 whole genomes across diverse ancestry groups

Authors: K. Karczewski[1], W. Lu[2], R. Carroll[3], Y. Wang[1], R. Grant[2], M. Solomonson[2], A. Kouame[4], J. Brogan[5], J. Qian[4], M. Basford[3], M. He[4], M. Lyons[6], J. Linder[3], W. Zhou[1], A. Musick[7], D. King[2], A. Martin[1], D. Roden[8], B. Neale[1]; [1]Massachusetts Gen. Hosp., Boston, MA, [2]Broad Inst., Cambridge, MA, [3]Vanderbilt Univ. Med. Ctr., Nashville, TN, [4]Vanderbilt Univ., Nashville, TN, [5]Vanderbilt Univ., Nashville, MA, [6]Vanderbilty Univ. Med. Ctr., Nashville, TN, [7]NIH - All of Us Program, Bethesda, MD, [8]VUMC, Nashville, TN

Abstract:

Whole genome sequencing of population biobanks provides opportunities for comprehensive phenome- and genome-wide ("all x all") association analyses, including rare burden and common variant testing of coding and non-coding variation, across many human phenotypes. The All of Us cohort is much more diverse than many other large biobanks (80% of the 245,400 participants are underrepresented in biomedical research), and this diversity has the potential to enhance gene discovery and pleiotropy relevant to human disease.
We built a comprehensive multi-ancestry phenome- and genome-wide analysis framework, using a generalized mixed model framework to perform common variant association tests and rare variant burden tests, followed by meta-analysis of each across

genetic ancestry groups. We applied this framework to perform association testing of 8,895 ancestry-phenotype pairs, spanning 3,416 anthropometric, biomarker, disease, and medication use phenotypes derived from electronic health records and surveys across 6 genetic ancestry groups for each quantitative trait and binary trait with over 200 cases. We have performed extensive quality control of our association tests, observing good calibration to simulated phenotypes as well as consistency with previous GWAS for a broad set of phenotypes.
We identify thousands of associations for rare variant burden analysis and common variant associations enabled by this diverse cohort. For instance, we identify an association between rare loss-of-function variants in TIMD4 and increased triglycerides, providing direct functional evidence that disruption of this gene is associated with heart disease. We find 9 associations between pLoF variants and human diseases that were not observed in the UK Biobank exome study, and highlight the impact of diversity in identifying these signals. Finally, we perform meta-analysis between gene burden results in AoU and UK Biobank to increase power further.
We present a roadmap for iterative releases of the data up to 1 million individuals. This will empower more well-powered association tests for diseases with less than 1% prevalence. We publicly release the full dataset of summary statistics on the Controlled Tier of the All of Us Researcher Workbench, which researchers can query or download, and develop an interactive public-facing browser for rapid querying of diseases, traits, and genes of interest.

Genome-wide association study and predictors of lymphocyte-related blood cell traits in Hispanic/Latino newborns

Authors: Y. Li[1], B. Alonzo[1], S. Myint[1], M. Ince[1], L. Kachuri[2], R. Lu[1], A. de Smith[3]; [1]Univ. of Southern California, Los Angeles, CA, [2]Stanford Univ., San Francisco, CA, [3]Univ. of Southern California, San Gabriel, CA

Abstract:

The contribution of heritable genetic variation to hematological traits in adults is well established, yet little is known regarding the genetic architecture of blood cell traits measured in early-life. We focused on Hispanic/Latino children due to higher risks of acute lymphoblastic leukemia (ALL), asthma, and more severe immune responses to COVID-19 infection, than in non-Hispanic white children. We conducted GWAS of 24 lymphocyte-related blood cell traits in 400 Hispanic/Latino newborns and examined associations with genetic ancestry and birth-related characteristics.

Cord blood (CB) aliquots were obtained from deidentified newborns of reported Hispanic/Latino ethnicity from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center. Hematopoietic stem cell- and lymphocyte-related blood cell traits were profiled by flow cytometry. Samples were genotyped using the Axiom Precision Medicine Diversity Research Array and imputed using the TOPMed reference panel. Blood cell profiles were transformed to z-scores and outliers were removed. GWAS models were adjusted for sex and principal components. Proportions of European, African, and Indigenous American ancestry were inferred using RFMix.
We identified a genome-wide significant signal at chr2p11.2, spanning the *CD8A* and *CD8B* genes, with the same lead SNP rs35505884 (minor allele frequency=0.105) observed for both total CD3+/CD4+/CD8+ (P=2.16x10^{-26}) and CD4+/CD8+ (P=5.66x10^{-17}) T-cells. SNP rs35505884 is an eQTL for CD8A in GTEx, and variants in this region were previously associated with T-cell traits in adults of European ancestry. Genome-wide significant loci were also detected for total CD45+, CD19+/CD56+CD16+, CD4+/8-, and CD4-/8+ measures. Proportion of global Indigenous American ancestry was associated with CD4+/8+ (beta=1.03 P=0.0058) T-cell measures after adjustment for sex and birth-related covariables. Further, delivery by C-section was associated with reduced total CD3 (P=0.0007) and CD3+/4+ (P=0.0008) T-cells compared to vaginal delivery.
Our GWAS of blood cell traits in Hispanic/Latino newborns identified strong signals for CD4+/8+ double-positive T-cell counts at birth, suggesting distinct immune development that may contribute to disparities in ALL incidence and response to infections observed in Hispanic/Latino children. Indigenous American ancestry and type of delivery may be associated with specific blood cell profiles. Larger studies are warranted to understand the genetic architecture of blood cell trait variation in early-life, which may inform mechanisms of disease and immune development in children.

Omic Risk Scores are Associated with Cross-Sectional and Longitudinal Chronic Obstructive Pulmonary Disease-Related Traits Across Three Cohorts

Authors: I. Konigsberg[1], L. B. Vargas[1], K. A. Pratte[2], K. Buschur[3], D. E. Guzman[3], T. D. Pottinger[3], A. Manichaikul[4], E. C. Oelsner[3], E. R. Bleecker[5], D. A. Meyers[5], V. E. Ortega[5], S. A. Christenson[6], D. L. Demeo[7], B. D. Hobbs[7], C. P. Hersh[7], P. J. Castaldi[7], J. L. Curtis[8], R. G. Barr[3], J. I. Rotter[9], S. S. Rich[4], P. G. Woodruff[6], E. K. Silverman[7], M. H. Cho[7], K. J. Kechris[1], R. P. Bowler[2], E. M. Lange[1], L. A. Lange[1], M. R. Moll[7]; [1]Univ. of Colorado - Anschutz Med. Campus, Aurora, CO, [2]Natl. Jewish Hlth., Denver, CO, [3]Columbia Univ., New York, NY, [4]Univ. of Virginia, Charlottesville, VA, [5]Mayo Clinic, Scottsdale, AZ, [6]Univ. of California - San

Francisco, San Francisco, CA, [7]Brigham and Women's Hosp., Boston, MA, [8]Univ. of Michigan, Ann Arbor, MI, [9]Lundquist Inst., Torrance, CA

Abstract:

Background: Individuals with chronic obstructive pulmonary disease (COPD) demonstrate marked heterogeneity with respect to lung function decline, emphysema, mortality, exacerbations, and other disease-related outcomes. Omic risk scores (ORS) estimate the cumulative contribution of omics, such as the transcriptome, proteome, and metabolome, to a particular trait. In this study, we aimed to assess the predictive value of ORS for COPD-related traits in both smoking-enriched and general population cohorts.
Methods: We developed and tested ORS in n=3,339 participants of the Genetic Epidemiology of COPD (COPDGene) study with blood RNA-sequencing, proteomic, and metabolomic data collected at the second study visit. On 80% of the data, we trained single- and multi-omic risk scores on a variety of traits using elastic net penalized regression with 10-fold cross-validation. We included 24 cross-sectional and 5 longitudinal traits (where the trait was measured approximately 5 years apart), enriched for measures of disease severity, exacerbations, and traits derived from spirometry and computed tomography scans. We used multivariable models to test association of ORS with outcomes in a held-out COPDGene testing set and externally validated findings in participants of SubPopulations and InteRmediate Outcome Measures In COPD Study (SPIROMICS) (n= 2,177) and Multi-Ethnic Study of Atherosclerosis (MESA) (n=1,000), adjusting for potential confounders and multiple testing.
Results: Among the 24 cross-sectional traits in the COPDGene testing set, there were significantly associations with 70 of 72 single-omic ORS (false discovery rate adjusted *p*-value < 0.05). Significant associations were found in 5 of 15 longitudinal ORS with changes in trait values between COPDGene visits, including with forced expiratory volume at one second (FEV_1) decline over 5 years and annually. We observed significant association with the relevant traits for all 38 cross-sectional ORS tested in SPIROMICS and for 15 of 24 in MESA. Generally, proteomic and metabolomic risk scores displayed stronger trait associations than transcriptomic risk scores, and multi-omic risk scores had higher predictive capacity than single-omic risk scores.
Conclusions: ORS constructed from blood-based omics can be leveraged to predict cross-sectional and future COPD-related traits in both smoking-enriched and general population cohorts. ORS for clinical use would require phenotype-focused risk score construction and replication.

Session 46: Diverse Epigenetic Marks in Health, Diagnosis, and Disease

Location: Room 505

Session Time: Thursday, November 7, 2024, 10:15 am - 11:45 am

H3K36 methylation - a guardian of epigenome integrity

Authors: G. Shipman[1], R. Padilla[1], C. Horth[1], e. Bareke[1], F. Vitorino[2], B. Garcia[2], J. Majewski[1]; [1]McGill Univ., Montreal, QC, Canada, [2]Washington Univ. Sch. of Med., St. Louis, MO

Abstract:

Histone modifications form a highly interdependent network and are essential for cells to see and read appropriate regions of the genome. While some are associated with heterochromatin and sequester DNA from transcriptional activators, others are associated with transcriptionally active euchromatin. Both types of modifications are essential for cells to maintain their identity and function. Some modifications are antagonistic and incompatible in the vicinity of others due to steric bulk or structural barriers. In the case of synergistic modifications, many chromatin effectors have reader domains that recognize specific marks and use their presence as a template to deposit their own catalytic products. Previously, we investigated the role of histone 3 lysine 36 methylation (H3K36me) in its function of maintaining an active chromatin state by influencing both repressive and activating modifications at H3K27, and by serving as a substrate for the deposition of DNA methylation (DNAme). Here, we establish mouse stem cell models with successive KOs of all H3K36 methyltransferases (KMTs): NSD1, NSD2, NSD3, SETD2, and ASH1L - which results in progressive depletion of all H3K36me, and allows us to identify the contributions of each KMT to H3K36me. Importantly, the H3K36me depleted cells allow us to assess its influence on other modifications, gene expression, and its ability to maintain active chromatin states. We find NSD3 deposits H3K36me2 at active enhancers, and this is critical for maintaining expression of many enhancer-dependent genes. Similarly, we show ASH1L deposits H3K36me2 at the regulatory elements of a small subset of developmentally important genes. We describe the abilities of H3K36me2/3 to exclude H3K27me2/3 (repressive marks) from transcribed genes, and find even in the complete absence of H3K36me, H3K27me3 does not freely invade genes, and that other factors, such as DNAme likely prevent its deposition. Unexpectedly, we uncover a novel relationship between H3K36me and H3K9me3. In our multi-KO cells, we find H3K9me3 is redistributed away from large heterochromatic domains and into euchromatin, resulting in

upregulation of previously silenced genes and transposable elements, while some previously active genes gain H3K9me3 and become silenced. We performed Hi-C and find that the total loss of H3K36me and redistribution of H3K9me3 results in drastic genome-wide decompartmentalization and diminishes long-range chromosomal contacts. Given the high number of developmental disorders and cancers driven by mutations in different KMTs, our work uncovers a broad range of crucial functions of H3K36me in the maintenance of epigenome integrity.

m^{6}A mediated epitranscriptomic dynamics in human brain development and disease

Authors: A. Shafik[1], Y. Peng[2], M. Chen[2], P. Jin[1]; [1]Emory Univ., Atlanta, GA, [2]Univ. of Chicago, Chicago, IL

Abstract:

N6-methyladenosine (m^{6}A) is a chemical modification prevalent in the brain that affects many aspects of RNA metabolism. Recent studies have linked m^{6}A to brain development and neuronal function, as well as many brain disorders. Here, we profiled m^{6}A in 100 normal postmortem brain samples from five regions (BA9, BA24, hippocampus, caudate, and thalamus). Genome-wide m^{6}A profiling reveals brain region- and age-specific m6A dynamics. Functional enrichment analysis showed region-specific methylation on genes typically upregulated in those regions. In hippocampus, unique methylation patterns were found on genes related to hippocampal morphology. Distinct methylation patterns were also observed on transcription factors, ion channels, and receptors across brain regions. Many potassium ion channels and receptors exhibited predominant hypermethylation in BA regions compared to other tested regions. Dysregulation of potassium channels has been implicated in neuropsychiatric disorders, which are most associated with BA9. Hypermethylation and upregulation of receptors could indicate increased activity or changes in signaling pathways related to various functions such as cognition, emotion, and sensory perception, mainly related to BA9. Additionally, many transcription factors exhibited hypermethylation in the thalamus, underscoring their critical role in regional patterning during brain development. Key transcription factors associated with thalamic development, such as *GBX2*, *FOXBP2*, *GLI1*, and *OTX2*, displayed hypermethylation and upregulation only in the thalamus. Age-related changes were most significant in the BA regions, where methylation typically decreased with age, correlating with reduced gene expression. Functional categorization of these differentially methylated genes revealed an enrichment of genes broadly involved in neurodevelopment. Enrichment analyses also

linked m^{6}A-dependent age changes to autism, nicotine dependence, schizophrenia, and epilepsy. Finally, integrating m^{6}A data with WGS identified 800 m^{6}A quantitative trait loci (m^{6}A-QTLs) enriched near lead SNPs. Cross-referencing m^{6}A-QTLs with GWAS highlighted disease associations, notably a m^{6}A-QTL identified within the *ZNF804A* gene, linked to schizophrenia with a high colocalization posterior probability (PP4 = 96.2%), indicating the same variant is causal for both the QTL and the disorder. Our comprehensive profiling of m^{6}A reveals intricate patterns of methylation across diverse brain regions and ages, underscoring its critical role in neurodevelopment and function, with implications for brain disorders.

Most genetic effects on DNA methylation are shared across tissues

Authors: S. Villicana Munoz[1], X. Zhang[1], C. Shore[1], E. Hannon[2], J. Bell[1]; [1]King's Coll. London, London, United Kingdom, [2]Univ. of Exeter Med. Sch., Exeter, United Kingdom

Abstract:

Background: DNA methylation (DNAm) is an epigenetic mark that regulates gene expression and maintains cell identity. Although DNAm levels are malleable, a proportion of sites are influenced by genetic variants, known as methylation quantitative trait loci (meQTLs). The extent to which meQTL effects are tissue-specific or shared is not well characterised. This study investigates cross-tissue genetic associations with the human methylome in blood, adipose and skin tissues. **Methods**: We analysed previously published DNAm data comprising 1,840 whole blood, subcutaneous adipose, and skin tissue biopsy samples, collected from 1,295 unique participants from the TwinsUK cohort. The methylome data included approximately 400,000 DNAm sites profiled with the HumanMethylation450 BeadChip. Independent association analyses were performed for each tissue, comparing genotypes with DNAm levels. For *cis*-meQTLs, all SNPs within 1 Mb of the DNAm sites were considered, while for *trans*-meQTLs, we focused on a LD-clumped set of SNPs. Summary statistics from the three tissues were combined and re-analysed in a two-stage empirical Bayes model to determine tissue-shared effects by magnitude and direction of association. Ongoing comparisons are being made with brain meQTLs from an external cohort, the Brains for Dementia Research (BRD) cohort, against meQTLs from TwinsUK. **Results**: Approximately half of the tested DNAm sites showed *cis*-meQTL associations in at least one of the three tissues, with adipose tissue exhibiting most significant effects. Approximately 70% of *cis*-meQTL associations were consistently observed across all three tissues, and a further 18% were shared across at least two tissues. In contrast, 99% of *trans*-meQTLs were significant across the three tissues. While most meQTLs had a consistent direction of association across tissues, their effect sizes

varied. Notably, adipose tissue and skin displayed more similar genetic effects on DNAm compared to blood. We validated 97% of our blood meQTLs, and are currently extending the cross-tissue comparison with samples from brain tissue. Finally, meQTL signals are enriched for genetic variants associated with complex diseases and are located in genes relevant to systemic disorders, such as *TRIT1*, *GCN2* or *CAV2*. **Conclusion**: Our study demonstrates evidence for extensive tissue-shared genetic regulation of DNA methylation, providing the first comprehensive investigation of cross-tissue meQTL effects in both *cis* and *trans*. These findings contribute to a deeper understanding of the mechanisms underlying DNA methylation variability, opening new avenues for the study of diseases.

Multi-platform long read genomics identifies methylation outliers in rare disease

Authors: T. Jensen[1], J. Gorzynski[1], J. Nguyen[1], A. Podesta[1], R. Kaur[2], D. Bonner[3], C. Reuter[1], R. Ungar[1], P. Goddard[1], K. Smith[1], J. Bernstein[4], M. Wheeler[1], E. Ashley[1], S. Montgomery[1]; [1]Stanford Univ., Stanford, CA, [2]Howard Univ., Washington D.C., DC, [3]Stanford Hlth.Care, Palo Alto, CA, [4]Stanford Univ. Sch. of Med., Stanford, CA

Abstract:

Whole-genome sequencing has revolutionized rare disease diagnosis, yet our incomplete ability to interpret genomes leave cases unsolved. Increasingly, additional omics assays are used to increase diagnostic yield. Profiling methylation has been shown to help diagnose imprinting disorders and define global “episignatures” of over 50 syndromes. Long-read genomics provides an opportunity to profile both sequence and DNA methylation in a single assay; however, lower accuracy, lower depth at CpG sites compared to arrays, and lack of population references for these technologies have limited their applications. Here, we present a new method to overcome these challenges with the goal of calling rare methylation outliers from long-read data. We apply this method to a cohort of rare disease genomes: 70 nanopore genomes from the Undiagnosed Disease Network (UDN) and 41 PacBio genomes from the GREGoR consortium, sequenced from whole blood (n=67) and cultured fibroblasts (n=44).

We compare methylation profiles between nanopore and PacBio genomes and find high correlation across segmented genome blocks. Technology-specific differences are largely explained by varying depth and accuracy. Applying batch correction techniques to adjust for depth and sequencing technology allows for joint analysis of PacBio and nanopore data.

To call methylation outliers, we compute a tissue reference by averaging methylation across samples. Then per CpG we calculate a score for how much each sample deviates from this reference. Finally, we segment this score to find regions of contiguously deviant

CpGs and compute corrected z-scores for each region. In total we find 7,936 methylation outlier regions (|z-score|>3) across all samples, with a median of 16 per sample from blood and 30 from fibroblast. 55% of these methylation outlier regions lie nearby protein coding genes, including 587 rare disease genes and 10 imprinting disorder genes. Integrating RNA-seq, we find methylation outlier genes are more likely to also be expression outliers. Among genes that are both methylation and expression outliers, we identify a strong negative correlation of methylation and expression, suggesting hypermethylation outliers are involved in silencing.
Using this approach, we detect a hypomethylation outlier in an imprinted region of GNAS in an individual diagnosed with Pseudohypoparathyroidism type IB. Using the same long-read data we detect a rare 3kb deletion of a cis-imprinting control element in the STX16 gene. In summary, with long-read technology, we detect methylation differences that are rare in the population, linked to complex structural variants, and involved in rare disease.

A novel single-cell sequencing method for *CHD2* variant classification in childhood epilepsies

Authors: N. Bodkin, J. Calhoun, G. L. Carvill; Northwestern Univ., Chicago, IL

Abstract:

Background: Childhood epilepsies often have a genetic basis, with over 100 genes implicated in their etiologies. However, determining the pathogenicity of variants within these genes remains a challenge. Variants in some epilepsy-related genes cause global epigenetic changes that can be measured through DNA methylation with modern genomic technologies. Leveraging these differences, “epi-signatures” have been designed to distinguish between pathogenic and benign variants in the clinical setting. However, this method is limited by the availability of clinical samples with inconclusive genetic findings. To overcome this limitation, we present a novel single-cell sequencing method to classify thousands of variants *in vitro*. **Methods:** Isogenic HAP1 cell lines harboring known pathogenic and benign *CHD2* variants were generated and underwent methylation profiling by array to establish a HAP1-specific epigenetic signature indicative of *CHD2* pathogenicity. The transcriptomes and proteomes of the similar lines were profiled to establish a transcriptomic and proteomic signature for pathogenicity. A comprehensive library of pegRNAs was designed to introduce thousands of unique *CHD2* variants into HAP1 cells. Each pegRNA will integrate into the genome with a unique barcode that will enable multi-modal profiling of cells. Using microfluidics, genomes will be digested with the methylation-sensitive enzyme HhaI at the single-cell

level, then amplified and sequenced. A similar workflow will be applied for sc-RNA seq to identify transcriptomic changes in individual cells. Following demultiplexing and variant calling, the methylation and transcriptomic profile of each cell will allow us to assign a machine learning-based pathogenicity score (using a random forest classifier) for each introduced variant. **Results:** We have observed global transcriptomic, epigenetic, and proteomic changes in *CHD2* knockout cells and HAP1 cells harboring pathogenic *CHD2* variants. At the transcriptome level, we observed significant downregulation of multiple genes in the presence of pathogenic variants, including *PXDN* and *MFAP* ($p<0.001$). We have also found significant changes in the methylation status of thousands of loci and a decrease in the abundance of Nestin protein in HAP1 cells harboring non-functional *CHD2*. **Conclusion:** This project proposes a novel method for classifying *CHD2* variants using epigenetic and transcriptomic profiling of engineered HAP1 cells. This high-throughput approach can significantly improve the accuracy and efficiency of variant classification, improve genetic diagnoses, and enable future targeted therapies in childhood epilepsies.

In-utero rescue of neurological dysfunction in a mouse model of Wiedemann-Steiner syndrome

Authors: T. Reynisdottir[1], K. J. Anderson[2], J. Ouyang[1], A. O. Snorradottir[3], A. Zuberi[4], V. B. DeLeon[5], H. T. Bjornsson[6]; [1]Louma G. Lab. of Epigenetic Res., Faculty of Med., Univ. of Iceland, Reykjavik, Iceland, [2]Dept. of Genetics and Molecular Med., Landspitali Univ. Hosp., Reykjavik, Iceland, Reykjavik, Iceland, [3]Dept. of Pathology, Landspitali Univ. Hosp., Reykjavik, Iceland, Reykjavik, Iceland, [4]The Jackson Lab., Bar Harbor, ME, [5]Dept. of Anthropology, Univ. of Florida, Gainesville, FL, [6]Univ. of Iceland, Reykjavik, Iceland

Abstract:

Wiedemann-Steiner syndrome (WDSTS) is a Mendelian disorder of the epigenetic machinery (MDEM) characterized by intellectual disability, growth retardation, hypotonia and hypertrichosis. WDSTS is caused by *de novo* heterozygous loss-of-function variants in the gene encoding the epigenetic regulator, lysine methyltransferase 2A (KMT2A). Genetic disorders now have potential to be diagnosed quite early (even *in-utero*) *and* several strategies are emerging to rescue genetic abnormalities at the cellular level. However, how does one figure out which disorders have the potential to be treated *in-utero*? Here we use one such possible strategy for WDSTS by developing a novel mouse model (*Kmt2a*$^{+/LSL}$) carrying a loss-of-function variant placed between two LoxP sites, allowing *in-utero* and postnatal genetic rescue. *Kmt2a*$^{+/LSL}$ mice demonstrate core features of WDSTS compared

to wildtype (WT) littermates including growth retardation ($p<0.001$), craniofacial abnormalities ($p<0.0001$), hypotonia ($p<0.05$) and hypertrichosis of the abdomen ($p<0.05$) as well as neurological defects with decreased dentate gyrus surface area ($p<0.05$) and visuospatial memory defects ($p<0.05$). A majority (60%) of *Kmt2a*$^{+/LSL}$ mice also show a white belly spot potentially indicating that neural crest function may also be disrupted in these mice. To test the *in-utero* malleability of WDSTS, we crossed this model with a Nestin-Cre model, restoring KMT2A levels in the nervous system during mid-to-late gestation (E12-P0). Compared to littermates without the Nestin-Cre we observe molecular rescue of gene expression abnormalities (RNA-Seq) and histone modification (H3K4me1) abnormalities (CUT&Run) in primary NPCs from these mice. Surprisingly, given KMT2As role as a transcriptional activator and histone methyltransferase, the H3K4me1 signal is elevated in the *Kmt2a*$^{+/LSL}$ mice and decreases upon rescue. This may indicate that heterozygous loss of KMT2A leads to a compensatory cellular response within the cells, that could potentially be the driver of the disease pathogenesis. Furthermore, on the phenotypic level we observe normalization of dentate gyrus area and visuospatial memory defects in mice with both alleles (LSL and Nes-Cre) compared to littermates but no effect on growth retardation or craniofacial defects, as expected. The frequency of the white belly spot also drastically decreased from 60% to 12%. Taken together, our results demonstrate genetic, molecular, and phenotypic malleability of WDSTS making it an ideal candidate for future therapeutic strategies.

Session 47: From Variant to Function: Prediction and Understanding Variants Function

Location: Mile High Ballroom 2&3

Session Time: Thursday, November 7, 2024, 10:15 am - 11:45 am

Defining the function of disease variants with CRISPR editing and multimodal single cell sequencing

Authors: Y. Baglaenko[1], M. Curtis[2], M. Al Suqri[2], R. Agnew[2], A. Nathan[3], H. M. Mire[4], A. Mah-Som[5], D. R. Liu[6], G. A. Newby[6], S. Raychaudhuri[2]; [1]Cincinnati Children's Hosp., Cincinnati, OH, [2]Brigham and Women's Hosp., Boston, MA, [3]Harvard Univ., Boston, MA, [4]Harvard Med. Sch., Boston, MA, [5]Havard Med. Genetics Training Program, Boston, MA, [6]Broad Inst., Boston, MA

Abstract:

Genetic studies have identified thousands of individual disease-associated non-coding alleles, but identification of the causal alleles and their functions remain critical bottlenecks. Even though CRISPR-Cas editing has enabled targeted modification of DNA, inefficient editing leads to heterogeneous outcomes within individual cells, limiting the ability to detect functional consequences of disease alleles particularly in primary human cells. To overcome these challenges, we present MINECRAFTseq: a multi-omic single cell sequencing approach that directly identifies genomic DNA edits, assays the transcriptome, and measures cell surface protein expression. Applying this novel approach in cell lines, we show CRISPR induced deletion of a regulatory region specifically associated with HLA-DQB1 gene expression, knockout of FBXO11 by splicing disruption alters CD40 mediated signaling, and CRISPR Cas base-editing identifies the causal variant in an expression quantitative trait locus of RPL8. We then apply this technique to primary human CD4 T cells. We induce an early stop codon with base editors in *PTPRC* and identity significant changes in gene and protein expression associated with hyperactivation and lineage skewing. Finally, we identify the state-specific effects of an *IL2RA* autoimmune variant in primary human T cells polarized under Treg but not Th1 conditions. Multimodal functional genomic single cell assays including DNA sequencing bridges a crucial gap in our understanding of complex human diseases by directly identifying causal variation in primary human cells.

Functional genomics applied to mapping the gene regulatory mechanisms downstream of neuron-astrocyte interactions

Authors: B. Li, K. T. Hagy, A. Safi, A. N. Pederson, S. J. Reisman, G. E. Crawford, C. Eroglu, C. A. Gersbach; Duke Univ., Durham, NC

Abstract:

Neurons and astrocytes are the most abundant cell types in the brain. They interact intricately, influencing crucial aspects of brain biology like synaptic biology, metabolism, and ion homeostasis. The disruption of their interactions is implicated in diverse pathologies, such as in Alzheimer's Disease.

Despite the importance of neuron-astrocyte interactions, their effects on gene expression and the epigenome remain largely unexplored. To elucidate these effects, we used an *in vitro* co-culture model of human pluripotent stem cell-derived excitatory neurons with mouse neonatal astrocytes. Using RNA-seq and ATAC-seq, we observed extensive changes in transcriptional profiles and chromatin landscapes in both cell types due to their interactions with the other, compared to when they are cultured in isolation. In neurons, ~25% of all expressed genes and ~10% of putative regulatory elements (pREs) responded to co-culture, impacting pathways such as synaptogenesis, metabolism, and transcriptional regulation. Notably, these changes encompassed genes linked to neuropsychiatric and neurodegenerative diseases, such as glutamate receptor subunit genes and APOE. These findings highlight the influence of neuron-astrocyte interactions on the epigenome and gene programs.

We then sought to understand how neurons regulate gene expression in response to astrocyte signals. First, we identified more than two hundred transcription factors (TFs) that altered expression in the neurons in response to astrocytes. They are the hypothetical regulators of the broader molecular phenotypes. We then implemented single-cell CRISPR epigenome perturbation screens of pREs surrounding the genes encoding these TFs to examine the molecular effects of each perturbation. We discovered dozens of perturbations at promoters and enhancers that altered the expression of local TFs, such as SIX3 and LHX1. These TFs, in turn, influenced downstream astrocyte-responsive genes and pathways, and cell fate decisions. Together, these results indicate these TFs to be central mediators of the molecular phenotypes in neurons following interactions with astrocytes.

In summary, we have shown that neurons and astrocytes impact each other's epigenome and transcriptome, mediated by TFs. Perturbations of genes encoding these TFs recapitulate aspects of neuron-astrocyte interactions. In the future, these perturbations could be used to rewire disease-relevant pathways in a neuron-specific manner. Altogether, these findings highlight the role of epigenetic changes in mediating the

molecular phenotypes of neuron-astrocyte interactions, crucial to diverse areas of neuroscience.

CircRNA mediated polyadenylation alteration contribute to Alzheimer's disease pathogenesis

Authors: F. Wang[1], Y. Li[1], Y. Feng[2], B. Yao[1]; [1]Dept. of Human Genetics, Emory Univ. Sch. of Med., Atlanta, GA, [2]Dept. of Pharmacology and Chemical Biology, Emory Univ. Sch. of Med., Atlanta, GA

Abstract:

Circular RNAs (circRNAs) are a class of covalently closed, single-stranded RNAs that are enriched in the brain and play critical roles in learning and memory by precisely modulating the availability of RNA-binding proteins (RBPs). Increasing evidence link circRNA dysregulation to age-related neurodegenerative disorders, such as Alzheimer’s disease (AD). However, how AD-associated circRNA alterations contribute to transcriptome changes at the molecular level remain poorly understood. In this study, we discovered genome-wide and cortical-specific circRNA alterations within the critical window for AD progression (between 5-month and 7-month of age) using a well-established 5xFAD mice and discovered brain-region specific circRNA dysregulations in AD mouse model. By systematically comparing circRNA dysregulation between AD mouse model and several specific human patient postmortem brain regions, we found hundreds of conserved circRNAs are consistently dysregulated across species. Among these common AD-related circRNAs, we identified circGigyf2 as the most and specifically repressed circRNA in 7-month of age with full AD manifestation. In human AD patients, circGigyf2 is also progressively downregulated and associated with human AD severity. Mechanistically, we identified multiple circGigyf2-interacting RBPs associated with AD, including the cleavage and polyadenylation factor 6 (CPSF6), an essential factor to regulate mRNA polyadenylation (pA). Pre-mRNAs undergo post-transcriptional pA process that is essential to produce functional mature mRNA with distinctive half-life and ability to translate into functional proteins. However, little is known about how circRNAs could contribute to AD-associated gene expression change via modulating alternative pA. We have experimentally validated the physical interaction between circGigyf2 and CPSF6 and showed downregulation of circGigyf2 in AD led to decreased miRNA targets expression and increased polyadenylation efficiency of many known CPSF6 targets, both contribute to neuron apoptotic process. Importantly, we demonstrated circGigyf2 knockdown in a mouse immortalized neuron cell line led to hyperactivation of miRNAs and CPSF6 with

increased cell apoptosis upon induction, strongly support the causative role of circGigyf2 deficiency in AD disease progression. Overall, our results unveiled pathological alteration of cortical circRNA landscape in AD and identified novel molecular mechanisms underlying dysregulation of conserved circRNA-CPSF6-polyadenylation pathways that lead to neuron apoptosis during AD pathogenesis.

In Silico Module Perturbation Analysis unlocks a functional understanding of the dynamic gene networks in single-cell data

Authors: Z. Shi[1], S. Morabito[1], V. Swarup[2]; [1]Univ. of California, Irvine, Irvine, CA, [2]Univ. of California Irvine, Irvine, CA

Abstract:

Biological functions are governed by gene regulatory networks that orchestrate a diverse array of dynamic cell states. These networks are altered throughout development, aging, disease, and in response to molecular stimuli, however such perturbations are exceedingly difficult to measure directly with technologies like single-cell RNA-seq (scRNA-seq). Here we propose a novel computational framework, CO-expression Module Perturbation Analysis for Cellular Transcriptomes (COMPACT), to perform in-silico gene expression perturbations in single-cell data, and to track downstream changes in cell state dynamics. Building off of our previous method high-dimensional Weighted Gene Co-expression Network Analysis (hdWGCNA), COMPACT applies direct perturbations to co-expression network hub genes, and uses the network structure to propagate the perturbation signal to other linked genes. This framework is highly flexible to perform knock-in, knock-down, or knock-out perturbations on different networks and sets of genes and in different cell lineages, allowing researchers to explore a wide range of strategies mimicking various experimental conditions and interventions. We demonstrate the efficacy of COMPACT across established paradigms of cellular dynamic response, including mouse oligodendrocyte differentiation, human disease-associated microglia, and existing single-cell perturbation data. Notably, our in-silico perturbation simulations with COMPACT have unveiled disease-relevant phenotypic outcomes resulting from co-expression module perturbations in human disease-associated microglia. These findings offer valuable insights into the molecular mechanisms underlying cellular states in disease, and identify novel targets for therapeutic intervention. In summary, our work introduces COMPACT as a powerful tool for functionally dissecting gene network perturbations and elucidating their disease-relevant impact. By leveraging the wealth of information contained within single-cell transcriptomic data, COMPACT facilitates comprehensive analyses of gene regulatory

networks, paving the way for advancements in our understanding of cellular dynamics and disease pathology.

De Novo Precise Splice Site Predictor Using Deep Learning and Integration with Minimap2 for Enhanced Long-Read Sequence Alignment

Authors: S. Yang[1,2], N. Huang[2], H. Li[1,2]; [1]Harvard Med. Sch., Boston, MA, [2]Dana Farber Cancer Inst., Boston, MA

Abstract:

INTRODUCTION: Splice sites play a pivotal role in pre-mRNA splicing to form mature mRNA, crucial for understanding gene structure and function. Current tools, like Minimap2, face challenges in precisely aligning spliced sequences due to static penalty matrices for non-canonical splicing events. Here, we propose a novel deep learning approach to predict splice site scores dynamically, potentially enhancing future integration with sequence alignment tools such as Minimap2, to improve genomic annotation and alignment accuracy. **METHODS**: We developed a novel splice site predictor using a CNN-biLSTM architecture to calculate splicing site probability scores. This model processes sequences around potential splicing signals through three distinct CNN blocks to capture local features, integrated with biLSTM layers that consider sequence interdependencies. Performance of the model was benchmarked against Splice AI and SPLAM through metrics such as accuracy, losses, and PR-AUC. We then aim to enhance Minimap2. Adjusting the penalty matrix to reward matches with canonical splice sites and dynamically penalize non-canonical splice sites based on integrated splice scores are expected to enhance splice junction precision. **RESULTS**: The CNN-biLSTM model demonstrated superior performance, with validation accuracies of 92.5% for donors and 92.0% for acceptors, and PR-AUCs of 0.90 and 0.89, respectively. It significantly outperformed Splice AI and SPLAM across all metrics. **CONCLUSION:** This study establishes the effectiveness of our CNN-biLSTM model in predicting splice sites with high precision. While this work currently focuses on the development and validation of splice site predictors, future work will explore integrating these predictions with Minimap2. The integration will adjust the penalty matrix dynamically based on predicted splicing scores, which is anticipated to enhance splice junction identification and the accuracy of long-read sequence alignments, thereby refining genome annotations and expanding our understanding of gene regulation and structure.

Classification of rare nonsynonymous variants to identify individuals at low risk of disease: introducing variants of potential risk

Authors: A. Bolze, K. M. Schiabor Barrett, M. E. Levy, N. Telis, N. L. Washington, W. Lee, E. T. Cirulli; Helix, San Mateo, CA

Abstract:

Background: Genetics can be used to identify individuals at lower risk of disease. Benefits of identifying individuals at lower risk of disease include: avoiding unnecessary screening in these individuals, better allocation of resources, and facilitating discussions about risk with individuals in the 'average' risk group. However, the fear of incorrectly classifying an individual at low risk has hindered the implementation of a 'low genetic risk' program.
A person may be classified at low genetic risk if they meet 2 conditions: (i) have a low polygenic risk score, and (ii) do not have a 'monogenic risk' for the disease. While existing polygenic risk scores enable the identification of a sub-population at low risk, there are no guidelines on how to classify variants in disease-causing genes in the context of ruling out any possible genetic risk of disease. Should all rare missense variants in genes established to be associated with the disease be considered of 'potential risk'? For example, is it ever appropriate to tell a woman with a rare, missense variant in BRCA2 that they are at low risk for breast cancer?
New concept: Here, we define Variants of Potential Risk (VPR) as the smallest group of rare variants that has a non-zero effect on disease risk.
Objective: Identify the best way to classify variants of potential risk (VPR), minimizing the risk of wrongly assigning a person to the low risk group, while keeping the number of VPRs to a reasonable number. If 100% of the population has a VPR, the genetic testing is not helpful.
Methods: We tested our concept on breast cancer. We sequenced and analyzed electronic health records of more than 50,000 participants in the US. Common variants are excluded from variants of potential risk (VPR) because they could be part of the polygenic risk score model. Rare variants were put in 12 mutually exclusive groups based on 3 criteria: genes analyzed, Minor Allele Frequency (MAF), and predicted functional impact.
Results: Groups of variants that collectively had a non-zero effect were: very rare (<0.01%) predicted damaging variants in 11 genes. Altogether, 1.6% of the population had a pathogenic variant for breast cancer and an additional 4.9% had a variant of potential risk (VPR). Excluding these individuals from a low-risk category, 9.3% of the population was considered at low risk (bottom 10% of the polygenic risk score distribution for each genetic similarity group). Women at low risk had a Hazard Ratio of 0.33 (CI, 0.25-0.43) for breast cancer compared to the rest of the population.

Conclusion: We propose a new approach for variant interpretation when the objective is to identify a subset of the population at low genetic risk of disease.

Session 48: Novel Genetic, Genomic, and Epigenetic Resources in the Era of Big Data

Location: Room 401

Session Time: Thursday, November 7, 2024, 10:15 am - 11:45 am

The developmental Genotype-Tissue Expression projects

Authors: K. Ardlie, dGTEx Consortium; Broad Inst., Cambridge, MA

Abstract:

Children are critically underrepresented in clinical and research studies and medically underserved, despite the numerous diseases that occur during gestation, development, or early in life. Many treatments for diseases in adults are inappropriate for children. Most large gene expression and cell atlas resources have focused primarily on adult tissues and lack information on human development and critical cell states that may only manifest during specific developmental windows. In parallel, little is known about the conservation of developmental programs across non-human primate (NHP) species, with implications for human evolution, drug development and clinical testing. The developmental Genotype-Tissue Expression projects, spanning both human (dGTEx) and non-human primates (NHP-dGTEx) aim to address this gap. The human dGTEx will consist of reference normal (non-diseased) samples from up to 74 tissue sites from 120 donors spanning four distinct age groups (postnatal, early childhood, prepubertal, and post pubertal). This is matched by the non-human primate NHP-dGTEx consisting of samples from age groups developmentally matched to the human study, with additional prenatal animals, from a total of 126 rhesus macaques and 72 common marmosets. Comprehensive histological review is being performed on all tissues, with data generation including whole genome reference sequences on all donors/animals, and extensive bulk RNA, single cell RNA, chromatin accessibility, and spatial gene expression profiles across most collected tissues. The dGTEx projects aim to produce reference datasets for human and NHP development, enabling research into developmental and childhood disorders, the effect of gene perturbations during development, and translational therapeutics.

The Clinical Genome Resource (ClinGen): Advancing Genomic Knowledge through Global Curation

Authors: S. Plon, The Clinical Genome Resource; Baylor Coll. Med., Houston, TX

Abstract:

Background: The Clinical Genome Resource (ClinGen) is a National Institutes of Health-funded program established eleven years ago that defines the clinical relevance of genes and variants for medical and research use. **Methods:** ClinGen working groups develop standards for data-sharing and frameworks for evaluation of evidence resulting in expert curation of genomic knowledge. Expert panels curate the validity of monogenic disease relationships, pathogenicity of genetic variation, dosage sensitivity of genes, oncogenicity of somatic variants, and actionability of gene-disease-interventions using ClinGen standards, infrastructure and curation interfaces. **Results:** ClinGen curated results are available on clinicalgenome.org and classified variants are also submitted to ClinVar, a publicly available database at NCBI/NLM/NIH. As of June 2024, there are more than 2400 active members from 69 countries. Over 2800 genes have been curated (2647 gene-disease relationships for validity, 1568 genes for dosage sensitivity, and 464 gene-condition pairs for actionability). 7038 unique variants have been classified for pathogenicity across 92 mitochondrial and nuclear genes for monogenic disorders. Using ClinGen processes, new curation frameworks and accompanying curation interfaces are near completion for HLA alleles for complex disease, polygenic risk score actionability and pharmacogenomics, in addition to our collaboration with the CIViC database team for somatic cancer variation. A systematic approach to address justice, equity, diversity, and inclusion in ClinGen has been developed. **Conclusion:** ClinGen's knowledge can be used to build evidence-based genetic testing panels, interpret copy number variation, resolve discrepancies in variant classification, guide disclosure of genomic findings to patients and assess new predictive algorithms. ClinGen's iterative improvement of standardized methods for evidence curation, ongoing growth of the curation ecosystem and incorporation of new technologies will support our continued efforts to provide rigorous, evidence-based curation that is reproducible, sustainable, and clinically relevant for people of all backgrounds.

The New York Genome Center ALS Consortium combines postmortem tissue transcriptomics with whole genome sequencing to empower biological discovery

Authors: J. Humphrey[1], A. Basile[2], U. Evani[2], A. Oku[2], M. Byrska-Bishop[2], A. Corvelo[2], W. Clarke[3], H. Geiger[2], R. Fu[2], S. Chang[2], K. BP[1], A. Real[2], D. Fagegaltier[2], Y. Kim[2], A. Runnels[2], S. Fennessey[2], N. Propp[2], G. Gursoy[4], D. Knowles[5], G. Narzisi[2], M. Zody[2], N. Robine[2], T. Raj[1], H. Phatnani[5], NYGC ALS Consortium; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]New York Genome Ctr., New York, NY, [3]Outlier Genomics, Saskatoon, SK,

Canada, [4]Columbia Univ., New York, NY, [5]New York Genome Ctr. & Columbia Univ., New York, NY

Abstract:

Amyotrophic Lateral Sclerosis (ALS) is a progressively fatal neurodegenerative disease striking in early middle age, leading to motor decline and death within 5 years. ALS is highly heritable, with 10% of patients having a family history. Rare mutations have been discovered in over 30 genes, providing genetic diagnoses to 25% of patients. In parallel, GWAS have identified 15 risk modifying loci in patients of European and East Asian ancestry. Critically, the mechanistic basis behind these genes is largely unknown, stalling progress in translating findings to therapeutics. While the latest omics technologies have potential to transform ALS research, the cost of generating omics data at scale is prohibitive for individual labs. Therefore, collaborative consortia play a fundamental role in generating large-scale resources for the research community.

We present the New York Genome Center (NYGC) ALS Consortium, a multi-center collaboration spanning 42 contributing sites with 73 participating members. Blood and postmortem tissue from patients with ALS and non-neurological controls, as well as a minority of other neurological diseases, were sent to NYGC for both whole genome sequencing (WGS) and RNA sequencing (RNA-seq). Although previous data freezes have already contributed to 16 publications, here we present an overall analysis of the full consortium cohort.

The WGS dataset consists of 4,857 donors sequenced to average 30x coverage, of which 488 were predicted to be non-European or admixed using genetic ancestry estimation. Using joint genotyping with GATK (v3.5), a total of 108,448,943 single nucleotide variants and 14,673,485 indels were identified. Additionally, 136,503 structural variants were discovered by combining insertion and deletion calls from Manta, Absinthe, and MELT with external PacBio long-read data from HGSVC. Known disease-associated short tandem repeats were quantified with ExpansionHunter (v5.0). Pathogenic expansions in C9orf72 were found in 337 donors and were validated by repeat-primed PCR.

The RNA-seq dataset consists of 2,725 tissue samples from 718 donors (419 ALS, 126 controls), across 15 different regions including frontal/motor cortex, cerebellum, and spinal cord. We identified large numbers of differentially expressed genes between ALS and control, and observed shifts in cell-type composition using MuSiC2 deconvolution.

We will present additional use cases combining the power of the two datasets together, including mapping expression and splicing quantitative trait loci, and associating rare genetic variation with expression outliers.

FILER 2.0: Unified access to >100,000 omics datasets across >1,000 cell types and tissues

Authors: P. Kuksa[1], F. Leung[1], J. Cifello[1], P. Gangadharan[1], L. Carter[1], S. Cole[1], E. Greenfest-Allen[2], O. Valladares[1], L-S. Wang[1]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Natl. Inst. on Aging Genetics of Alzheimer`s Disease Data Storage Site (NIAGADS), Philadelphia, PA

Abstract:

Background: Functional genomic (FG) experimental and annotation data such as gene-regulatory or enhancer elements, transcription factor binding sites, open chromatin regions, and 3D chromatin interactions are widely used in genetic analyses. Such analyses can identify potentially causal GWAS variants, their target genes, affected cell types and biological mechanisms through tissue- and cell type-specific annotations. The FILER repository is a database developed by the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) with the most comprehensive harmonized, indexed, and searchable human FG data collection.
Method: FILER (PMID: 35047815) is built to provide scalable, streamlined access to massive FG data collections across diverse omics datatypes/assays in consistent, BED-based formats. FILER 2.0 integrates >100,000 genomic data tracks across >1,100 cell types curated, harmonized, and integrated from 54 data sources and 52 experimental assay types. All data in FILER 2.0 is organized into >200 indexed data collections and can be queried by cell type/tissue, assay type, genomic feature type and other data attributes, allowing researchers to customize annotations for their analyses. Importantly, genomic interval queries are supported with high efficiency using Giggle-based genomic indexing (PMID: 29309061).
Result: FILER 2.0 resource has been updated to include new data types such as uniformly processed chromatin interaction datasets (e.g., Hi-C, Capture-C from SRA, 4DN, ENCODE) and quantitative trait loci (QTL) data (e.g., splicing, methylation, histone, protein expression) with >28,000 newly integrated tracks from 33 data sources. Metadata has been improved to include standardized cell types, data categories, and provides flexible key-value-based sample and track descriptions. Additionally, for all FG datasets that were originally available only in GRCh37/hg19 build, their lifted or re-processed GRCh38/hg38 native datasets have been added to FILER. The FILER workflow has also been updated to use an improved high-throughput hipFG pipeline (PMID: 37947320) for data and meta-data harmonization and integration.
Availability: FILER is freely available (https://lisanwanglab.org/FILER) and is also deployable in cloud or HPC environments.

Whole exome sequencing of 44,028 British South Asians in Genes and Health uncovers 2,917 genes with putative human knockouts for systematic characterization

Authors: H. Kim[1], K. Walter[2], G. Kalantzis[2], E. Fauman[1], M. Miller[1], J. Gafton[3], K. Hunt[3], Genes and Health Industry Consortium, Genes and Health Research Team, R. Trembath[4], S. Finer[3], H. Martin[2], d. van heel[3]; [1]Pfizer, Cambridge, MA, [2]Wellcome Trust Sanger Inst., Hinxton, United Kingdom, [3]Queen Mary Univ. of London, London, United Kingdom, [4]King's Coll. London, London, United Kingdom

Abstract:

Individuals homozygous for complete loss-of-function variants (referred to as human knockouts) can provide valuable insights into new biological discoveries and aid in drug development. However, identifying such individuals has been challenging due to their rarity. Genes and Health is a cohort of British South Asians with high parental relatedness and autozygosity. Genes and Health Industry Consortium conducted exome sequencing of 44,028 volunteers, allowing the identification of homozygous carriers of predicted loss-of-function (pLOF) variants and evaluation of their phenotypic profiles in the electronic health record (EHR) data.We identified 122,347 rare (MAF < 1%) high confidence pLOF variants in 15,792 genes with 3,978 pLOF variants in 2,917 genes having at least one homozygous carrier. Notably, 2,208 genes were uniquely found to have homozygous pLOF carriers in Genes and Health exomes when compared to 1,649 genes found in 396,651 European-ancestry exomes in UK Biobank despite the smaller sample size. The number of genes with homozygous pLOF carriers in Genes and Health continues to increase linearly with the sample size, suggesting the potential benefits of sequencing additional individuals.The increased number of genes with homozygous pLOF carriers enabled us to assess the characteristics of genes with human knockouts. As expected, genes that are constrained, essential, developmental, and ubiquitously expressed were less likely to have homozygous pLOF carriers. Analyses using Open Targets drug database revealed that drug target genes are relatively depleted of genes with homozygous pLOF carriers (OR=0.66, p-value=4.6e-6), consistent with previous reports on pLOF variation. Interestingly, among the drugs with antagonistic actions, there was a trend that drugs and drug target genes with homozygous pLOF carriers were more likely to advance beyond phase I in clinical trials (OR=2.1, p-value= 6.2e-5 and OR=5.8, p-value=0.082, respectively). This observation aligns with the hypothesis that genes with human knockouts may be safer for therapeutic antagonism, but caution is needed as many unknown factors can influence the clinical success of drugs. This abstract provides an overview of the putative human knockouts identified in Genes and Health exomes. Ongoing and future efforts aim to further characterize pLOF variants

and carriers and derive biological insights, including 1) manual curation of pLOF variants and genotypes, 2) investigating the molecular effects of homozygous pLOF genotypes, 3) reviewing EHR data of homozygous pLOF carriers of high interest, and 4) recalling such individuals for further phenotypic characterization.

Enhanced Genetic Insights from Brain Region-Specific GWAS Using Deep Unsupervised Learning Derived Endophenotypes on UK Biobank T1-Weighted MRI Data

Authors: S. Islam[1], Z. Xie[2], H. Wei[1], **D. Zhi**[2]; [1]McWilliams Sch. of BioMed. Informatics, Univ. of Texas Hlth.Sci. Ctr., Houston, TX, [2]UTHlth.Houston, Houston, TX

Abstract:

Genetic studies of brain imaging endophenotypes offer many new insights into genetic underpinning of brain structures and functions. Unsupervised approaches are a new trend that enables the discovery of new phenotypes beyond human biases. However, existing unsupervised approaches are mostly limited to focusing uniformly on the whole brain. Focusing on brain region-specific GWAS studies enhances sensitivity and specificity in detecting genetic associations, which otherwise might be diluted in whole-brain analyses. This targeted approach may increase the resolution of genetic discovery by targeting distinct brain regions, crucial for understanding and treating specific neurological conditions. Previously we developed a 3D convolutional autoencoder to extract phenotypes from brain MRIs for genome-wide association studies. Here, we trained a new autoencoder model with weighted reconstruction loss that is preferentially incorporating information about predefined target regions. The autoencoder was trained on 4,597 subjects to compress the linearly registered T1-weighted MRI data into a compact, low-dimensional representation and then reconstruct it, with increased emphasis on selected target regions. We used white matter (WM), gray matter (GM), and cerebrospinal fluid(CSF) as target regions, training three separate autoencoders with bottleneck dimensions of 128, and conducted GWAS on the latent representations in 22,880 UK Biobank subjects.
By GWAS of our region-weighted endophenotypes, we discovered 94, 162, and 447 new significant SNPs for white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF), respectively, with 47, 39, and 310 of these SNPs not overlapping with any loci from our earlier whole brain study. There are 5, 6, and 14 of these loci, respectively, being new compared to our earlier whole brain study. Importantly, we identified genes related to white matter ('MRVI1', 'RFX4', 'RIC8B'), gray matter ('ZIC4', 'ZIC1', 'L3MBTL3', 'NHSL1'), and CSF ('AS3MT', 'CNNM2', 'NT5C2', 'NAV3', 'DOCK9', among other 15 genes). These findings

suggest that by designing targeted weighting to the deep learning loss function, we can direct the model to discover region-specific genes.

Session 49: Polygenic Risk Scores: Novel Methods for Modeling Risk

Location: Room 501

Session Time: Thursday, November 7, 2024, 10:15 am - 11:45 am

JointPRS: A Comprehensive Framework for Genetic Prediction Across Populations Incorporating Genetic Correlation and Combining Meta-Analysis and Tuning Strategies

Authors: L. Xu[1], G. Zhou[1], W. Jiang[1], H. Zhang[2], L. Guan[1], H. Zhao[3]; [1]Yale Univ., New Haven, CT, [2]Natl. Cancer Inst., Bethesda, MD, [3]Yale Univ. Sch. of Publ. Hlth., New Haven, CT

Abstract:

Background: Genetic prediction accuracy for non-European populations is hindered by the limited sample size of GWAS data. Many studies overlook the potential or risk of small individual-level datasets for parameters tuning in non-European populations.
Method: We propose JointPRS, a comprehensive framework that models multiple populations and estimates chromosome-wise cross-population genetic correlation using GWAS summary statistics. JointPRS has robust performance even without individual-level datasets for tuning parameters. When non-European individual-level data are available, we adopt a data-adaptive approach combining meta-analysis and tuning strategies and inheriting the merits from both strategies, further enhancing prediction performance and robustness.
Results: JointPRS was applied to predict 22 continuous traits in five populations (European; East Asian; African; South Asian; and Admixed American) and evaluated using the UK Biobank (UKB) and All of Us (AoU) cohorts. We considered three data scenarios: [1] no tuning data, [2] tuning and testing data from the same cohort, and [3] tuning and testing data from different cohorts, and compared with six other state-of-the-art methods: PRS-CSx, SDPRX, XPASS, PROSPER, MUSSEL, and BridgePRS. When no tuning datawere available, Joint PRS outperformed three other applicable methods (PRS-CSx, SDPRX, XPASS), with R-square improvements ranging from 2.32% to 53.6% in East Asians, 21.2% to 64.6% in Africans, 22.7% to 74.5% in South Asians, and 20.9% to 88.5% in Admixed Americans. When tuning and testing data were available from the same cohort (UKB), JointPRS showed improvements over all six methods, with relative gain from 2.91% to over 100% in East Asian, 31% to over 100% in Africans, 2.57% to 100% in South Asians, and 27.3% to over 100% in Admixed Americans. When tuning and testing data were from different cohorts (UKB and AoU), JointPRS outperformed the six methods, with

improvements from 7.96% to over 100% in Africans, and 10.8% to over 100% in Admixed Americans.
Conclusion: JointPRS consistently outperforms other cutting-edge methods across different data scenarios, while maintaining model simplicity with computational efficiency comparable to PRS-CSx. This framework leverages GWAS summary statistics and individual-level tuning datasets, offering a robust solution for accurate genetic risk prediction in diverse populations.

A Novel Polygenic Risk Scoring Framework Integrating Common and Rare Variants for Enhanced Genetic Prediction Across Ancestries

Authors: J. Williams[1], T. Chen[2], X. Hua[1], W. Wong[1], K. Yu[1], P. Kraft[1], X. Li[3], H. Zhang[1]; [1]Natl. Cancer Inst., Rockville, MD, [2]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA, [3]Univ. of North Carolina at Chapel Hill, Chapel Hill, NC

Abstract:

Background: Polygenic risk scores (PRS) predict complex diseases and traits by aggregating multiple genetic variants. Current PRS models focus on common variants, missing the potential of rare variants (minor allele frequency < 1%) to uncover the hidden heritability of complex traits. Recent advances in whole exome and genome sequencing (WES/WGS) allow for the inclusion for rare coding and noncoding variants. Here we introduce RICE (polygenic Risk predictions Integrating Common and rarE variants), a novel PRS framework optimized for biobank-scale sequencing data, aimed at improving genetic risk prediction accuracy across diverse genetic ancestries.
Method: RICE incorporates common and rare variant PRSs. The common variant PRS integrates PRSs from multiple existing methods. For the rare variant PRS, RICE leverages genomic functional annotations to identify significant rare variant gene masks with STAAR. These masks are used to create multiple PRSs using penalized regression, which are then combined using an ensemble machine learning algorithm. We assessed our proposed methods using simulations and analyzed UK Biobank WGS data from 137,012 independent individuals with diverse genetic ancestries on 11 complex traits—six continuous traits and five binary traits.
Results: Simulation studies demonstrate that RICE yields a significant 71% increase in the average effect size of PRS per standard deviation (SD) compared to top existing common variant methods. Within RICE, the inclusion of the rare variant PRS contributes 38% to the total PRS effect size. In real data analyses of six continuous traits, RICE leads to an average increase of 50% in effect size of PRS per SD, with the highest gain of 81% observed in

individuals of African ancestry. Notably, for HDL, RICE outperforms existing common variant PRS methods by 48%. For the five binary traits, we observe significant improvements of RICE, including a 51% increase of effect size of PRS per SD for type 2 diabetes in South Asian participants.
Conclusion: RICE significantly advances PRS by incorporating rare variants, offering a more accurate and inclusive approach to genetic risk prediction.

Modeling diagnostic code dropout of schizophrenia in electronic health records improves phenotypic data quality and transferability of polygenic risk scores for a diverse Veteran cohort

Authors: D. Burstein[1,2], S. Tomasi[1], S. Venkatesh[1,2], M. Rizk[1], P. Roussos[1,2], G. Voloudakis[1,2]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]James J. Peters VA Med. Ctr., Bronx, NY

Abstract:

Background To identify cases in large electronic health record (EHR) linked-biobanks, researchers typically leverage diagnosis code counts to infer phenotypic status. Nevertheless, such methodologies disregard temporal changes within EHR data and could incorrectly label misdiagnosed patients. Prior research has also indicated that Black patients in the United States disproportionately receive misdiagnoses for a range of neuropsychiatric conditions, including schizophrenia. Consequently, we need rigorous phenotyping approaches that account for diagnostic code dropout to promote equitable genomic research in EHR linked-biobanks. **Methods** We utilize XGBoost to predict future diagnostic code dropout using prior diagnosis and medication codes on patients with a history of schizophrenia from EHR data in the Million Veteran Program (n = 14,046). We benchmark the performance of our model on an independent holdout set against two methodologies that leverage diagnosis code counts to infer phenotypic status (PMIDs: 34465180, 31613361). We then investigate the utility of leveraging diagnostic code dropout as a proxy for misdiagnosis by evaluating our model on a second independent holdout set (n = 145), where diagnoses were attained through chart review. Finally, we construct data-driven definitions for schizophrenia by formulating additional exclusionary criteria based on the top negative diagnostic predictors from our XGBoost model. We generate polygenic risk scores (PRS) using existing GWAS data from the PGC (PMID: 35396580) with PRS-CS to quantify the transferability of the GWAS results from a predominately European ancestry cohort to a diverse Veteran cohort with our new data-driven definitions for schizophrenia. **Results** The XGBoost model strongly predicts future diagnostic code

dropout in schizophrenia on the independent holdout set (AUROC = 91.2%, AUPRC = 94.9%). When predicting schizophrenia diagnoses labeled through chart review, our XGBoost model increases the AUPRC by 9.6% compared to competing methodologies. Finally, by leveraging our new data-driven definition for schizophrenia from our XGBoost model, we observed an 85% and 34% increase in the PRS log odds ratios for African ancestry Veterans (log(OR) increases from .118 to .218, $p = 1.74 \times 10^{-3}$) and European ancestry Veterans (log(OR) increases from .519 to .697, $p = 8.25 \times 10^{-10}$), respectively. **Conclusions** We anticipate that modeling diagnostic code dropout in electronic health record linked-biobanks will be critical for both mitigating disparities in genomic research and boosting the translational utility of polygenic risk scores across all demographic groups.

A Deep Ensemble Encoder Network Method for Improved Polygenic Risk Score Prediction

Authors: O. B. Ozdemir, R. Chen, R. Li; Cedars Sinai Med. Ctr., Los Angeles, CA

Abstract:

Background Polygenic risk scores (PRS) have been developed to combine effects from multiple genetic variants to improve genetic risk predictions of complex traits. However, current PRS methods linearly combine genetic variants into a single score that could limit their ability to capture potential complex genetic interactions and non-linear genetic effects predictive of the phenotypes.
Method In this study, we introduce DeepEnsembleEncodeNet (DEEN), a novel method that leverages an autoencoder (AE) for latent genetic feature representation and a fully connected deep neural network (FCNN) for prediction. DEEN utilizes multiple autoencoder structures to capture both linear and non-linear SNP relationships, generating a lower-dimensional latent representation of genetic data. These representations are then fed into a fully connected deep neural network for prediction, allowing for the learning of non-linear effects and variable weighting across genomic regions. The model is trained using both binary and continuous phenotypes in UK Biobank (UKBB), with performance evaluated internally on the UKBB holdout testing dataset and externally on the All of Us (AoU) dataset.
Results DEEN achieved improved predictive performance across all tested phenotypes compared to traditional PRS methods such as Lasso, PRS-CS, and PCA-based FCNN. In the UKBB, DENN showed an increase in the area under the curve (AUC) between 2.58% to 3.67% for Type 2 Diabetes (T2D) and 1.12% to 1.85% for Hypertension compared to existing PRS methods. Similarly, for continuous phenotypes, DEEN consistently yielded lower

mean squared error (MSE) values compared to existing PRS methods. Specifically, DEEN lowered on average 10% MSE for Body Mass Index, 12-14% for cholesterol, 15-17% for High-Density Lipoprotein, and 13-14% for Low-Density Lipoprotein. Independent validation using the AoU dataset confirmed DEEN's robustness. DEEN trained in UKBB improved AUC for binary phenotypes between 1.5-3% and decreased MSE for continuous traits between 3 to 9%. We also evaluated the clinical significance of DEEN by comparing the case enrichment between the high (top 5 or 10%) and low-risk (bottom 5%) groups for the two binary phenotypes. DEEN can better stratify the risk groups with a maximum improvement of the odds ratio of enrichment of 13.51% for T2D and 9.13% for hypertension compared to standard PRS methods.
Conclusion These results across two independent biobanks show that DEEN holds substantial promise for improving genetic risk predictions by leveraging autoencoders to generate flexible latent genetic representations and fully connected neural networks to create powerful predictive models.

Integrative polygenic score modeling with tissue-specific annotation improves polygenic scores transferability

Authors: X. Tian[1], T. Fabiha[2], W. F. Li[1,3], K. Dey[2], M. Kellis[1,3], Y. Tanigawa[1,3]; [1]Computer Sci. and Artificial Intelligence Lab., Massachusetts Inst. of Technology, Cambridge, MA, [2]Mem. Sloan Kettering Cancer Ctr., New York, NY, [3]Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

Systematic characterization of functional annotations like histone modification, tissue-specific expression, and transcription factor binding profiles could enhance PGS transferability across genetic ancestry groups by prioritizing putatively causal alleles in predictive models. However, consensus on the best practices for combining such large-scale resources in PGS modeling has yet been reached.
We hypothesize that large-scale integration of tissue-specific functional annotations through statistical learning improves PGS transferability by incorporating biological priors, by which we introduce candidate variant-to-function (cV2F) informed inclusive PGS (iPGS). We analyze 406,659 ancestrally diverse individuals in UK Biobank and develop predictive models for four traits with clear causal tissues, focusing on genome-wide functional annotations across 1.3 million genetic variants.
In cV2F-iPGS, we use gradient boosted trees to learn the optimal combination of ENCODE4 functional annotations from fine-mapped variants across ~100 traits. We accordingly prioritize likely causal variants when fitting penalized regression on individual-level data

from ancestry-diverse individuals. We assess model improvement of those with cV2F-informed priors compared to the vanilla iPGS, which only considers statistical correlations in fitting PGS models.
We show ancestry- and tissue-matched cV2F scores are most effective in improving prediction. Specifically, for predicting lymphocyte count in Africans, ancestry- and tissue-matched cV2F-iPGS show the best predictive performance (R^2=.0073), a 35.13% improvement over the vanilla model (R^2=.0054), a 25.86% improvement over the ancestry-mismatched model (R^2=.0058), and a 23.72% improvement over the tissue-agnostic model (R^2=.0059). The highest improvement is seen in the spirometry measure FEV1/FVC ratio, where tissue-matched cV2F-iPGS (R^2=0.0067) showed a 36.21% improvement over the vanilla model (R^2=0.0049). Overall, we found ancestry- and tissue-matched cV2F improve transferability of iPGS scores by an average of 13.5% (95% CI: [3.6%, 23.4%], p-value: .004) across the four selected traits.
Lastly, we provide locus-level biological interpretations in cV2F-iPGS models. We find, for example, tissue-matched eQTL signals help improve PGS transferability.
Overall, our approach is the first to leverage multimodal biological priors in fitting PGS models on individuals across the continuum of genetic ancestry, improving PGS accuracy and specificity.

Functional gene embeddings improve rare variant polygenic risk scores

Authors: S. Londhe[1], J. Lindner[1], Z. Chen[2], F. Hölzlwimmer[3], E. Holtkamp[4], F. Casale[5], F. Brechtmann[6], J. Gagneur[1]; [1]Technical Univ. of Munich, Garching, Germany, [2]Deutsches Herzzentrum München, Klinik an der Technischen Univ. München, Munich, Germany, [3]Technical Univ. Munich, Garching b. München, Germany, [4]Technical Univ. Munich, Garching bei München, Germany, [5]Helmholtz Zentrum München, Munich, Germany, [6]Technical Univ. Munich, Garching, Germany

Abstract:

Rare variants can exhibit large effects, aiding the discovery of effector genes underlying traits and attractive drug targets. Moreover, modeling rare variant effects in addition to common variants can help construct phenotype predictors that generalize better across populations and better identify individuals at high disease risk. However, rare variant association testing (RVAT) is statistically challenging due to sparsity and a high multiple-testing burden. Gene-set RVATs address this issue by performing burden tests or variance-component tests such as SKAT on a-priori-defined groups of genes such as pathways. However, this approach does not naturally lead to pinpointing individual genes or deriving phenotype predictors.

Here, we present FuncRVAT, a model that leverages gene function information using a Bayesian neural network trained to predict phenotypes from rare variants. FuncRVAT learns a prior on the magnitude of gene impairment effects on traits as a function of multidimensional gene embeddings derived from genome-wide assays [1]. The embeddings mitigate the bias towards well-studied genes while retaining functional information, and the model is not restricted by the limitations of the gene-set RVATs. We used DeepRVAT [2] as gene impairment scores.
We trained FuncRVAT on 41 quantitative traits from 161,850 whole-exome sequencing samples from the UK Biobank. FuncRVAT matched or surpassed existing phenotype prediction methods on most traits in held-out test data, significantly outperforming an association test-based phenotype predictor on 11 traits. To assess the accuracy of the gene effect estimates, we considered as ground truth the effects computed by ordinary least squares on 209,492 further unrelated individuals. Remarkably, gene effects estimated on the discovery cohort with FuncRVAT correlated better with this ground truth than effects estimated by ordinary least squares, even when restricted to burden-test significant genes. We found that FuncRVAT captures modest-to-small effect sizes that are often missed by association tests. Moreover, FuncRVAT identifies associations missed by previous rare variant association studies but are known through experimental validation, such as the association of *CPXM2* with blood pressure and *TPCN1* with calcium levels.
Collectively, these results show that FuncRVAT provides increased sensitivity in gene discovery, more robust gene effect size estimates, and more accurate phenotype prediction by borrowing information across functionally related genes.
[1] Brechtmann et al. *NARGAB* 5.4 (2023)
[2] Clarke, Holtkamp et al. *bioRxiv* (2023)

Session 50: The Context of All in Which We Live: Gene by Environment Interactions

Location: Four Seasons Ballroom 4

Session Time: Thursday, November 7, 2024, 10:15 am - 11:45 am

The Genetic Basis of Environmental Exposures in the Personalized Environment and Genes Study (PEGS)

Authors: C. Campbell[1], R. N. Noga[2], F. S. Akhtari[1], A. DiFrank[3], J. Mack[4], A. Burkholder[1], D. C. Fargo[1], C. P. Schmitt[1], J. Hall[1], J. S. House[5], W. Rick[1], A. A. Motsinger-Reif[1]; [1]NIEHS, Durham, NC, [2]UNC, Chapel Hill, NC, [3]NC State, Raleigh, NC, [4]NIH | Univ. of Cambridge, Cambridge, United Kingdom, [5]NIEHS, RTP, NC

Abstract:

Genome-wide association studies (GWAS) have been widely used to assess the genetic basis of behavior, with a large focus on addiction (e.g. cigarette smoking and alcohol use). However, a gap in knowledge exists beyond these behaviors, leaving a lack of understanding of how genetics influence an individual's likelihood of environmental exposures. We report here the results of examining the genetics of exposure in the Personalized Environment and Genes Study (PEGS). PEGS is a North Carolina-based, ethnically diverse research cohort with extensive questionnaire data on health and exposures from medications, lifestyles, and occupational and recreational activities. Genetic variation in PEGS has been captured via whole genome sequencing on nearly 5,000 participants, making it ideal for exploring previously uncharacterized genetic components of unique exposures. We conducted overall, sex- and ancestry-stratified GWAS via logistic regression using exposures as outcomes while adjusting for age, age^2, sex, and PCs. We mapped variants at a nominal $p<1x10^{-3}$ to their nearest unique gene and conducted gene set enrichment on Reactome pathways. Finally, we assessed whether exposure-associated variants were more likely to occur in regulatory regions. The GWAS analysis returned genome-wide significant hits in 26 of 85 exposures (e.g. female-only analysis, "residence treated for pests regularly", $p=4.80x10^{-9}$, rs117035805; full analysis, "pet ownership", $p=9.37x10^{-9}$, rs11626746). Additionally, several neural genes had genome-wide suggestive hits in multiple exposures: RBFOX1 (8 exposures), NEDD4L (5), and CSMD1 (4). Gene pathways falling into the neuronal system and developmental biology Reactome categories were most often enriched for variants correlated with likelihood of exposure (20/85 and 28/85 exposures respectively), results that were confirmed via extensive permutation tests. Variants associated with all exposures were significantly more

likely to be found in gene promoters as defined by both gene proximity (3kb before TSS, OR=1.81, $p=3.8x10^{-64}$) and FANTOM CAGE annotations (OR=1.45, $p=9.6x10^{-3}$). These results highlight that neuronal and developmental processes and gene promoter regions contribute to the genetic component of the likelihood of environmental exposure. While much attention has been paid to gene-environment interactions, our study addresses a key, and often ignored, aspect: the impact of genetics on exposure likelihood, laying the groundwork for a complete picture of the intersection of individual genetics, environmental exposures, and health and disease outcomes.

Nature versus nurture of glucose homeostasis trajectories in children

Authors: I. Gamache[1], K. Fagbemi[1], N. Timpson[2], C. Greenwood[3], D. Manousaki[1]; [1]CHU Sainte-Justine, Montreal, QC, Canada, [2]Bristol Univ., Bristol, United Kingdom, [3]Jewish Gen. Hosp., Montreal, QC, Canada

Abstract:

Introduction: The etiology of dysglycemia in youth, encompassing type 2 diabetes (T2D) and its prodrome prediabetes, is multifactorial, involving genetics, environment, lifestyle, and their potential interactions. However, the understanding of their respective contribution to the early changes in glucose metabolism leading to T2D remains limited. **Methods:** We utilized the longitudinal ALSPAC study, which included 8,783 children of European descent. First, we assessed their genetic risk using 3 polygenic risk scores (PRS): two PRS for T2D in adults (Bell and al. (P+T) and Khera and al. (LDPRED)) and one for T2D in children (Srinivasan and al. (P+T)). Dysglycemia was evaluated using phenotypes such as fasting and postprandial glucose and insulin levels, and binary diagnoses of insulin resistance and prediabetes. Second, we performed univariable and multivariable regression analyses for each visit (for the latter we retained significant lifestyle and environmental variables in the univariate analysis). Third, we evaluated the interaction between PRS quintiles and environment/lifestyle at each visit and their trends over multiple visits using mixed models. **Results:** In ALSPAC, the prevalence of prediabetes was up to 25% in some visits. Compared to models using only environment/lifestyle factors, models including PRS had a better predictive performance. For instance, the AUROC for predicting prediabetes increased by 10%, rising from 0.68 [0.62-0.74] to 0.78 [0.72-0.83] by including the Bell et al. T2D PRS. In our interaction models, both cross-sectional and across visits, we observed that some environmental and lifestyle factors behave differently between individuals at low versus high genetic risk. For instance, consumption of sweet fruit juice and low birth weight accentuate risk of dysglycemia in individuals at high genetic risk. Meanwhile, high total energy intake and low income

attenuate the protective effect of being at lower genetic risk. **Conclusion:** Using data from multiple visits spanning pre-adolescence to young adulthood, we demonstrated that genetic liability to type 2 diabetes predicts dysglycemia in the ALSPAC study. The effect of genetic liability can be mitigated by environmental and lifestyle factors. Our results emphasize the importance of early lifestyle interventions to prevent dysglycemia in youth.

Decomposing sex-different phenotypic correlations in the UK Biobank into genetic and environmental components

Authors: A. Fritz[1,2,3], L. Darrous[2], K. Bønnelykke[4], A. Pedersen[1], Z. Kutalik[2]; [1]Danish Technical Univ., Kgs. Lyngby, Denmark, [2]Univ. of Lausanne, Lausanne, Switzerland, [3]COPSAC Copenhagen Prospective Studies of Childhood Asthma, Herlev-Gentofte, Denmark, [4]

Abstract:

Sexual dimorphism has been observed for disease prevalence and physical stature. The influence of genetic versus environmental contribution to these sex-differences, particularly how risk factors are linked to diseases (e.g., BMI ~ cardiovascular disease), remains unclear yet is crucial for understanding disease aetiology. We developed a method to calculate environmental correlation (rE) of trait pairs leveraging phenotypic and genetic correlations (rG) along with heritabilities. We analyzed 299 trait pairs with significant sex-different phenotypic correlations in the UK Biobank. Overall, we observed a predominance of environmental contribution to sex-different effects with 146 trait pairs (69%) exhibiting only sex-different rE, while 32 (15%) show sex-differences in both rE and rG, and only 6 (3%) are affected solely by sex-specific rG. Specifically, we detected differing rG and rE across blood biomarkers, e.g. C-reactive protein ~ BMI, showing sex-different rE (rE(men) = 0.14; 95% CI = [0.135, 0.143], rE(women) = 0.31; 95% CI = [0.307, 0.315]) and no sex-different genetic effects (rG(men) = 0.55; 95% CI = [0.47, 0.63], rG(women) = 0.64; 95% CI = [0.57, 0.70]) indicating that sex-different environmental effects are the main contributor to sex-different phenotypic correlations. Observations that women generally engage less in sports than men could explain differing environmental impacts. Additionally, glycated hemoglobin and LDL cholesterol only show rG in women (rG(women) = 0.16; 95% CI = [0.07, 0.24]) and not in men (rG(men) = 0.01; 95% CI = [-0.08, 0.11]). Whereas no rE is seen in women (rE(women) = 0.001; 95% CI = [-0.003, 0.005]), but in men (rE(men) = -0.16; 95% CI = [-0.158, -0.155]), indicating a sex-specific interplay of genetic and environmental components in diseases such as T2D, which are more prevalent in men. Some of these findings are supported by significant intra-trait rGs between men and women that differ

from 1, suggesting different genetic mechanisms across sexes. Significant sex-specific differences were also observed in trait pairs involving testosterone with SHBG, urate, waist-hip ratio, and triglycerides. In conclusion, while predominantly environmental components contribute to sex differences, this varies widely, indicating that disease-specific mechanisms may influence the outcomes of certain risk factors across sexes. Future research will focus on identifying missed indirect genetic effects mediated by parental nurturing and sibling effects and detecting confounders responsible for sex-different correlations in trait pairs.

Neanderthal introgression modifies the response to environmental stimuli in modern humans

Authors: C. Boye[1], Y-L. Lin[2], A. Findley[1], A. Alazizi[1], A. Dumaine[2], L. Barreiro[2], R. Pique-Regi[1], F. Luca[1]; [1]Wayne State Univ., Detroit, MI, [2]Univ. of Chicago, Chicago, IL

Abstract:

Approximately 2% of the modern Eurasian genome contains variants that humans obtained through interbreeding with Neanderthals. Neanderthal genetic variants may have been beneficial for modern humans (adaptive introgression), for example by providing a selective advantage in new environments. We previously showed that Neanderthal-introgressed variants regulate genes that are important for the transcriptional response to environmental challenges. As we cannot study Neanderthal-derived cells directly, we used a massively parallel reporter assay called Biallelic Targeted STARR-seq to study the gene regulatory function of Neanderthal-introgressed variants in human cells in response to three environmental stimuli: dexamethasone (a synthetic glucocorticoid, similar to the endogenous cortisol which is an anti-inflammatory hormone released in times of stress), vitamin D (sun exposure), and vitamin A/retinol (dietary compound). We designed a library of synthetic biallelic DNA constructs (targets), containing over 12,000 introgressed variants predicted to alter gene regulation, and their human counterpart. We transfected the constructs into human lymphoblastoid cells and measured their regulatory activity by quantifying self-transcribed enhancer sequences via RNA-seq. A total of 1,238 targets were differentially active in response to dexamethasone, 1,565 to vitamin D, and 3,105 to vitamin A. We observed differences in human and Neanderthal gene regulatory activity for 40 variants (FDR < 10%). The majority of these variants exhibited lower activity for the Neanderthal allele, similar to previous reports. For 21 of these variants, we observed differences between human and Neanderthal gene regulatory activity in the molecular response: 9 variants for dexamethasone, 9 for vitamin D, and 3 for vitamin A (FDR < 10%).

Out of 48 adaptively introgressed variants tested, we identified 2 that contribute to differences between the Neanderthal and modern human response to vitamin D (FDR < 10%). One of these variants is within a binding site for SPI1, which is a pioneer factor that opens chromatin, allowing the vitamin D receptor to bind to DNA. This suggests that some Neanderthal introgressed alleles modulating the response vitamin D exposure (such as sunlight) were beneficial to modern humans. Our results highlight the importance of the environmental context in understanding why humans retained some Neanderthal-introgressed alleles.

Assessing cellular contexts of type 2 diabetes-associated variants at scale

Authors: A. Tovar[1], K. Nishino[1], A. Etheridge[2], J. Rosen[2], K. Sun[2], S. Vadlamudi[2], Z. Chen[3], D. Dicorpo[4], J. Meigs[5], A. Manning[5], A. Kundaje[3], K. Lorenz[6], B. F. Voight[6], S. Schoenrock[2], M. Stitzel[7], R. Tewhey[8], K. Mohlke[2], J. Kitzman[1], S. C. J. Parker[1]; [1]Univ. of Michigan, Ann Arbor, MI, [2]The Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [3]Stanford Univ., Palo Alto, CA, [4]Boston Univ., Boston, MA, [5]Massachusetts Gen. Hosp., Boston, MA, [6]Univ. of Pennsylvania, Philadelphia, PA, [7]The Jackson Lab., Farmington, CT, [8]The Jackson Lab., Bar Harbor, ME

Abstract:

Type 2 diabetes (T2D) is a common metabolic disorder characterized by dysregulation of glucose metabolism. Genome-wide association studies have identified >660 loci associated with T2D. While much of this genetic risk is predicted to act through insulin-producing pancreatic islets, heritability is also distributed across regulatory regions active in other important metabolic tissues including adipose, liver, and skeletal muscle. Beyond specific tissues of action, there is mounting evidence that genetic effects on gene regulation are influenced by environmental context. High-throughput variant characterization assays such as massively parallel reporter assays (MPRAs) represent a tractable method to survey context-dependent regulatory activity of disease-associated variants at scale. Previously, we demonstrated widespread condition-specific allelic bias in a small MPRA library delivered to the LHCN-M2 human skeletal muscle myoblast cell line across four relevant states: (1) undifferentiated, or differentiated with (2) basal media, (3) AICAR to mimic exercise or (4) palmitate to induce insulin resistance. Specifically, 31.8% of tested variants (90/283) displayed allelic bias in only a single condition (FDR < 0.05) compared to just 9 variants across all conditions. We have since constructed an MPRA library to assess regulatory activities of >23k common variants in high linkage with 667 independent T2D association signals (R^2 > 0.7) and a set of ~1.5k TOPMed-contributed rare variants, comprising the largest single disease-associated MPRA to date. Given previous

evidence of enhancer-promoter regulatory specificity, we generated parallel versions of this library with several housekeeping and tissue-specific promoters. We delivered this library paired with either the potent synthetic promoter SCP1 or the skeletal muscle-specific MYBPC2 promoter to differentiated LHCN-M2 cells (n = 4 for both promoter contexts) and observed high interreplicate reproducibility (Pearson's r = 0.91-0.99). As an example, rs2037407 displayed stronger allelic bias when paired with the MYBPC2 promoter compared to SCP1 (p = 1.3 x 10^{-5} vs. 1.4 x 10^{-4}). This variant overlaps a skeletal muscle open chromatin peak and is in high LD (R^2 = 0.9-1) with the multi-ancestry T2D signal variant rs11114560 in non-European populations. Together, our work showcases the profound impact of cell type and context on gene regulation. We are currently completing analogous studies with the EndoC-βH3 human pancreatic beta cell line and performing integration with other genomic datasets to annotate all ~25k variants and uncover novel disease mechanisms.

Alternative polygenic score approaches aid in detecting genetic modification of the relationship between adiposity and cardiometabolic risk

Authors: K. Westerman, J. Gervis, A. Manning; Massachusetts Gen. Hosp., Boston, MA

Abstract:

Optimal use of genetics for precision medicine requires polygenic scores (PGS) that predict not just risk of disease, but also response to specific pharmaceutical or lifestyle interventions, detectable in observational datasets as PGS-by-environment interactions. A common practice is to use standard main effect PGS (mPGS) in these interaction tests, based on the strong statistical assumption that genetic main and interaction effects are proportional. We sought to test whether alternative scores built from variant-specific interaction effects (iPGS) or variance effects (vPGS) could better capture genetic modification of the association between adiposity and a series of cardiometabolic risk factors. The unrelated, European-ancestry subset of the UK Biobank (N = 323,802) was divided into training (50%), optimization (25%), and testing (25%) groups. For each of 10 serum cardiometabolic biomarkers (lipid, inflammatory, glycemic, and liver-related), genome-wide analyses in the training set separately tested genetic main effects, interaction effects with body mass index (BMI), and variance-quantitative trait locus effects. PRSice-2 was used to develop mPGS, iPGS, and vPGS from these summary statistics at a series of p-value thresholds. In the optimization set, one threshold was selected per PGS-biomarker pair to maximize the significance of the PGSxBMI product term. Finally, in the testing set, each optimized PGS was evaluated based on the estimate and significance of the same PGSxBMI interaction. Replication of specific PGS interactions was sought in the All of Us cohort (N = 122,678). For 7 of the 10 biomarkers, at least one

optimized PGS type reached significance ($p < 0.05 / 10$ biomarkers) for PGSxBMI interaction in the UKB testing set. Of these, the iPGS was the strongest for 4 biomarkers, compared to 3 for the mPGS. For example, the iPGS performed best for the log-transformed alanine aminotransferase (ALT; optimized p-value threshold of $5x10^{-8}$), with the positive BMI-ALT association being 72% larger in the highest versus the lowest iPGS decile ($p_{int} = 2.2x10^{-35}$). This effect modification was weaker using the mPGS ($p_{int} = 2.6x10^{-26}$). In All of Us, this interaction replicated in both a European-ancestry subset (82% larger association across extreme deciles; $p_{int} = 6.9x10^{-13}$) and in the full dataset ($p_{int} = 4.4x10^{-23}$). Our results describe PGS modifying the degree of adiposity-associated cardiometabolic risk, which could prioritize genetically susceptible individuals for weight loss interventions. They further suggest that alternative PGS types could enable more actionable, response-focused PGS for clinical and behavioral decision making.

Session 51: 3D Chromatin and Epigenomics

Location: Four Seasons Ballroom 4

Session Time: Thursday, November 7, 2024, 1:15 pm - 2:15 pm

Dissecting the genetic underpinnings of chromatin loops and their relationship to transcriptional regulation

Authors: E. Kharitonova[1], L. Lee[2], S. Mishra[2], M. Hu[2], Y. Li[1]; [1]UNC Chapel Hill, Chapel Hill, NC, [2]Cleveland Clinic, Cleveland, OH

Abstract:

BACKGROUND: 3D chromatin organization plays an essential role in regulating gene expression. For example, enhancers interact with their target genes via the formation of chromatin loops. Although these chromatin structures are tightly regulated, little is known about the genetic variants that influence them. Previous mammalian studies that examined the relationship between DNA sequence and 3D chromatin organization had limited findings due to small sample sizes (less than 20 individuals). Here, we conducted the first systematic chromatin loop quantitative trait loci (QTL) search in a sizeable number of mammals.
METHODS: We obtained liver tissue samples from 60 adult Collaborative Cross Recombinant Inbred Intercross (CC-RIX) mice, three female animals per litter from 20 distinct genetic lines. We generated high-resolution Hi-C, RNA-seq, and ATAC-seq data. We used linear mixed models to identify cis-QTL associated with gene expression, chromatin accessibility, and chromatin loop strength (eQTLs, caQTLs, and loopQTLs, respectively). Additionally, we conducted a mediation analysis using the DACT test to determine the contribution of caQTL and loopQTL to target gene expression.
RESULTS: We discovered 541 eQTLs, 7,800 caQTLs, and 1,507 loopQTLs that were significant with a Bonferroni adjusted p-value below 0.05. There was significant overlap between eQTLs and both caQTLs and loopQTLs, suggesting that genetic variants associated with gene expression tend to also be associated with chromatin accessibility and chromatin loop strength. Our mediation analysis found 191/1,162 QTLs where chromatin accessibility/loop mediates the effect of the variant on gene expression at an FDR level of 0.05. Only 4% of the target genes were shared between the significant chromatin accessibility and loop mediators, suggesting the presence of distinct chromatin regulatory pathways for different genes.
CONCLUSION: We have discovered novel associations between genetic variants and chromatin accessibility and loop strength, uncovering some of the genetic underpinnings

of 3D chromatin organization. These loopQTLs provide a new annotation for GWAS variants to help with generating mechanistic hypothesis for how SNPs affect complex human diseases. Results of our mediation analysis further support the claim that some genetic variants affect gene expression levels via the formation of chromatin loops. However, further functional validation experiments, such as CRISPR inhibition experiments, are needed to validate these findings. Overall, this study helps shed light on the functional mechanisms between chromatin spatial organization and genetic regulation.

Comprehensive Single-Nucleus Analysis of Genetic Regulation on Gene Expression and Chromatin Accessibility in Human Kidneys to understand of genetic basis of chronic kidney disease

Authors: E. Ha[1], H. Yuan[2], A. Abedini[1], F. Huang[2], K. Kloetzer[1], A. Sanchez Navarro[1], D. Hirohama[1], D. Kelley[2], K. Susztak[1]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Calico Life Sci., South San Francisco, CA

Abstract:

Chronic kidney disease (CKD) is a complex, multifactorial disorder affecting millions globally. Genome-wide association studies (GWAS) have identified numerous genetic loci associated with estimated glomerular filtration rate (eGFR), a key diagnostic marker for CKD. Given that most GWAS loci reside in non-coding regions, quantitative trait loci (QTL) analyses using bulk tissue have been employed to uncover their causal variants and functional impacts on gene expression and regulatory elements. However, these approaches may obscure cell-type-specific effects, limiting our understanding of the precise genetic regulation involved in CKD. In this study, we profiled transcriptomes and chromatin accessibility from >500,000 nuclei across 97 healthy and CKD human kidney tissue samples. We mapped expression and chromatin accessibility QTLs (eQTLs and caQTLs) across nine distinct kidney cell populations, achieving cell-type-specific resolution. Utilizing the 10X Multiome platform, we simultaneously captured gene expression and chromatin accessibility within the same nuclei, facilitating the linkage of regulatory elements to their target genes. We identified >7,000 eGenes from single-nucleus eQTL analysis. 70% were exclusive to a single cell type, suggesting cell-type-specific genetic regulation of gene expression. For instance, two variants rs11259952 and rs34882080 within an eGFR GWAS loci were found to specifically affect the expression of the genes *WHAMM* and *UMOD* in the proximal tubule (PT) and thick ascending limb (TAL) cells, respectively. The caQTL analysis identified >100,000 caPeaks. To integrate these multimodal datasets, we leveraged the simultaneous RNA-seq and ATAC-seq profiles to

infer genetic regulatory networks for the identified eGenes. In PT cells, we mapped 5,955 regulatory elements to 936 eGenes with high confidence. Further investigation combining eQTLs, caQTLs, and GWAS signals within these regulatory networks is essential to uncover the genetic foundations of kidney function. This approach aims to identify potential therapeutic targets and pathways for CKD treatment.

Genetic and Epigenetic Insights into the Aging of the Human Retina

Authors: T-Y. Lin[1,2], J. Advani[1], M. English[1], S. Mehrotra[3,4,5], P. Mehta[3,4,5], Y. Luo[3,4,5], D. Ferrington[6], A. Segre[3,4,5], A. Swaroop[1]; [1]Neurobiology, Neurodegeneration and Repair Lab., Natl. Eye Inst., NIH, Bethesda, MD, [2]Doctoral Degree Program of Translational Med., Natl. Yang Ming Chiao Tung Univ. and Academia Sinica, Taipei, Taiwan, [3]Ocular Genomics Inst., Dept. of Ophthalmology, Massachusetts Eye and Ear, Boston, MA, [4]Dept. of Ophthalmology, Harvard Med. Sch., Boston, MA, [5]Broad Inst. of Harvard and MIT, Cambridge, MA, [6]Dept. of Ophthalmology and Visual NeuroSci.s, Univ. of Minnesota, Minneapolis, MN

Abstract:

We have developed a framework to longitudinally explore the molecular basis of human physiological aging, quantifying advanced age's and genetics' relative contributions to gene expression, regulation, and methylation in the retina. We performed genotyping and RNA-sequencing on 926 postmortem human retina samples spanning nine decades of life, with nearly half (49.7%) consisting of non-diseased controls and half of different grades of age-related macular degeneration (AMD). 487 significant age differentially expressed genes (DEG) in discovery (222 control samples) were highly replicable for protein-coding (ρ;;= 0.88, p = <1E-05) and lincRNA (ρ;;= 0.84, p = <1E-05) in replication (513 control and different AMD grade samples) studies. A combined analysis (full study) identified 2,205 significant aging-associated genes; these include *RFX4, FCGR2B, NR4A2, WLS, LYPD1*, and *IL20RA*, which have been linked to neurogenerative diseases. Genes upregulated with age were enriched in antigen processing and presentation pathways, whereas downregulated genes were augmented in oxygen transport and haptoglobin binding (FDR<0.05). To assess genetic regulation of expression in retina across a wide age span, we computed expression quantitative trait loci (eQTL) correcting for age, sex, genotype PCs, hidden covariates, and AMD grade and found that the eQTLs are highly replicable between discovery and replication studies (ρ;;= 0.86, p = <1E-05); 3,623/10,593 genes are unique to our full study compared to previously reported eQTLs from an older sample set. The unique genes were significantly enriched for targets of transcription factor *TWIST1*, *GPCR* ligand, peptide hormone, and cysteinyl leukotriene receptor activity. Finally, we computed age-

interaction eQTLs (ieQTLs) by adding an interaction term between genotype age to the eQTL model and identified 11 age-ieQTLs (FDR<0.25). These ieQTLs regulate genes previously associated with retinal health, age-related ocular disorders, neurodegenerative disease, and mediator of AMD-risk factors (*IFI6, LINC02145, FRS3, THUMPD3, ZNF581, RPE, KHDC4, CASKIN1, CABP5, RGS19, VEPH1*). Notably, 5/11 age-ieQTL were not found with the eQTL model. Our study highlights critical genes and pathways that govern age-related changes in gene expression and genetic regulation in the retina. Changes in genome-wide DNA methylation with age will be presented.

Single-cell genomics, QTLs, and regulatory networks for 388 human brains

Authors: M. Jensen[1], P. Emani[1], J. Liu[1], D. Clarke[1], J. Warrell[1], C. Gupta[2], R. Meng[1], C. Lee[3], S. Xu[3], C. Dursun[1], S. Lou[1], Y. Chen[1], T. Galeev[1], A. Hwang[3], Y. Li[1], Z. Chu[1], P. Ni[1], X. Zhou[1], PsychENCODE Consortium, D. Lee[4], M. Gandal[5], E. Lein[6], P. Roussos[4], N. Sestan[1], Z. Weng[7], K. White[8], H. Won[9], M. Girgenti[1], J. Zhang[3], D. Wang[10], D. Geschwind[11], M. Gerstein[1]; [1]Yale Univ., New Haven, CT, [2]Univ. of Wisconsin, Madison, WI, [3]Univ. of California, Irvine, Irvine, CA, [4]Icahn Sch. of Med. at Mount Sinai, New York, NY, [5]UCLA, Los Angeles, CA, [6]Allen Inst Brain Sci., Seattle, WA, [7]Univ. of Massachusetts Med. Sch., Worcester, MA, [8]Natl. Univ. of Singapore, Singapore, Singapore, [9]Univ. of North Carolina, Chapel Hill, NC, [10]Univ. of Wisconsin - Madison, Madison, WI, [11]Univ California Los Angeles, Santa Monica, CA

Abstract:

Single-cell genomics approaches represent a powerful tool for understanding the role of genomic variants towards gene expression and regulation, especially for heterogeneous tissues with many diverse cell types such as the brain. Using these technologies, we can refine our understanding of how variants and gene regulation affect brain phenotypes, including neuropsychiatric disorders such as schizophrenia, autism, and bipolar disorder. However, population-scale cohorts with a wide range of disease phenotypes and traits are needed to make statistically meaningful associations between variants, regulatory elements, and expression, and to develop comprehensive models of brain gene regulation at the single-cell level. Using a harmonized cell-typing scheme, we uniformly processed genotype, expression, and chromatin accessibility sequencing data in >2.8M single nuclei from the pre-frontal cortex of 388 individuals with brain-related disorders and controls. We identified population-level variation in expression and chromatin for 28 cell types across multiple gene families, including neurotransmitters and drug targets. We also found >550K cell-type-specific cis-regulatory elements and >1.4M single-cell expression QTLs, applying novel Bayesian methods to identify QTLs in rare cell types. By including ~300 additional

samples from external cohorts, we further increased our power to discover larger sets of single-cell eQTLs for both SNVs and structural variants (SVs), and also found limited sets of chromatin, isoform, and cell fraction QTLs. We next built cell-type regulatory networks and cell-to-cell communications networks from the expression, open chromatin, and QTL data. These networks detail extensive cell-type-specific regulatory changes across traits, including enrichment of TF motifs during aging, and disorders, such as altered WNT and FGF signaling in schizophrenia and bipolar disorder. We finally constructed an integrative deep-learning model to accurately impute single-cell expression and simulate network perturbations, which explained a higher proportion of phenotype variance than models built with bulk datasets or polygenic risk scores. The model prioritized ~250 known and novel disease-risk genes and drug targets along with their associated cell types, such as *MEF2A* and *LINGO2* in excitatory neurons for bipolar disorder. Overall, our population-scale single-cell resource for the human brain can help facilitate discovery of precision medicine approaches and variant prioritization for neuropsychiatric disorders.

Session 52: Computational Methods for Causal Variant Prioritization

Location: Mile High Ballroom 2&3

Session Time: Thursday, November 7, 2024, 1:15 pm - 2:15 pm

Footprint quantitative trait loci (fpQTLs) reveal non-coding causal variants associated with transcription factor binding for liver traits

Authors: M. Dudek[1,2], B. M. Wenz[1], C. D. Brown[1], S. F. Grant[2,1], L. Almasy[2,1]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Children's Hosp. of Philadelphia, Philadelphia, PA

Abstract:

Genome-wide association studies (GWAS) have revealed multiple loci for chronic liver disease, a trait area representing a major healthcare burden worldwide, most commonly due to non-alcoholic fatty liver disease. Given that GWAS signals are enriched in cis-regulatory elements harboring binding motifs of transcription factors (TFs), we sought to determine variants associated with TF-binding, or 'footprint quantitative trait loci' (fpQTLs) using the largest sample size for this approach reported to date, in ATAC-seq data generated from liver. Specifically, TF binding can be detected in ATAC-seq experiments where bound TFs block the transposase Tn5, leaving a pattern of relatively depleted Tn5 insertions known as a 'footprint'. We used computational tools to scan ATAC-seq reads from 170 human liver samples with genotyped variants and calculated 'footprint scores' to estimate TF-binding likelihood at these variants across samples. Regressing footprint score onto genotype, we observed 693 fpQTLs significantly associated with footprint-inferred TF binding at 5% FDR. Given that variant footprint scores are not affected by linkage disequilibrium, fpQTLs can aid GWAS fine-mapping by precisely locating TF activity within broad disease-associated loci which typically include several candidate variants.
Liver fpQTLs were strongly enriched across ChIP-seq peaks for known liver TFs (SOX6 odds ratio (OR)=11.1, $P=2.7x10^{-3}$; FOXA1 OR=2.6, $P=3.7x10^{-5}$; FOXA2 OR=2.6, $P=1.5x10^{-5}$) as well as for TF binding motifs, liver expression QTLs (eQTLs, OR=4.0, $P=2.5x10^{-20}$), and GWAS loci for liver-related traits (LDL lipids OR=3.4, P=0.031). Notably, the measured effect of a fpQTL on TF binding was highly concordant with its effect on an underlying sequence motif in 91% of overlaps. When applied to the established lipid-associated *SORT1* locus, our fpQTL method re-discovered the experimentally validated causal variant (rs12740374), and implicated several other potentially causal variants at additional liver trait-associated loci. We conclude that our findings reveal novel insights in the genetic etiology of liver-related traits. We show that fpQTLs can be leveraged both to reveal causal GWAS-implicated non-

coding variants through studying the role of TF binding site disruption and to provide functional insights that can yield novel treatments for common diseases.

A new variant-to-disease score prioritizing causal variants in GWAS

Authors: S. Cheng, S. Gazal; USC, Los Angeles, CA

Abstract:

The genetic architecture of human complex diseases is largely driven by non-coding variants. It is thus critical to predict the function of these variants to prioritize causal common variants in genome-wide association studies (GWAS). However, existing methods for prioritizing non-coding variants (e.g., CADD) have demonstrated only modest utility in complex diseases. Translating GWAS into biological insights requires more powerful methods for identifying non-coding variants impacting genome function and disease risk. Here, we propose to estimate the expected disease effect sizes of all variants based on a hundred of functional annotations and GWAS results of independent traits, and to use these effects as variant-to-disease (V2D) scores to prioritize causal variants in GWAS. Specifically, we propose a machine learning framework that 1) transform GWAS marginal effect sizes (i.e. including tagging effects due to linkage disequilibrium) into true effect sizes by leveraging fine-mapping algorithm on 15 independent polygenic traits of the UK Biobank, 2) model these effects using diverse functions (e.g., linear model, neural networks, XG-boost) on a hundred of annotations, 3) predict disease effects of each variant using a leave-one-chromosome-out approach.
First, we benchmarked our approach using simulations, comparison of heritability enrichment using a linear function and S-LDSC outputs (gold standard), and linear trees (which provide interpretable results). Linear trees notably identified non-linear relationships between annotations related to distal regulation (enhancers) and evolution (primate-constraint), which were underestimated by S-LDSC.
Second, we created a V2D score for every variant by leveraging neural networks. V2D outperformed estimates from a linear model and from CADD v1.7 in identifying bases validated by MPRA and in heritability analyses on the 79 GWASs (not genetically correlated to the 15 traits).
Finally, we illustrated the benefits of V2D in prioritizing disease variants. By leveraging V2D as a prior for fine-mapping studies, we were able to nominate known causal variants (e.g., the FTO-rs1421085 in BMI, and the SMAD3-rs17293632 in asthma), and to identify new disease causal mechanisms.In conclusion, we have developed a machine learning framework leveraging GWAS and a hundred functional annotations to prioritize non-coding variants in GWAS.

Robust fine-mapping in the presence of LD mismatch

Authors: W. Zhang[1], T. Lu[2], J. Dupuis[3], G. Lettre[4,1]; [1]Montreal Heart Inst., Montreal, QC, Canada, [2]Univ. of Wisconsin-Madison, Madison, WI, [3]McGill Univ., Montreal, QC, Canada, [4]Université de Montréal, Montreal, QC, Canada

Abstract:

Fine-mapping methods based on GWAS summary statistics and linkage disequilibrium (LD) information are widely used to identify potential causal variants and prioritize therapeutic targets. However, most GWAS, especially meta-analyses, do not provide in-sample LD information due to logistical and ethical concerns. LD mismatch between the external LD reference panel and the GWAS population is common and can lead to compromised accuracy of fine-mapping.

We developed RSparsePro, a novel errors-in-variables model with an efficient variational inference algorithm, to simultaneously perform LD mismatch detection and robust fine-mapping. In UK Biobank-based simulations, in the absence of LD mismatch, RSparsePro had comparable performance as state-of-the-art fine-mapping methods, such as SuSiE and SparsePro. The metric for detecting LD mismatch was also well-calibrated. With increased LD mismatch, RSparsePro achieved improved accuracy in identifying mismatched variants compared to methods designed for LD mismatch detection, such as DENTIST and SLALOM, and significantly outperformed existing fine-mapping methods in identifying causal variants.

We applied RSparsePro to fine-map the largest GWAS meta-analyses for LDL levels conducted by GLGC. We constructed ancestry-specific LD reference panels using individual-level data from the UK Biobank European, South Asian, East Asian, and African populations. In 496 genome-wide significant loci, we first performed fine-mapping using ancestry-matched LD reference panels for each population, where the degree of LD mismatch was expected to be modest. As a result, variants in the 95% credible sets identified by RSparsePro were 1.89-fold (95% CI: 1.32-2.71) more likely to have protein-altering effects than those identified by SuSiE. Next, for each of the non-European ancestry populations, we performed fine-mapping using the European LD reference panel, which artificially induces severe LD mismatch. Importantly, 62.4%, 51.9%, and 68.4% of the credible sets obtained in South Asian, East Asian, and African populations by RSparsePro using the incorrect, European LD reference panels were still consistent with those based on ancestry-matched LD reference panels. In contrast, when SuSiE was applied to the South Asian, East Asian, and African populations with the European LD reference panels, the consistency rates were 42.3%, 22.5%, and 16.2%, respectively.

In summary, RSparsePro can reliably identify causal variants in the presence of LD mismatch and will greatly expand the applicability of fine-mapping analyses, especially in increasingly larger GWAS involving multiple cohorts and diverse populations.

Do deep genome language models help pinpoint causal variants in statistically fine-mapped loci?

Authors: M. Sweeney, H. Kang; Univ Michigan, Ann Arbor, Ann Arbor, MI

Abstract:

Current deep learning models to understand genomic regulation have shown potential in predicting gene expression levels from genotypes, offering new opportunities for understanding the genetic basis of complex traits. However, recent studies suggest that these genome language models, while useful, do not yet match the prediction accuracy of linear regression models trained from population-scale transcriptome-wide association studies, such as PrediXcan.
In this study, we aim to evaluate whether current genome language models and other deep learning approaches can help pinpoint causal variants in statistically fine-mapped molecular quantitative trait loci (QTLs). The precision of statistical fine-mapping is often limited by linkage disequilibrium, where highly correlated variants make it difficult to confidently identify a single causal variant. We hypothesized that existing deep learning models may help prioritize likely causal variants among statistically indistinguishable genetic variants due to linkage disequilibrium by leveraging the functional prediction accounting for their sequence contexts.
We evaluated three deep learning and/or genome language models: Sei, Enformer, and Enformer's successor, Borzoi, using simulated and real eQTL datasets. Simulation studies quantified how the accuracy of these models improves the power to pinpoint causal variants among credible sets in fine-mapped loci. We also compared the benefits of these models in prioritizing replicated variants across multiple studies, including GTEx, MAGE, and TOPMed. When focusing on fine-mapped loci with a credible set size of 2, where only one variant was replicated in an independent study, using posterior inclusion probabilities (PIPs) to prioritize causal variants achieved 74.8% accuracy. Sei-based prediction resulted in 71.3% accuracy, while Enformer and Borzoi achieved 75.6% and 78.3%, respectively. Furthermore, ensemble methods such as multiple logistic regression models which provide combined PIP and Borzoi-based predictions can further improve accuracy, up to 79.7%. The benefit of deep learning models was particularly evident when statistical fine-mapping yielded lower posterior probability for the top variant (PIP <0.75).

Our findings demonstrate that the most recent deep genome language models can outperform current best practices in statistical fine-mapping for pinpointing causal variants. As the accuracy of these models continues to improve, we expect their benefits in aiding causal variant identification to grow, ultimately enhancing our understanding of the genetic architecture of complex traits and diseases.

Session 53: Dysfunction at the Powerhouse: Molecules, Models, and Organisms

Location: Room 405

Session Time: Thursday, November 7, 2024, 1:15 pm - 2:15 pm

Investigating the role of seryl-tRNA synthetase (*SARS2*) in mitochondrial biology and human recessive disease

Authors: C. Del Greco, A. Antonellis; Univ. of Michigan, Ann Arbor, MI

Abstract:

The human nuclear genome encodes 17 mitochondrial aminoacyl-tRNA synthetases (ARSs) responsible for charging tRNA with amino acids in the mitochondria. This process is required for translation of the 13 mitochondrial-encoded proteins. All 17 mitochondrial ARSs have been implicated in recessive diseases with a broad range of clinical phenotypes. The gene encoding mitochondrial seryl-tRNA synthetase (*SARS2*) has been implicated in two distinct disease phenotypes: (**1**) progressive spastic paresis; and (**2**) HUPRA syndrome (hyperuricemia, pulmonary hypertension, renal failure in infancy, and alkalosis). Interestingly, the clinical heterogeneity of *SARS2*-related disease is currently unexplained and there are few functional studies on pathogenic *SARS2* variants. To address these gaps in knowledge, we are interrogating the molecular effects of *SARS2* dysfunction in a human cell culture model. Toward this, we generated a Hap1 cell line that contains: (**a**) a randomly integrated, doxycycline-inducible wild-type copy of *SARS2*; and (**b**) a 2kb deletion at *SARS2* that ablates endogenous gene function. This cell line is maintained in low-dose doxycycline and then grown in the absence of doxycycline to assess the effects of a loss of *SARS2* expression over time. RNA-sequencing data from cells grown in the absence of doxycycline show upregulation of cellular stress response-associated genes compared to cells grown in the presence of doxycycline, including genes found in the integrated stress response (*e.g.*, *GDF15* and *CHAC1*) and ER stress (*e.g.*, *ATF5* and *CHOP*), and pathway analysis indicates upregulation of stress responses consistent with mitochondrial stress (*e.g.*, stress associated with the HRI kinase). In addition, seahorse mitochondrial stress tests of cells grown with and without doxycycline show significant decreases in mitochondrial oxygen consumption upon loss of *SARS2*; loss of *SARS2* expression for 6 days resulted in decreased basal respiration (81% decrease) and maximal respiration (77% decrease) compared to controls. In addition, when these cells were transduced with lentiviruses containing known pathogenic *SARS2* variants under the control of a constitutively expressed promoter and grown in the absence of doxycycline, we find that

we can quantify differences in oxygen consumption dependent on which variant was transduced. Additional studies are being performed to assess the effects of all reported pathogenic *SARS2* variants on mitochondrial function. In sum, our studies provide insight into the role of *SARS2* in cells and the differential effects of pathogenic *SARS2* variants on protein function, which will inform genotype-phenotype correlations.

COXFA4 Dysfunction Leads to ODC Dysregulation: A Link to Mitochondrial Disease Mechanism

Authors: J. Marquez[1,2], S. Viviano[3], E. Beckman[2], J. Thies[2], C. T. Lam[1], E. Deniz[3], E. M. Shelkowitz[1,2]; [1]Univ. of Washington, Seattle, WA, [2]Seattle Children's Hosp., Seattle, WA, [3]Yale Univ. Sch. of Med., New Haven, CT

Abstract:

COXFA4 deficiency has been implicated as a cause of mitochondrial disease. Individuals with putatively pathogenic variants in *COXFA4* experience cardiomyopathy, neurodevelopmental differences, microcephaly, and lactic acidosis. We identified a family with 2 siblings with a history of developmental delays. One sibling had neonatal cardiomyopathy, while the other had atypical MRI findings. A sibling of theirs had previously died due to neonatal cardiomyopathy. This prompted evaluation via exome sequencing that identified that both living siblings were homozygous for a deletion of the entire coding region of *OCXFA4*. We reviewed previously reported cases of *COXFA4*-related mitochondrial disease along with these unpublished cases. Our findings further support the previously described phenotypic spectrum of this form of mitochondrial disease including early onset cardiomyopathy. The molecular disease mechanisms underlying *COXFA4* deficiency and its impact on complex IV are not fully understood. To this end, we believed a vertebrate model of *COXFA4* dysfunction was crucial for furthering our understanding of the mechanism by which loss of *COXFA4* leads to disease. In a *Xenopus tropicalis* model, *coxfa4* knockout led to multiple phenotypes that recapitulated patient disease. These included impaired craniofacial development and decreased cardiac function as demonstrated by craniofacial cartilage staining and optical coherence tomography imaging of cardiac contractility respectively. In addition, we confirmed that mitochondrial complex IV activity was reduced in our model through in-gel activity assays. To understand potential developmental pathways that may contribute to observed phenotypes, we analyzed RNA-sequencing data from a *COXFA4* knockout induced pluripotent stem cell line. Through this approach, we identified deficits in the ornithine decarboxylase (ODC) pathway that governs the biosynthesis of polyamines. ODC signaling

is characterized by a rapid turnover rate of the ODC protein as a dynamic pathway in response to stimuli. Modulation of ODC signaling through polyamine supplementation in our model appeared to improve cardiac function based on measures of cardiac contractility. In summary, we demonstrated that depletion of *coxfa4* leads to both complex IV dysfunction and ODC pathway dysregulation. Furthermore, this work suggests that downstream polyamine supplementation may improve COXFA4-associated cardiac disease and supports further investigation of this as a therapeutic modality.

A multiomic approach to elucidate muscle-specific pathogenesis of SUCLA2-deficient mitochondrial myopathy

Authors: M. Lancaster[1], C. Matias[1], A. Law[1], C. Ferreira[2], B. Jeffrey[1], B. Graham[1]; [1]Indiana Univ. Sch. of Med., Indianapolis, IN, [2]Purdue Univ., Lafayette, IN

Abstract:

Mitochondrial diseases are a prevalent cause of multisystem disorders. Among these, early onset mitochondrial encephalomyopathy is linked to biallelic pathogenic variants in subunits of the TCA cycle enzyme succinyl-CoA synthetase (SCS). This condition typically presents in early childhood with symptoms such as intellectual disability, hypotonia, progressive muscle weakness, growth deficits, and failure to thrive; however, the pathogenesis of this disease remains unclear. To investigate the mechanisms in the energy-dependent and highly affected skeletal muscle, we developed a muscle-specific mouse model of SCS-deficient mitochondrial myopathy. Interestingly, though model validation confirmed nearly identical patterns of SCS loss, hindlimb muscles soleus (SOL) and extensor digitorum longus (EDL) exhibit stark differences in response to SCS deficiency. The EDL appears largely unaffected, whereas the SOL, inherently rich in type-I oxidative fibers, exhibits severe myopathic phenotypes, including 40% reduction of specific tetanic force, slowed contractile kinetics, and fatigue resistance. To elucidate the molecular foundations of these distinct responses, we have employed a multiomic approach analyzing transcriptomic, proteomic, and metabolomic differences between SOL and EDL muscles. First, we performed spatial metabolomics on muscle cross sections using desorption electrospray ionization (DESI) mass spectrometry. This mass spectrometry imaging (MSI) technique maps the spatial distribution of metabolomic and lipidomic profiles across tissue sections. Our preliminary data indicate that only 50 metabolites showed genotype-specific differential abundance in sections containing EDL and tibialis anterior muscles, while over 400 differentially abundant metabolites were observed in the triceps surae (TS, generally comprising SOL, plantaris, and gastrocnemius

muscles). Initial pathway analysis integrates the metabolomic results in the TS and 5,773 differentially expressed transcripts within the SOL. Among the most highly enriched pathways include the TCA cycle, PPAR signaling, and muscle contraction, among other potentially noteworthy pathways. Ongoing studies aim to integrate high-resolution SOL and EDL MSI with transcriptomic and proteomic data to identify *fiber-type* specific pathway perturbations and unravel the divergent mechanisms of SCS deficiency in distinct cellular contexts. This context-specific multiomic approach will yield crucial insights, having significant impact as we advance towards development of therapeutic interventions for mitochondrial dysfunction.

Identification and targeting of ABHD18 as a strategy to alleviate TAZ mutant phenotypes

Authors: S. Masud[1], P. Mero[2], K. Brown[2], V. Saba Echezarreta[1], J. Wei[1], D. Thomson Taylor[1], N. Mikolajewicz[3], A. Granda Farias[1], L. McDonald[2], O. Sneizek Carney[4], M. Niphakis[5], K. Chan[2], O. Sizova[2], A. Habsid[2], L. Nedyalkova[6], A. Shaw[7], G. Tan[6], S. Mital[2], H. Vernon[8], M. Billmann[7], B. Andrews[6], C. Myers[9], I. Scott[2], C. Boone[6], J. Moffat[3]; [1]Univ. of Toronto, Hosp. for Sick Children, Toronto, ON, Canada, [2]Hosp. for Sick Children, Toronto, ON, Canada, [3]The Hosp. for Sick Children, Toronto, ON, Canada, [4]John Hopkins Univ. Sch. of Med., Baltimore, MD, [5]Lundbeck La Jolla Reseach Ctr., San Diego, CA, [6]Univ. of Toronto, Toronto, ON, Canada, [7]Univ. of Bonn, Bonn, Germany, [8]Dept. of Genetic Med., Johns Hopkins Univ. Sch. of Med., Baltimore, MD, [9]Univ. of Minnesota, Minneapolis, MN

Abstract:

Mitochondrial dysfunction has pleiotropic effects and can contribute to the pathophysiology of cardiac and neurodegenerative diseases. Understanding the complex networks of genetic interactions (GIs) that regulate mitochondrial function is critical to understanding disease mechanisms and therapeutic opportunities. The pathogenesis of Barth Syndrome (BTHS) is driven by mutations in TAFAZZIN, encoded by the gene *TAZ*, and is required for the maturation of cardiolipin (CL), the signature phospholipid of the mitochondria. BTHS is characterized by mutations along the entire open reading frame of *TAZ*, with drastically different clinical manifestations and no targeted therapeutics. We have leveraged our expertise in functional genomics to generate the first global loss-of-function GI profile for *TAZ* in human cells using optimized genome-wide CRISPR screens/ gene-trap screening platforms and have identified genetic suppressors/modifiers of mutant *TAZ*. We have discovered and characterized that perturbation of ABHD18 (C4orf29), an uncharacterized abhydrolase, not only suppresses the cellular phenotypes associated

with *TAZ* mutations, but is the missing lipase in human cells, the functional homolog of yeast CLD1 which functions to convert nascent CL to monolysocardiolipin (MLCL), the substrate for tafazzin. We have characterized the mode of action of ABHD18 in both patient and animal models and have developed a novel covalent inhibitor providing an opportunity for the first targeted therapeutic for BTHS and other CL-related diseases. Furthermore, our GI profile has generated a list of novel genetic modifiers that may explain the phenotypic pleiotropy caused by mutation in *TAZ*, providing insight into buffering mechanisms for mitochondrial fitness and lipid homeostasis. We present a multi-layered approach including functional genomics, proteomics, metabolomics, and chemical biology to identify and validate a new target for a rare genetic disease.

Session 54: Expanding the Table: Considerations for Inclusion in Genetics and Genomics

Location: Room 505

Session Time: Thursday, November 7, 2024, 1:15 pm - 2:15 pm

Equity-focused implementation illuminates diverse perspectives in rare disease research

Authors: M. Wojcik[1], E. Martinez[2], G. VanNoy[2], M. O'Leary[2], J. Serrano[2], S. Abouhala[2], B. Mangilog[2], G. Shah[2], I. Holm[1], Y. Fraiman[1], H. Rehm[3], A. O'Donnell-Luria[1]; [1]Boston Children's Hosp., Boston, MA, [2]Broad Inst. of MIT and Harvard, Cambridge, MA, [3]Massachusetts Gen. Hosp., Boston, MA

Abstract:

Background: Engagement of diverse populations in rare disease genomic research has been challenging, with particular underrepresentation of minoritized racial and ethnic groups in large genomic sequencing studies. Other social informants of health such as primary language, household resources, and location of primary residence may also impede access. We therefore report on our efforts to improve diverse representation in the Rare Genomes Project (RGP) and participant-reported outcomes related to the process and context of genomic research.
Methods: We implemented a multi-faceted intervention over two years to support recruitment of underserved populations for rare disease research. This included expanded outreach to community providers, non-English language support, proactive and flexible participant contact and engagement, and use of mobile phlebotomy. Participants in RGP were offered a survey upon enrollment to assess values and priorities as well as stress measured by the Perceived Stress Score.
Results: Since the launch of our equity initiative, 111/152 (73%) participants previously underrepresented in RGP were successfully enrolled in our study, with an additional 28/152 (18%) still in the process of enrollment. Of the 111 enrolled, 48% reported Hispanic/Latino ethnicity, 33% reported non-white race, 20% had a primary language other than English, 25% reported household income under the federal poverty level, 21% had an education level of high school or less, and 23% live in rural areas. Most (85/111, 77%) have submitted samples for genome sequencing (GS), with GS analysis completed for 56/85 (66%) and identifying a diagnosis or strong candidate in 15/56 (27%).
Perceived importance of a genetic diagnosis significantly differed between the underserved cohort and the remainder of the RGP participants, with higher perceived importance in the

underserved cohort (median of 5 vs 4.5 on 5 point Likert scale, p = 0.03). Perceived stress category also varied between those from the underrepresented cohort compared to the remainder of the RGP cohort: 9/61 (15%) high, 41/61 (67%) moderate, 11/61(18%) low versus 6/94 (6%) high, 81/94 (86%) moderate, 7/94 (7%) low, p = 0.023.
Conclusions: Our two-year equity-focused initiative was able to successfully enroll over 100 participants previously under-represented in our rare disease genomic sequencing study, with these participants reporting both higher stress and higher perceived importance of a genetic diagnosis upon entry to the study. Future efforts will evaluate whether identification of a diagnosis impacts both perceived importance and stress.

Use of exclusion criteria to select critically ill newborns for rapid genome sequencing captures precise genetic diagnoses missed by use of conventional inclusion criteria

Authors: T. Wenger[1], A. Keefe[1], L. Kruidenier[2], A. Scott[3], M. Sikes[4], J. Love-Nichols[1], K. Anderson[1], O. Sommers[1], J-H. Yu[5], K. MacDuffie[6], K. Retterer[7], K. McWalter[8], D. Doherty[9], D. Veenstra[9], C. Davis[1], K. Shively[1], H. Gildersleeve[9], J. Juusola[10], K. Buckingham[1], J. Chong[1], P. Kruszka[11], K. Dipple[12], M. Bamshad[1]; [1]Univ. of Washington, Seattle, WA, [2]Univ. of Texas Hlth.Sci., Houston, TX, [3]Seattle Children's Hosp., Seattle, WA, [4]Seattle Children s, Seattle, WA, [5]Univ. of Washington Sch. of Med., Seattle, WA, [6]Seattle Children's Res. Inst., Seattle, WA, [7]Geisinger, Danville, PA, [8]GeneDx, Honolulu, HI, [9]Univ of Washington, Seattle, WA, [10]GeneDx, Gaithersburg, MD, [11]GeneDx, Alexandrira, VA, [12]Seattle Children s Hosp. and Univ. of Wash, Seattle, WA

Abstract:

SeqFirst-neo is a project to develop and test approaches to centering equity for a precise genetic diagnosis (PrGD) at the initial point of care of infants with a critical illness using broad exclusion criteria to identify infants eligible for rapid whole genome sequencing (rWGS). Infants admitted to the NICU at Seattle Children’s Hospital with clinical findings were not fully explained by a previously known PrGD, prematurity, infection or trauma were offered rWGS. This resulted in significantly increased access to a PrGD, more equitable access to a PrGD, and fewer missed diagnoses compared to use of a conventional workflow and stratified testing. Herein we sought to explore the clinical characteristics of the 28 infants with a PrGD who were missed by the conventional workflow but diagnosed by rWGS. Medical records for each infant were evaluated to identify one or more common themes underlying a missed diagnosis. Themes included provider assessment that an infant’s clinical findings were fully explained by: 1) an “isolated” birth defect (11/28; 39%);

2) complications of prematurity (11/28; 39%); 3) "self-limited" non-genetic medical problem (5/28; 17%); 4) abnormal lab values due to critical illness (4/28; 14%)I 5) postsurgical complications (2/28; 7%); 6) infection (2/28; 7%); 7) trauma (2/28; 7%); 8) limitations of nongenetic testing (2/28; 7%); and 9) an incorrect diagnosis (1/28; 4%). Illustrative vignettes were selected to highlight themes: *TTC7A*-related disorder in infant thought to have postsurgical complications of isolated colonic atresia repair (themes 1 and 5); cardiofaciocutaneous syndrome in infant transferred briefly from birth hospital for VCUG for “isolated” hydronephrosis and feeding problems attributed to 33-week prematurity (themes 1 and 2); Hemophilia A and Glanzmann thrombasthenia in a late-preterm infant with GI bleeding in infant with enterovirus-related liver failure, thrombocytopenia and meningoencephalitis (themes 2, 4, 6, 8); hemophilia B in late-preterm infant with subgaleal bleed and hypovolemic shock after traumatic delivery with 6 hours of pushing with face presentation (themes 2, 4, 6, 7); hereditary hemorrhagic telangiectasia in a premature infant with pulmonary hypertension and bronchopulmonary dysplasia (themes 2, 4, 7, 8); Rubenstein-Taybi syndrome in infant with a clinical diagnosis of CHARGE syndrome (theme 9). In 21/28 (75%) of these cases a PrGD led to changes to management. These results demonstrate that eligibility for rWGS based on broad exclusion criteria can capture missed PrGD particularly in premature infants and those with isolated birth defects with low clinical suspicion for a genetic condition.

Reprogenomics, Ethics and Inclusivity: Perspectives from Sex and Gender Diverse Communities

Authors: M. Michie[1], H. Custer[1], C. Collart[2], B. Gillani[3], S. E. Moore[3], R. Ponsaran[1], S. Rubeck[1], R. Farrell[2]; [1]Case Western Reserve Univ. Sch. of Med., Cleveland, OH, [2]Cleveland Clinic, Cleveland, OH, [3]Case Western Reserve Univ., Cleveland, OH

Abstract:

Research on genomic interventions has the potential to yield major advances in reproductive outcomes. For same-sex couples and transgender/gender-diverse (TGD) individuals, interventions such as gamete modification, mitochondrial replacement, and in-vitro gametogenesis could enable genetic and/or gestational parenthood. These groups have historically experienced marginalization in medical settings, particularly in reproductive health care and assisted reproduction. As “reprogenomic” interventions move toward human trials, it is essential to engage with people who may participate during pregnancy or become pregnant through research participation. The unique perspectives and priorities of sex- and/or gender-diverse (SGD) communities must be included early in

research translation, in order to avoid exclusionary practices and other future harms. As part of a larger study on reprogenomic research ethics, we aimed to capture SGD perspectives and integrate them into an ethical framework for research governance. Methods: Interviews with 51 multidisciplinary experts, including 5 experts in SGD health, were followed by 6 focus groups with individuals who self-identified as transgender/gender diverse and/or lesbian/gay/bisexual. Results: Interviews and focus groups revealed concerns regarding the ethics of reprogenomic clinical trials, including concerns regarding future uses of reprogenomics. Participants discussed the need for gender-neutral or inclusive language in research design and recruitment, also highlighting the importance of intersectional identities and the ways that these interventions may intersect or interfere with gender-affirming care. Some participants mentioned benefits for achieving parenthood or for avoiding serious disease in offspring. However, participants also expressed concerns that these technologies could increase discrimination, including efforts to eliminate sexuality and gender diversity related genes and future “double discrimination” against children born into SGD families via novel forms of assisted reproduction. Conclusions: This qualitative study examined the thoughts and opinions of SGD individuals about future reprogenomic trials, connecting these to existing barriers to healthcare and research participation among SGD communities and identifying some ways of addressing concerns. These findings provide insights into opportunities for increased inclusivity in governance for future human trials of reprogenomic interventions.

The NIH INCLUDE Project: Over five years of transformational research for people with Down syndrome

Authors: M. Parisi[1], S. Bardhan[1], M. Brown[1], K. Davis[2], L. Garcia[1], H. Li[3], A. Mazzucco[2], M. Oneill[1], L. Ryan[4], C. Schramm[3], B. Schwartz[3], E. Tarver[4]; [1]NICHD/NIH, Bethesda, MD, [2]OD/NIH, Bethesda, MD, [3]NHLBI/NIH, Bethesda, MD, [4]NIA/NIH, Bethesda, MD

Abstract:

Introduction: The National Institutes of Health (NIH) launched the INCLUDE (INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndromE; https://www.nih.gov/include-project) Project in 2018. INCLUDE promotes research on conditions that affect those with Down syndrome (DS) and the general population, such as Alzheimer’s disease, autism, obstructive sleep apnea, celiac disease, congenital heart disease, and diabetes. INCLUDE also aims to expand the diversity of individuals participating in research and the number of investigators engaged in DS research. Methods: The INCLUDE Project has three components, including: (1) Conduct

targeted, high-risk, high-reward basic science studies on chromosome 21; (2) Assemble a large study population of individuals with DS; and (3) Conduct clinical trials research inclusive of individuals with DS. There are currently 18 active Funding Opportunities, including seven focusing on training, career development, and fellowship awards, and several focusing on diversity. A major cohort and biobanking initiative is being launched in FY2024. The INCLUDE Data Coordinating Center (DCC; https://includedcc.org/) is promoting data sharing by creating tools for clinical data visualization and analysis of extant clinical datasets through the INCLUDE Data Hub, with an Experimental Models Portal containing iPSC and animal model data under development. The DCC has launched a training initiative for diverse data scholars through a summer internship program. In addition, the DS-Connect® registry is being revamped to enhance research participation by connecting families with research opportunities. Results: NIH has invested $348 million over the past 6 years by issuing grants to support 330 awards. In FY2023, the INCLUDE investment was $90 million of a total NIH DS funding of $133 million. Over 75 trainees have been supported during the past 6 years. DS-Connect® has over 5900 registered participants and has promoted research enrollment for 100 projects, 18 INCLUDE-funded, and 5 clinical trials. The INCLUDE DCC Data Hub has clinical and ‘omics data from over 9000 participants in multiple studies, and includes over 2700 whole genomes and 499 transcriptomes. Conclusions: The INCLUDE Project has stimulated DS research funding since its launch, while the DS-Connect® registry and INCLUDE DCC are facilitating subject recruitment and data sharing. Efforts to increase the diversity of research participants and investigators through community outreach are essential to create a more representative cohort to improve quality-of-life for those with DS.

Session 55: Insights into Somatic Mosaicism and Human Diseases

Location: Room 501

Session Time: Thursday, November 7, 2024, 1:15 pm - 2:15 pm

A personalized multi-platform assessment of somatic mosaicism in the human frontal cortex

Authors: W. Zhou[1], C. Mumm[1], Y. Gan[1], J. Switzenberg[1], P. De Oliveira[2], J. Wang[1], K. Kathuria[2], B. Bessell[1], T. McDonald[1], K. Van Deynze[1], M. McConnell[2], A. Boyle[1], R. Mills[1]; [1]Univ. of Michigan, ANN ARBOR, MI, [2]Lieber Inst. for Brain Dev., Baltimore, MD

Abstract:

Somatic mutations in individual cells lead to genomic mosaicism, contributing to the intricate regulatory landscape of genetic disorders and cancers. To assess the capabilities for detecting somatic mosaicism across different technologies, we obtained tissue from the dorsolateral prefrontal cortex (DLPFC) of a post-mortem neurotypical 31-year-old individual. We sequenced bulk DLPFC tissue using a PromethION2 (P2) Solo (~60X), linked-read sequencing (~53X), and matched fibroblasts using NovaSeq (~30x). Additionally, we coupled Cas9 capture methodology with long-read sequencing (TEnCATS, targeting active transposable elements (TE). We further isolated and amplified DNA from 120 flow-sorted single DLPFC neurons using MALBAC and sequenced 115 neurons on MinIONs and 94 neurons by NovaSeq (89 matched). We then constructed a haplotype-resolved assembly to facilitate cross-platform analysis of somatic genetic variations against a personalized reference that comprised a total length of 2.81 Gb and a 92.07% heterozygous phased SNV consistency compared to 10X linked-read data. We generated a catalog of phased germline SNVs, CNVs, and TEs for benchmarking from the assembled genome. We applied standard approaches to recall these variants across sequencing technologies, achieving aggregated recall rates of 97.8%, 79.1%, and 91.7% respectively, providing an upper bound for detection limits. We next examined somatic variation using long-reads in 115 individual neurons and identified 1950 somatic heterozygous large deletions (50kb-7.7Mb), four of which possessed putative breakpoint signals. We observed an increase in phase rate from 10% to 45% between short- and long-read technologies, improving our somatic CNV detection. We further identified 39 somatic TEs (36 Alus, 3 LINE-1s) in the single cell data, 90% (35/39) of which were missed by short reads. Collectively, we have developed a multiplatform genomic resource for the human frontal cortex genome and anticipate that

this infrastructure will propel tool development and methodology benchmarking within the community.

Reconstructing Cell Lineage in Human Brain Using Somatic Mutations in Microsatellites

Authors: D. Snellings, E. Goodman, C. Walsh; Boston Children's Hosp., Boston, MA

Abstract:

Lineage tracing is the process of mapping a cell's ancestry through developmental lineages, all the way back to the zygote. Lineage tracing in primary human tissue is made possible by tracking somatic mutations that arise during normal development at a rate of ~2-4 mutations per division which are stably inherited by daughter cells and serve as endogenous lineage markers. However, this method is limited by the immense cost of sequencing the entire genome of individual cells. Here, we present preliminary data on a method that circumvents this limitation by targeting sequencing to microsatellites. Microsatellites and other repetitive regions of DNA mutate faster than any other regions of the genome and therefore provide a rich source of somatic mutations to serve as endogenous lineage markers. The low cost of targeted sequencing makes it feasible to assay hundreds of cells from a single sample yielding high resolution lineage information for a fraction of the cost of single-cell whole genome sequencing. We present data from 486 neurons and oligodendrocytes isolated from two postmortem brain samples. In each cell we genotyped 20,000-140,000 microsatellites and leveraged somatic mutations shared by multiple cells to infer lineage relationships.

Somatic genomic changes in single ischemic human heart cardiomyocytes

Authors: Z. An[1,2,3], N. Hilal[1,2,3], M. Prondzynski[1,2], M. A. Trembley[1,2], E. Wang[1], S. Araten[1], I. Sivankutty[1,2,3], M-H. Chen[1,2], W. Pu[1,2], Y. Huang[1,2,3], S. Choudhury[1,2,3]; [1]Boston Children's Hosp., Boston, MA, [2]Harvard Med. Sch., Boston, MA, [3]Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

Heart failure (HF) is a complex chronic condition wherein the heart struggles to pump blood efficiently, often resulting from diseases that compromise myocardial function. Ischemic heart disease (IHD) results from reduced blood flow to the heart caused by narrowed or obstructed coronary arteries, potentially leading to HF due to the gradual

weakening of the heart muscle from chronic ischemia. The risk of developing IHD is related to a complex interplay between genetic, lifestyle, and environmental factors.To evaluate the impact of somatic changes in the heart muscle cell genome and the direct DNA damage, we performed single-cell whole-genome sequencing from 33 diploid cardiomyocytes extracted from the left ventricles of 13 individuals with ischemic heart disease and age-matched healthy controls. Our findings reveal that in healthy individuals, normal cardiomyocytes primarily accumulate somatic single-nucleotide variants (sSNVs) in an age-related manner (SBS5, a “clock-like” signature observed in almost all cell types), reflecting the gradual accumulation of somatic mutations over time. In contrast, individuals with IHD exhibit a significantly higher burden of sSNVs driven by distinct mutagenesis mechanisms (SBS30, SBS32, and SBS44), which are associated with the deficiency in base excision repair, immune suppression, and defective DNA mismatch repair, respectively. These alterations are notably marked by specific nucleotide changes like C>T and C>A substitutions, which offer insights into the molecular pathways involved in IHD progression. Analysis of single-nucleus RNA sequencing (snRNAseq) indicates that IHD affects the coding exons of cardiomyocytes, exerting an influence on their contractility function.Our results suggest that recognized pathogenic mechanisms in IHD may cause genomic damage, subsequently resulting in a gradual impairment of function in cardiomyocytes. Investigating the abnormal buildup of DNA alterations in the ischemic heart sheds light on the cascade of molecular and cellular events that occur during the development of IHD.

Genotype-Informed Single-Cell RNA-Seq Reveals Somatic Loss of Heterozygosity in Hemimegalencephaly with *PIK3CA* Mutations

Authors: M. Gade[1], D. Lai[1], A. Poduri[2], H. Won[3], W. Ma[4], D. Wu[5], E. Heinzen[1,3]; [1]Eshelman Sch. of Pharmacy, Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [2]Boston Children's Hosp, Boston, MA, [3]Dept. of Genetics, Sch. of Med., Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [4]Dept. of Genetics, Sch. of Med., Univ. of North Carolina at Chapel Hill,, Chapel Hill, NC, [5]Dept. of Biostatistics, Gillings Sch. of Publ. Hlth., Univ. of North Carolina at Chapel Hill, Chapel Hill, NC

Abstract:

Hemimegalencephaly (HMEG) is a rare pediatric neurodevelopmental disorder characterized by the asymmetric enlargement of one cerebral hemisphere. This condition is linked to somatic mutations in the PI3K-AKT-mTOR signaling pathway. Studying surgically resected mosaic tissue from individuals with HMEG with pathogenic somatic variants

offers a unique opportunity to explore disease mechanisms. In this study, we use SoMoSeq (Somatic Mosaicism Sequencing), a novel protocol to analyze both DNA and RNA from the same single cell. We focus on individuals with a recurrent somatic mutation associated with HMEG (*PIK3CA*: E545K) to identify specific cell types carrying these pathogenic variants and determine cell type specific transcriptional changes associated with the pathogenic variant. In this study, we performed SoMoSeq on 4,300 nuclei extracted from three individuals with HMEG, who had a variant allele frequencies (VAF) between 25-14%, and compared them to three age, sex, and region matched neurotypical controls. Single-cell genotyping revealed three distinct populations: heterozygous (~44-28% cells), homozygous variant (~6-2% cells), and homozygous wild type (~50-70% cells). Orthogonal confirmation using whole genome amplification (Picoplex) followed by PCR amplification of the *PIK3CA* locus, and Sanger sequencing confirmed the presence of these genotypes, suggesting possible loss of heterozygosity (LOH) on chromosome three, in at least 2 out of the three cases (third case pending analysis). We are currently conducting whole genome sequencing with primary template amplification and genome-wide SNP array-based analysis to determine the LOH coordinates. Preliminary cell type specific differential gene expression analyses between heterozygous vs homozygous wild type populations revealed cell-type specific activation of the mTOR signaling pathway across major cell types within cases. Analysis utilizing a mixed effects model that incorporates genotype, case-control status , and cell type to look at gene expression across the three populations are currently ongoing. The detection of LOH in 2-6% of cells aligns with Knudson's two-hit hypothesis, suggesting that a secondary genetic event (LOH) contributes to the pathogenesis of HMEG in addition to the initial somatic mutation. This finding provides a framework for understanding the cell-autonomous and non-cell-autonomous effects of *PIK3CA* mutations, potentially offering new insights into the cellular mechanisms driving the pathophysiology of HMEG.

Session 56: Neurogenomic Approaches Translating Risk Variants to Disease

Location: Four Seasons Ballroom 2&3

Session Time: Thursday, November 7, 2024, 1:15 pm - 2:15 pm

Complex Structural Genome Variation in the Genetic Architecture of Neuropsychiatric disorders: Insights from Human Population Analysis and from Postmortem Brains of Individuals with Psychiatric Disorders

Authors: B. Zhou[1], J. G. Arthur[1], H. Guo[1], T. Kim[2], Y. Huang[1], R. Pattni[1], T. Wang[1], S. Kundu[1], J. X. J. Luo[1], H. Lee[1], D. Nachun[1], C. Purmann[1], E. M. Monte[1], A. Weimer[1], P. Qu[1], J. F. Fullard[3], J. Bendl[4], K. Girdhar[3], L. Duncan[1], H. P. Ji[1], X. Zhu[5], G. Song[2], D. Palejev[6], S. B. Montgomery[1], H. Dohna[7], P. Roussos[3], A. Kundaje[1], J. F. Hallmayer[1], M. P. Snyder[1], W. H. Wong[1], A. E. Urban[1]; [1]Stanford Univ., Stanford, CA, [2]Pusan Natl. Univ., Pusan, Korea, Republic of, [3]Icahn Sch. of Med. at Mount Sinai, New York, NY, [4]Icahn Sch. of Med. at Mount Sinai, New York City, NY, [5]The Pennsylvania State Univ., University Park, PA, [6]Bulgarian Academy of Sci., Sofia, Bulgaria, [7]American Univ. of Beirut, Beirut, Lebanon

Abstract:

Psychiatric disorders such as schizophrenia and bipolar disorder have a strong but complex genetic component to their molecular etiology. Multiple candidate loci have been identified by large GWAS, but individual variants are expected to have small effect sizes and to act in combination with other variants. Oftentimes, the functional variant will not be the GWAS marker SNP, but a nearby variant of a different nature (e.g., structural variant). All human genomes contain many complex structural variations (cxSVs), but their functions in genome biology are mostly unknown as they are effectively excluded from standard genome analyses, due to the technical limitations in accurate detection. We developed Automated Reconstruction of Complex Structural Variations (ARC-SV), a probabilistic and machine-learning method (which leverages the new Human Pangenome Reference) that permits the characterization of cxSVs on a population scale with unprecedented accuracy from standard short-read whole-genome sequencing (WGS). From 4,262 genomes spanning all continental populations, we identified unique 8,493 cxSVs belonging to more than 12 subclasses. We found cxSVs that are rare in the human population (i.e. those with signatures of negative selection) to be especially enriched, more than other variant types, for neural genes and for loci that underwent rapid human-specific evolution, including those that regulate human-specific corticogenesis. From 119 PsychENCODE postmortem brains including those from individuals with schizophrenia and bipolar disorder, by

leveraging single-nuclei multiomics coupled with WGS, we found, across multiple brain regions, significant associations between cxSVs and differentially expressed genes and accessible chromatin. Furthermore, we found that the cxSVs identified from the psychiatric cases of these 119 brains to show enriched linkage with psychiatric GWAS risk alleles (but not for the healthy controls) where linkage is much stronger than that of simple SVs. By developing a rigorous statistical approach that overcome sample size limitations, we integrated multi-modal data dimensions (genotype, phenotype, single-nuclei RNA-seq/ATAC-Seq) from these 119 brains where our results show significantly decreased expression of cxSV-associated genes among psychiatric cases across multiple brain regions and different cell types, thus implicating cxSVs as a previously unknown contributing factor in the complex genetic architecture of human neuropsychiatric disorders.

Sex differences of the spatiotemporally dynamic FMRP-RNA interactome in the human brain

Authors: A. Lee, X. Guo, Y. Li, S. Yu, Z. Qin, A. Shafik, P. Jin; Emory Univ., Atlanta, GA

Abstract:

Fragile X syndrome (FXS) is the most common inherited cause of intellectual disability and monogenic cause of autism spectrum disorder (ASD). Other features of FXS include developmental delay, anxiety, and seizures. FXS is caused by a CGG trinucleotide repeat expansion in the 5'-UTR of the *FMR1* gene on the X chromosome that leads to the functional loss of fragile X messenger ribonucleoprotein (FMRP), an RNA binding protein involved in diverse cellular processes regulating translation, transcription, and protein function. FXS disproportionately affects males, and affected females typically tend to have milder symptoms, which has been interpreted to be due to X inactivation. **Using enhanced crosslinking and immunoprecipitation followed by high-throughput sequencing (eCLIP-seq), we investigated the FMRP-RNA interactomes in postmortem male and female human brains of different ages and regions.** Included in our dataset were a total of 55 samples from the dorsolateral and medial prefrontal cortex (Brodmann area 9), caudate nucleus, anterior cingulate cortex (Brodmann area 24), hippocampus, and thalamus of 31 individuals. **We found that the FMRP-RNA interactomes are spatially and temporally dynamic. Moreover, sex-dependent binding of FMRP was observed**. Comparing male and female samples across the infant (<1 year of age), young adult (30s), and elderly (80s) groups of a brain region (BA9 or caudate nucleus), we identified female-specific FMRP binding peaks on genes *TMEM245*, *ATP1A1*, and *SCAMP5*. Similarly, FMRP

binding peaks on genes *LHFPL2* and *TEX41* were found to be unique to female caudate nucleus samples. Gene ontology and gene-disease-association analyses revealed distinct features of these sex-by-region group-specific targets. Furthermore, as ASD and seizures are present in many individuals with FXS, we examined FMRP binding to mRNAs of genes known to be associated with ASD or epilepsy. The levels of FMRP binding mRNAs of the 233 genes with a SFARI Gene Score of 1 (high confidence genes associated with ASD risk) displayed statistically significant differences between males and females in both BA9 and caudate nucleus. FMRP binding to RNAs of a previously published list of genes highly associated with epilepsy also exhibited sex differences in BA9. **Our findings that FMRP could bind to different regions or sets of RNAs in male and female brains expand on the explanation for differences in clinical phenotypes between male and female FXS patients. Overall, our results reveal the dynamic nature of the human FMRP-RNA interactome that is dependent on brain region and sex and highlight the complexity of the pathophysiology of fragile X syndrome.**

ASXL1 mutations drive mitochondrial dysfunction, resulting in disrupted mTOR signaling and cellular proliferation in Bohring Opitz Syndrome

Authors: V. Arboleda[1], A. Krall[2], H. Christofk[2], B. Russell[3], I. Lin[2]; [1]UCLA, ENCINO, CA, [2]UCLA, Los Angeles, CA, [3]UCLA David Geffen Sch. of Med., Los Angeles, CA

Abstract:

Bohring-Opitz Syndrome (BOS, OMIM#605309) is a rare neurodevelopmental disorder and one of the causes are heterozygous truncating mutations in the *ASXL1* gene. Patients are identified at birth with significant neurodevelopmental impairments and distinctive craniofacial anomalies. While previous research has primarily focused on the genetic and molecular dimensions of BOS, our study reveals novel insights into the metabolic and mitochondrial dysfunctions defects in BOS. Eight BOS individuals and eight sex-matched controls provided skin punch biopsies for patient-derived fibroblast cultures. We conducted a comprehensive study using patient-derived fibroblasts to investigate the metabolic and mitochondrial consequences of *ASXL1* mutations in BOS. We first assessed whether BOS cells were more sensitive to nutrient deplete using scratch wound assay, a high throughput assay measuring cellular growth and migration. Only after growth in media that is depleted of non-essential amino acids (NEAA) did BOS cells showed a significant 40-50% decrease in growth and migration rates compared with control cells. Repletion with L-asparagine partially rescued the wound closure rate (40-50%). Given this sensitivity to NEAA depletion, we next wanted to measure differences in oxygen consumption rate (OCR)

and extracellular acidification rate (ECAR). BOS cells showed a significant 48% increase in ECAR (padj <0.001) highlighting a preference for glycolysis over more efficient energy generation via the electron transport chain. We next performed targeted polar metabolomic assays by mass spectrometry to identify differentially regulated metabolites. Consistent with the energy preference for glycolysis, we saw significant increases in 8 of the 10 glycolysis intermediates measured. Since pyruvate was also significantly upregulated, we asked whether the main mitochondrial pyruvate transporter comprised of MPC1 and MPC2 proteins were decreased nearly 80% in BOS cells compared to controls. No significant changes were seen in mitochondrial mass or morphology that might explain our findings. Finally, given the link to nutrient depletion, we asked whether BOS cells had defective nutrient sensing through the mTOR pathway. Western blot analysis revealed that NEAA depletion decreased mTORC signaling, evidenced by reduced S6K phosphorylation. The decreased phosphorylation was rescued by L-asparagine repletion. These findings highlight the novel effects of *ASXL1* mutation on metabolism driving cell growth and proliferation. Our work provides a novel link between *ASXL1* mutation and metabolic dysregulation in BOS.

Translating *IGHMBP2* variants with a patient-specific neuromuscular junction system: Personalized medicine rescue

Authors: C. Tyner[1], S. Smieszek[1], B. Przychodzen[1], **H. Bai**[1], C. Johnson[1], C. Polymeropoulos[1], G. Birznieks[1], W. Hagan[2], C. Niccum[2], R. Brighton[2], X. Guo[3], R. Aiken[3], A. Nawaz[3], K. Hawkins[3], J. Hickman[2], M. Polymeropoulos[1]; [1]Vanda Pharmaceuticals Inc., Washington, DC, [2]Hesperos Inc., Orlando, FL, [3]Univ. of Central Florida, Orlando, FL

Abstract:

Charcot-Marie-Tooth disease Type 2S (CMT2S) is a rare Charcot-Marie-Tooth disease subtype caused by immunoglobulin mu-binding protein 2 (*IGHMBP2*) variants that result in abnormal RNA processing leading to alpha-motor neuron degeneration. A patient was reported with pathogenic variants within *IGHMBP2*. Whole genome sequencing revealed a cryptic splice site variant (c.1235+894 C>A) deep in intron 8, which leads to nonsense-mediated decay resulting in IGHMBP2 haploinsufficiency. We designed an antisense oligonucleotide, VCA-894A, that targets this specific cryptic splice site to restore IGHMBP2 protein levels. Our objective was to develop a patient-specific *in vitro* model to characterize the morphology and electrophysiology of CMT2S motor neurons (CMT2S-MNs) and neuromuscular junctions (NMJs), and to test the efficacy of VCA-894A to restore observed deficits. CMT2S-MNs were differentiated from an induced pluripotent stem cell (iPSC) cell

line generated from the patient's fibroblasts. Patch clamp electrophysiology, phase imaging, and immunocytochemistry were utilized to characterize CMT2S-MNs. To model the NMJ, CMT2S-MNs and wild-type (WT) control iPSCs were integrated into a dual-chamber NMJ platform with WT iPSC-derived skeletal muscle myofibers. NMJ functional defects were analyzed by NMJ number per chamber, fidelity, and fatigue index (FI). NMJ systems were then incubated with VCA-894A (10 nM, 100 nM, and 1 μM). Morphology differences observed in CMT2S-MNs included thinner processes with a higher number of varicosities in the processes and longer axonal lengths. Patch-clamp electrophysiology revealed hyperexcitability and spontaneous firing of CMT2S-MNs and reduced resting membrane potential, capacitance, and cell body area compared to WT-MNs. CMT2S-NMJ analysis revealed no fidelity or NMJ number differences but a higher FI than WT-NMJ with quick fatigue, characterized as tetanus followed by decay. We demonstrate rescue of NMJ functioning following ASO treatment, captured by a FI decrease and reduction in decay and sporadic responses. NMJ FI showed high and quick fatigue simulating the clinical phenotype of muscle fatigue and weakness. Tetanus observed in the CMT2S NMJ system was predominantly tetanus followed by decay; irregular tetanus may be responsible for frequent falls of CMT2S patients. We showed rescue of NMJ functioning post VCA-894A treatment, captured by FI decrease and reduction in decay and sporadic responses. This may lead to clinical motor control restoration and an improvement in muscle fatigue. We are further analyzing this patient-specific model to continue phenotyping CMT2S caused by *IGHMBP2* variants.

Session 57: Population Genetics Methods Matter

Location: Four Seasons Ballroom 1

Session Time: Thursday, November 7, 2024, 1:15 pm - 2:15 pm

Characterizing features affecting local ancestry inference performance in diverse admixed populations

Authors: J. Honorato Mauer[1], N. Shah[1], A. Maihofer[2], C. Zai[3], C. Nievergelt[2], S. Belangero[4], M. Santoro[4], E. Atkinson[1]; [1]Baylor Coll. of Med., Houston, TX, [2]Univ California San Diego, La Jolla, CA, [3]Ctr. for Addiction and Mental Hlth., Toronto, ON, Canada, [4]Univ.e Federal de Sao Paulo, Sao Paulo, Brazil

Abstract:

In recent years, there has been significant effort in developing and improving methods for the genomic study of admixed populations using Local Ancestry Inference (LAI), including applications in medical genetics (e.g. producing ancestry-specific allele frequencies), statistical genetics (e.g. ancestry-informed gene discovery), and population genetics (to understand past admixture events and demography). In all these applications, it is of the highest importance that LAI results are accurate so that downstream methods output results that best reflect the genetic ancestry of research participants. We comprehensively tested analytic strategies for LAI to provide guidelines for parameter setting and reference panel compositions with particular attention to three-way admixed populations reflective of Latin America's primary continental ancestries: African (AFR), Amerindigenous (AMR), and European (EUR). We also examined miscall frequencies in these tests and compared error modes and true positive calling rates. After simulating LD-informed admixed haplotypes under a variety of 2 and 3-way admixed demographic models, we implemented a commonly used LAI pipeline (RFMix v1.25), testing various reference panel compositions, using 1000 Genomes and Human Genome Diversity Project samples, to quantify their relative overall and ancestry-specific accuracy (true-positive) rates. We observed that AMR ancestry tracts suffer notably reduced accuracy compared to EUR and AFR tracts in all tested comparisons - true positive rate means for AMR tracts ranged from 88-94%, while EUR ranged from 96-99% and AFR 98-99% - many of which had statistically significant differences (Wilcoxon test p-adj < 0.05). When miscalls occured, LAI error rates arise most frequently in the direction of erroneously calling EUR ancestry in true AMR sites. We observed that using a reference panel composition that is well-matched to the target population, even with a low number of samples, produces true-positive estimates that are not statistically different from a high sample, mismatched reference (Wilcoxon test p-adj >

0.05), while being more computationally efficient. Though our investigations were focused on Latin American admixture compositions, the trends we characterize allow us to broadly provide recommendations for researchers when analyzing diverse populations to produce higher local ancestry inference accuracy results across diverse admixed populations. The trends observed in LAI miscalls reinforce the need for more underrepresented populations' inclusion in sequencing efforts for improving reference panels.

A genealogy-based approach for revealing ancestry-specific structures in admixed populations

Authors: J. Tang[1], C. Chiang[2]; [1]Univ. of Southern California, Los angeles, CA, [2]Univ. of Southern California, Los Angeles, CA

Abstract:

Revealing ancestry-specific structures in admixed populations is critical for understanding the population history and adjusting for population stratifications in genome-wide association studies. Existing methods for elucidating the ancestry-specific structures generally rely on frequency-based estimates of genetic relationship matrix (GRM) among admixed individuals after masking segments from ancestry components not being targeted for investigation. A variant of principal component analysis (PCA) possibly supplemented with a uniform manifold approximation and projection (UMAP) analysis on the resulting GRM would then be used to investigate the ancestry-specific structure. However, these methods continue to ignore linkage information between markers and thus are expected to have decreased resolution in revealing structure within an ancestry component. Building on our previous work devising a genealogy-based estimator of GRM, the eGRM, to reveal fine-scale structures, we propose a framework named as-eGRM. This approach intersects information from eGRM with local ancestry callset to infer ancestry-specific structure in admixed populations. We evaluated the as-eGRM framework in extensive simulations, including models based on (1) a simple two-population split, two-way admixture model, (2) a grid-like stepping-stone model with variable admixture, and (3) a realistic Latino 3-way admixture model previously inferred. We assessed the effectiveness of elucidating fine-scale structure as measured by the Separation Index (SI), which assesses the proportion of nearest neighbors that are in the same subpopulation in simulated "ground truth" multi-dimensional space. Compared to current alternatives that either ignore admixture (based on eGRM alone) or ignore the linkage information (based on ancestry-specific masking in constructing the GRM), we found that as-eGRM is better at elucidating ancestry-specific fine-scale structure. For example, in the grid-like stepping-stone model with admixture

proportions between 0.1-0.3 across nine demes, PCA+UMAP of as-eGRM achieved a SI = 0.75, compared to SI = 0.62 and 0.32 when using eGRM and Missing DNA PCA (mdPCA), respectively. Under the Latino demographic model, PCA+UMAP of as-eGRM achieved a SI = 0.88, compared to SI = 0.62 and 0.50 when using eGRM and mdPCA, respectively. Taken together, as-eGRM has the promise to better reveal the fine-scale structure within an ancestry component of admixed individuals, which can help improve the robustness and interpretation of findings from association studies of disease or complex traits for these understudied populations.

Deep learning-augmented models of gnomAD v4 enable estimation of LoF mutational constraint for all human genes

Authors: J. Guez[1], K. Laricchia[1], S. Parsa[2], J. Goodrich[2], K. Chao[1], gnomAD Aggregation Consortium, C. Seed[3], H. Rehm[4], M. Daly[4], B. Neale[4], H. Finucane[5], K. Samocha[4], K. Karczewski[6]; [1]Broad Inst. of MIT and Harvard, Cambridge, MA, [2]Broad Inst., Cambridge, MA, [3]Broad Inst, Cambridge, MA, [4]Massachusetts Gen. Hosp., Boston, MA, [5]Broad Inst. of MIT and Harvard, Boston, MA, [6]Massachusetts Gen. Hosp., Medford, MA

Abstract:

The accurate estimation of mutational constraint is critical for prioritizing genes in rare disease diagnosis. In the gnomAD project, we have previously developed constraint metrics (e.g. pLI and LOEUF), which are widely used in both research and diagnostics. However, these methods are currently limited by their reliance on only pLoF mutations, which are exceedingly rare. These metrics are therefore underpowered for genes with low expected mutation counts (in v2, 28% of genes), and increasing power for constraint metrics will reveal genes under recessive constraint.
Here, we increase the resolution of LOEUF by increasing sample size, incorporating the full site frequency spectrum from multiple genetic ancestry groups, and leveraging information from protein language models that encode elements of species conservation and thus missense deleteriousness. In updated constraint metrics based on 730K exomes in gnomAD v4, we find 22% of genes with strong evidence of selective constraint (LOEUF < 0.6). However, 15% of genes remain underpowered, with an expected number of pLoF variants < 10. Thus, we further improve the method by leveraging the full site frequency spectrum (SFS), compared to the original LOEUF which considers only the presence or absence of mutations, by weighting mutations based on their frequency. Furthermore, gnomAD v4 is more diverse than previous releases, with a 2.9x increase in non-European genetic ancestry groups. With this diversity, we extend LOEUF by calculating observed and

expected SFS for each ancestry group, enabling the assessment of selective pressure on genes under diverse demographic histories, thereby increasing power. Finally, we leverage recent transformer models trained on a large number of sequences from different species, whose embeddings contain rich information about variant deleteriousness. We decompose LOEUF as a nonlinear function of these embeddings for genes with sufficient power and then apply this formula to underpowered genes.
We employ these approaches concomitantly to refine LOEUF in a composite method we term LOEUF-ALL (LOEUF augmented with language layers). This new method combines allele frequency information within human diversity in a large dataset with information about conservation across species and missense deleteriousness to increase the precision of estimating selection on human genes. We demonstrate increased power to detect LoF constraint for all genes, and compare our approach to the other methods (e.g. original LOEUF, sHet). This method highlights the importance of including information from human diversity as well as multiple model inputs to understanding natural selection.

Genotype Representation Graphs: Enabling Efficient Analysis of Biobank-Scale Data

Authors: X. Wei, D. DeHaas, Z. Pan; Cornell, Ithaca, NY

Abstract:

Computational analysis of a large number of genomes requires a data structure that can represent the dataset compactly while also enabling efficient operations on variants and samples. Current practice is to store large-scale genetic polymorphism data using tabular data structures and file formats, where rows and columns represent samples and genetic variants. However, encoding genetic data in such formats has become unsustainable. For example, the UK Biobank polymorphism data of 200,000 phased whole genomes has exceeded 350 terabytes (TB) in Variant Call Format (VCF), too large to fit into hard drives in uncompressed form. To mitigate the computational burden, we introduce the Genotype Representation Graph (GRG), an extremely compact data structure to losslessly present phased whole-genome polymorphisms. A GRG is a fully connected hierarchical graph that exploits variant-sharing across samples, leveraging on ideas inspired by Ancestral Recombination Graphs. Capturing variant-sharing in a graph format compresses biobank-scale data to the point where it can fit in a typical server's RAM (5-26GB per chromosome), and enables graph-traversal algorithms to trivially reuse computed values, both of which can significantly reduce computation time. We have developed a command-line tool and a library usable via both C++ and Python for constructing and processing GRG files which

scales to a million whole genomes. It takes 160GB disk space to encode the information in 200,000 UK Biobank phased whole genomes as a GRG, more than 2000 times smaller than the size of VCF. Moreover, the size of GRG increases sublinearly with the number of samples stored, making it a sustainable solution to the increasing number of samples in large datasets. We show that summaries of genetic variants can be computed on GRG via graph traversal that runs 230 times faster than on VCF. We anticipate that GRG-based algorithms will improve the scalability of various types of computation and generally lower the cost of analyzing large genomic datasets.

Session 58: Scaling Structural Birth Defects

Location: Room 401

Session Time: Thursday, November 7, 2024, 1:15 pm - 2:15 pm

Whole genome analysis of 137 trios with CHARGE-phenotype overlap

Authors: B. Muchmore[1], M. Dulchavsky[1], A. Nguyen[2], S. Regan[2], J. Kidd[2,3], S. Bielas[2], D. Martin[1,2,4]; [1]Univ. of Michigan, Div. of Pediatric Genetics, Metabolism and Genomic Med., Ann Arbor, MI, [2]Univ. of Michigan, Dept. of Human Genetics, Ann Arbor, MI, [3]Univ. of Michigan, Dept. of Computational Med. and Bioinformatics, Ann Arbor, MI, [4]Univ. of Michigan, Dept. of Pediatrics, Ann Arbor, MI

Abstract:

CHARGE syndrome is an autosomal-dominant, multiple congenital anomaly condition characterized by vision and hearing loss, congenital heart disease, and malformations of craniofacial and other structures. Pathogenic variants in *CHD7* are present in most individuals who have a clinical diagnosis, however, 5-30% of individuals with a clinical diagnosis have no identifiable *CHD7* variant. The purpose of this study is to assess the value of whole-genome sequencing (WGS) for diagnosis of cases with CHARGE-phenotype overlap, many of whom have already had undiagnostic whole-exome sequencing.
We performed WGS on 137 trio families in which the proband had at least one feature that overlapped CHARGE syndrome major criteria (coloboma, choanal atresia, cleft palate or ear abnormality) or minor criteria (cranial nerve dysfunction, dysphagia, structural brain abnormalities, developmental delay, intellectual disability, or autism spectrum disorder, hypothalamo-hypophyseal dysfunction or genital/heart/esophageal/renal/skeletal/limb anomalies). Analyses were run separately for protein-coding and non-protein coding regions: Protein coding analysis used a combination of LIRICAL, EXOMISER, and automated ACMG scoring while non-coding analysis used GENOMISER and GREEN-DB. Per-sample variants from the protein coding analysis were then filtered to keep those that were associated with a pathogenic ClinVar entry, had an ACMG score >5 or VEP impact was predicted to be “high”, and variants were prioritized based on the consensus rank of the LIRICAL, EXOMISER, and ACMG scores. Additionally, copy-number variants were analyzed using a paralog detector, QuicK-mer2.
Pathogenic variants in *CHD7* were present in 7/137 individuals (5.1%), and 3/137 (2.2%) individuals had pathogenic variants in a CHARGE-phenocopy (*KMT2D*). On preliminary review, a further 51/137 (37.2%) individuals have pathogenic variants in 36 additional genes associated with autosomal dominant conditions that have phenotype overlap with the proband. Additionally, 35/137 (25.6%) had one pathogenic variant found in a recessive

disorder that had overlap with the proband's phenotype, and multiple genes were found that are candidates for functional follow up.
These results demonstrate that phenotypic features of CHARGE syndrome overlap with multiple other rare single-gene syndromes. We expect integration of non-coding and structural variants to increase diagnostic yield, especially in cases in which a single pathogenic protein-coding variant has already been uncovered and is associated with an autosomal recessive disease that fits the proband's phenotype.

Functional validation of a novel gene associated with orofacial clefts

Authors: S. Murray[1], I. Welsh[1], M. Jarmusz[1], H. Brand[2], R. Green[3], S. Weinberg[4], J. Shaffer[3], E. Leslie[5], M. Marazita[6]; [1]The Jackson Lab., Bar Harbor, ME, [2]MGH, Boston, MA, [3]Univ. of Pittsburgh, Pittsburgh, PA, [4]Univ of Pittsburgh, Pittsburgh, PA, [5]Emory Univ., Atlanta, GA, [6]Ctr. for Craniofacial and Dental Genetics, Unive, Pittsburgh, PA

Abstract:

Orofacial clefts (OFCs) are one of the most common structural birth defects, affecting approximately 1 in 700 newborns worldwide. While association studies and rare variant sequencing efforts have advanced our understanding of the genetic causes of OFCs, only a small fraction of cases have been solved. The availability of large whole genome sequencing (WGS) datasets, including those generated through the Gabriella Miller Kids First Initiative, provide an opportunity to discover additional novel variants associated with OFCs and to provide insight into key pathways critical for normal facial development. Functional validation of these variants is essential to both corroborate the genetic findings and to better understand the biological mechanisms that underpin variant effects. The mouse is an ideal model system to understand the genetics of mammalian physiology and development, and thus serves as an important platform for interrogating novel variants associated with structural birth defects. We have developed a system that leverages the high efficiency of CRISPR/Cas9 genome editing and our high-throughput phenotyping capabilities developed for the Knockout Mouse Phenotyping Program (KOMP2) to rapidly test the effect of pathological variants in founder "F0" mouse embryos. In proof-of-principle experiments, we show that we can faithfully recapitulate known mutant phenotypes efficiently and rapidly. To expand our understanding of the genetic architecture of OFCs, we are interrogating the mutation spectrum in a WGS dataset of over 2,000 case/parent trios, investigating single nucleotide variants (SNVs), indels, and structural variants (SVs), which will then be validated in mice. In one example from this program, we recently identified a 100Kb deletion overlapping four protein coding genes, including *EIF3G*. This gene is expressed in the developing mouse palate, but to date no

mouse models have been reported to corroborate its role in OFCs. We show that deletion of this gene in mouse embryos results in a range of developmental abnormalities at midgestation, including cleft palate, abnormal lung development, microphthalmia, and limb defects. We generated germline mutants of this deletion that demonstrate significant haploinsufficiency that is highly sensitive to genetic background. Together, these studies corroborate the role of *EIF3G* in OFCs and underscore the value of combining human genetic discovery and animal model validation pipelines into a single program.

Analysis of rare *de novo* variants in 5707 congenital heart disease (CHD) trios identifies three novel CHD genes

Authors: K. Ng[1], N. Lake[1], S. Morton[2], S. DePalma[3], M. Wagner[4,5], J. Chen[4,5], P. Dexheimer[4], B. Gelb[6], Y. Shen[7], M. Tristani-Firouzi[8], J. Seidman[9], C. Seidman[3,10], M. Lek[1], M. Brueckner[1], Pediatric Cardiac Genomics Consortium; [1]Yale Univ., New Haven, CT, [2]Boston Children`s Hosp. /Harvard Med. Sch., Boston, MA, [3]Harvard Med. Sch., Boston, MA, [4]Cincinnati Children's Hosp. Med. Ctr., Cincinnati, OH, [5]Univ. of Cincinnati Coll. of Med., Cincinnati, OH, [6]Icahn Sch. of Med. at Mount Sinai, New York, NY, [7]Columbia Univ. Med. Ctr., New York, NY, [8]Univ. of Utah Sch. of Med., Salt Lake City, UT, [9]Harvard Med Sch, Boston, MA, [10]Brigham and Women's Hosp., Boston, MA

Abstract:

Congenital heart disease (CHD) is the most common birth defect, affecting approximately 1% of all live births. Genetic causes have been identified for ~34% of CHD cases, with potential genetic involvement in many more. Increasing sample sizes in genomic studies should significantly enhance our ability to uncover novel CHD genes. The Pediatrics Cardiac Genomics Consortium (PCGC) has recruited 17,064 CHD probands for study, with exome sequencing (ES) completed for 12,200 probands including 5,707 trios. In this study, our analysis of ES of 5,707 trios significantly increased the sample size compared to previous CHD studies, leading to the discovery of additional CHD genes. Using a calibrated mutation model, we assessed the mutational burden of rare loss-of-function and damaging missense (REVEL score > 0.5) de novo variants (DNVs). This model aggregates germline trinucleotide mutation rates of all damaging DNVs per gene, comparing expected versus observed DNVs using a Poisson test. Our analysis identified twenty genes significant at a false discovery rate of 0.05, including three novel genes not previously associated with CHD in humans: SETD5, EDNRA, and SLIT3. Notably, these three novel genes reside in the upper quartile of gene expression in E14.5 mouse embryo hearts. These genes are functionally implicated in transcriptional regulation during development via chromatin

modification or signaling pathways. Murine knockout models for SETD5, EDNRA, and SLIT3 recapitulated CHD phenotypes, reinforcing their role in CHD. Importantly, our findings expand the phenotypic spectrum of SETD5, which is known to be associated with Intellectual disability-facial dysmorphism syndrome. This study reveals a novel role for SETD5 in CHD pathogenesis, underscoring the multi-organ manifestations of this syndromic gene. To further strengthen the novel gene associations, we assessed the protein-protein interactions of the twenty significant genes using STRING, revealing significant network connectivity (p-value = 1.89E-10). Louvain clustering identified three subnetworks: chromatin modifiers, DNA-binding transcription factors, and the RAS-MAPK pathway. SETD5 falls within the chromatin modifier cluster, aligning the gene with previously known syndromic CHD genes. In conclusion, the substantial increase in sample size has led to the discovery of three novel CHD genes, providing valuable mechanistic insights into CHD pathogenesis and demonstrating significant phenotypic expansion for syndromic genes.

Unraveling the Diverse Genetic Architecture of Structural Birth Defects

Authors: A. Sanchis-Juan[1,2], E. Shin[1,2], N. E. Kurtas[1,2], A. S. Lee[1,2], J. Fu[1,2], X. Zhao[1,2], J. Lim[1,2], Y. Mostovoy[2], G. Lemire[2], GREGoR Consortium, Broad Center for Mendelian Genomics, GMKF OFC working group, Broad Genomics Platform, A. Jelin[3], T. H. Beaty[4], F. A. High[5], P. Donahoe[5], S. Ware[6], I. Krantz, D[7], J. Gleeson[8], A. Butali[9], E. Engle[10,11], A. Gharavi[12], A. O'Donnell Luria[2], H. L. Rehm[1,2], W. Chung[13], E. J. Leslie[14], M. L. Marazita[15], M. E. Talkowski[1,2], H. Brand[1,2]; [1]Dept. of Neurology, Harvard Med. Sch.; Ctr. for Genomic Med., Massachusetts Gen. Hosp., Boston, MA, [2]Broad Inst. Ctr. for Mendelian Genomics, Broad Inst. of MIT and Harvard; Program in Med. and Population Genetics, Broad Inst. of MIT and Harvard, Cambridge, MA, [3]Dept. of Gynecology and Obstetrics, The Johns Hopkins Med. Inst.s, Baltimore, MD, USA., Baltimore, MD, [4]Dept. of Epidemiology, Johns Hopkins Bloomberg Sch. of Publ. Hlth., Baltimore, MD, [5]Pediatric Surgical Res. Lab., Dept. of Surgery, Massachusetts Gen. Hosp., Harvard Med. Sch., Boston, MA, [6]Dept. of Pediatrics and Med. and Molecular Genetics, Indiana Univ. Sch. of Med., Indianapolis, IN, [7]Div. of Human Genetics, The Children's Hosp. of Philadelphia, Perelman Sch. of Med. at Univ. of Pennsylvania, Philadelphia, PA, [8]Dept. of NeuroSci.s, Univ. of California San Diego, La Jolla, CA, [9]Dept. Oral Pathology, Radiology and Med., Coll. of Dentistry, Univ. of Iowa, Iowa City, IA, [10]Dept. of Neurology and Ophthalmology, Boston Children's Hosp. and Harvard Med. Sch., Boston, MA, [11]Howard Hughes Med. Inst., Chevy Chase, MD, [12]Ctr. for Precision Genetics and Genomics, Dept. of Med., Columbia Univ., New York, NY, [13]Dept. of Pediatrics, Boston Children's Hosp., Harvard Med. Sch., Boston, MA, [14]Dept. of Human

Genetics, Emory Univ. Sch. of Med., Atlanta, GA, [15]Ctr. for Craniofacial and Dental Genetics, Dept. of Oral and Craniofacial Sci., Univ. of Pittsburgh, Pittsburgh, PA

Abstract:

Structural birth defects (SBDs) affect 1 in 33 newborns in the US and are a leading cause of neonatal death. Discovering the genetic basis of an SBD phenotype is critical for understanding its underlying mechanism, providing better genetic diagnostic screening, and improving therapeutic development. A primary challenge in gene discovery across SBDs has been the rarity of individual disorders, leading to underpowered cohorts. Although typically examined in isolation, numerous SBDs exhibit comorbidity in syndromic presentations, implicating common genetic underpinnings, pathways, and mechanisms. To explore the distinct and shared genetic architecture of SBDs, we aggregated genomic data across eleven international consortia, and amassed a cohort of 20,993 SBD trios (15,793 exomes and 5,200 genomes). The most common phenotypes include: orofacial clefts, congenital diaphragmatic hernia, and kidney abnormalities. We harmonized SNVs/indels and structural variants (SVs) across all cohorts using GATK and developed filtering pipelines for the sensitive detection of *de novo* variants across genomes and exomes. Overall, we discovered 30,730 *de novo* coding SNVs, 2,411 indels, and 2,680 SVs. We performed association analysis through application of the transmission and de novo association (TADA) framework, a Bayesian statistical model that improves gene discovery power by combining evidence across diverse genomic data types and variant classes. Application of TADA revealed 217 genes significant at a false discovery rate of less than 0.01. Most of these genes (n = 195) had previously been associated with syndromic SBDs, including *SATB2, KMT2A* and *ADNP*. The remaining 22 genes have been previously unreported as associated with an SBD. Among the significant genes we observed causative variants driven by multiple variant types, including several loss-of-function SVs disrupting *MED13L, EHMT1, TAOK1*, and *SATB2*. Importantly, *NIPBL, AKT3*, and *TNRC18*, with *TNRC18* being a potentially novel gene, would not have reached the significance threshold without SV evidence. The significance of 56 genes was primarily driven by damaging missense variants, including variants in *RAB11A* in 3 individuals with craniofacial dysmorphism and neurodevelopmental disorders. Additionally, variants in 6 of the 195 genes were in individuals presenting with an SBD phenotype not previously reported with the gene, further expanding their phenotypic spectrum. These findings underscore the potential benefits of aggregating extensive large-scale data from diverse disease cohorts to empower novel genetic discoveries and unravel the genetic underpinnings of SBD and other rare diseases.

Session 69: Complex Traits and Other Omics

Location: Mile High Ballroom 2&3

Session Time: Friday, November 8, 2024, 10:15 am - 11:45 am

Genetically predicted leukocyte telomere length from 800,000 individuals identifies novel phenotypic associations

Authors: K-H. Wu[1], M. Kessler[2], J. Tang[3], S. Alvarez[4], N. Banerjee[5], D. Sharma[1], J. Staples[6], D. Li[7], GHS-RGC DiscovEHR Collaboration, MAYO-RGC Project Generation, UCLA-RGC ATLAS Collaborations, Colorado Center for Personalized Medicine-RGC Collaboration, Penn Medicine BioBank (PMBB), L. Lotta[8], G. Abecasis[8], A. Ferrando[9], E. Jorgenson[8]; [1]Regeneron Pharmaceuticals, Tarrytown, NY, [2]Regeneron Genetics Ctr., Emerson, NY, [3]Regeneron Genetics Ctr., Tuckahoe, NY, [4]Regeneron, New York, NY, [5]Regeneron, Tarrytown, NY, [6]Regeneron Genetic Ctr., Ossining, NY, [7]Rengeneron Phaarmaceuticals, Tarrytown, NY, [8]Regeneron Genetics Ctr., Tarrytown, NY, [9]Regeneron Pharmaceuticals, Inc., New York, NY

Abstract:

Studies have reported complex relationships between leukocyte telomere length (LTL) and disease, with bidirectional effects and confounding due to age. It is known that telomere length shortens with age, longer in younger individuals and shorter in older population. Additionally, longer LTL has been associated with a decreased risk of pulmonary and hematologic diseases, but also with increased cancer risk. Here, we leveraged electronic health record data from 791,838 individuals of recent European ancestry across 6 biobanks and used the genetically predicted LTL to better understand the associations between telomere length with 1,643 disease phenotypes. First, we used experimentally derived measures of LTL and genetic data for UK Biobank participants to conduct a Genome-Wide Association Study in 432,479 individuals and build a polygenic score (PGS) of telomere length. Then, we calculated the PGS of telomere length (LTL-PGS) in 5 cohorts to conduct Phenome-Wide Association Study (PheWAS) and meta-analyzed the results to disentangle the relationship between leukocyte telomere length and human health at scale. In the meta-analyzed LTL-PGS PheWAS, a total of 112 disease codes (7% of those tested) were significantly associated with genetically predicted telomere length ($P<3x10^{-5}=0.05/1,643$). Association signals with LTL-PGS suggested that longer telomere length was associated with increased risk of cancer, consistent with previous reports and suggesting that cells with longer telomeres might be better placed to proliferate and expand in cancerous and pre-cancerous states. All 40 significant cancers were positively associated with telomere

length, replicating previous nominal results, and providing definitive support for the relationship between longer telomeres and increased cancer risk. Among cancer traits, prostate (OR: 1.08; p: $5x10^{-18}$), thyroid (OR: 1.17; P: $2x10^{-22}$), and melanoma (OR: 1.13; P: $1x10^{-21}$) cancers were most strongly associated. Shorter telomere length was associated with increased risk of pulmonary disease, including both restrictive (e.g. fibrosis) and obstructive (e.g. COPD) disease types, potentially suggesting that longer telomeres reflect improved ability of lung cells to regenerate and avoid disease. Finally, shorter telomere length was also associated with several cardiovascular disease (e.g. myocardial infarction and vascular events). These results demonstrate the utility of using genetic instruments at scale to disentangle age related disease mechanisms and provide strong support for previously reported relationships.

Multiomics approach identifies novel genes for Skeletal Class III malocclusion

Authors: A. Alade[1], K. Almpani[1], J. Lee[2]; [1]NIH/NIDCR, Rockville, MD, [2]NIH/NIDCR, Bethesda, MD

Abstract:

Skeletal Class III malocclusion (Class III) is a dentofacial deformity resulting from a prognathic/hyperplastic mandible, retrognathic/hypoplastic maxilla, or a combination of both. Affected Individuals present with aesthetic and functional problems. Early orthopedic treatment can mitigate the developing malocclusion and prevent associated psychosocial issues. However, the effectiveness of this approach depends on early diagnosis, which remains challenging. Etiology is multifactorial, with genetics playing a major role (estimated heritability ~40%). To identify genes/variants that could be used as biomarkers for early case identification, we conducted whole genome sequencing (WGS) on Class III-affected families. Three Class III-affected families, Family 1 with Ashkenazi Jewish ancestry and Families 2 and 3 with mixed Jewish and European ancestry were included in the study. Family 1 included an affected proband and an unaffected parent. Family 2 included an affected proband, an unaffected fraternal twin, and unaffected parent. Family 3 included an affected proband and father and an unaffected mother. We screened the WGS data for low-frequency and potentially pathogenic variants in genes expressed in craniofacial tissues during human face development using bulk mRNA sequencing of human craniofacial tissues. Subsequently, we used the SysFACE database to examine facial tissue expression of the mouse orthologs of these genes and prioritized genes expressed in the mandible/maxilla during mouse facial morphogenesis. We identified six potential candidate genes: *TSHZ3* and *ANTXR2* in Family 1, *PAX3* in Family 2 and *INSR*, *DCTD*, and *CCDC40* in Family 3. We found a de novo missense variant in

the *TSHZ3* gene and compound heterozygous (missense and splice site) variants in the *ANTXR2* gene. A homozygous recessive missense variant in the *PAX3* gene. Additionally, the affected proband presented with other facial features (broad nasal root, dystopia canthorum) previously reported for PAX3 gene mutations. Compound heterozygous variants (missense and synonymous) in the INSR gene, an autosomal dominant splice variant in the *CCDC40* gene, and an autosomal dominant missense variant in the *DCTD* gene. These genes have been previously implicated in conditions that present with craniofacial anomalies or are involved in processes crucial for craniofacial development.
Our findings demonstrate the importance of facial phenotyping and the power of integrating transcriptomics in the analysis of whole genome sequencing data to identify candidate genes/variants for Class III. Experiments to functionally validate the variants are ongoing.

Complex interactions of copy number variants on rare and common disorders

Authors: C. Smolen, S. Girirajan; Pennsylvania State Univ., State College, PA

Abstract:

Copy number variants (CNVs) are deletions or duplications of relatively large pieces of the genome. In recent decades, CNVs and the dosage-sensitive genes within them have been associated with a wide range of complex disorders and traits. However, while studies of rare and common single nucleotide variants have begun exploring the role of epistatic variant interactions, studies of CNVs have thus far been restricted to effects of individual CNVs or dosage-sensitive genes and the additive effects of total CNV burden. To bridge this gap, we used logistic regression models to examine pairwise multiplicative interactions between deletions and duplications of about 11,000 genes across the genome for their effect on 380 medical record-derived phenotypes on 373,894 participants from the UK Biobank. We identified 38,882 significant interactions between 24,899 pairs involving 847 genes across 351 phenotypes. Further examinations of these pairs revealed characteristics of CNV interactions. For example, interactions between duplicated genes were enriched and interactions involving deleted genes were depleted compared to the initial set of CNV genes (Chi squared p=0). Further, while most (66%) pairs had effects on a single phenotype, 8,476 pairs affected multiple phenotypes, such as epistatic interactions of *SPPL2C* and *PSG11* deletions influencing parathyroid disorders (FDR=1.59×10^{-5}), anxiety (FDR=8.15×10^{-4}), GI tract ulcers (FDR=8.67×10^{-5}), and three other phenotypes (FDR<0.013). CNV genes interacted with an average of 49.1 other genes, but the distribution had a positive skew with a few genes interacting with many other genes (median=15, std=68.7,

range=1-294), including *SLC2A14,* which had significant interactions with 294 other genes across 181 phenotypes. Interacting genes were more likely to be both deleted and duplicated in UK Biobank samples compared to the initial set of genes (Chi squared $p=2.21\times10^{-15}$), suggesting that these CNV genes have bi-directional interactions. Interestingly, many of these bi-directionally interacting CNV genes interacted with different genes when deleted and duplicated, such as *GTF2H2*, which interacts with 51 genes when deleted and 7 genes when duplicated, but only interactions with *SCART1* are consistent between dosages. Bi-directionally interacting genes shared an average of 14.2% of interaction partners between deletion and duplication interactions. These results suggest a complex landscape of dosage-sensitive CNV interactions across a wide spectrum of human disorders and underscore the importance of examining epistatic interactions between CNV genes in human disease.

Uncovering the nuclear genetic basis of mitochondrial DNA heteroplasmy

Authors: R. Alhariri[1], W. Shi[1], D. Puiu[1], N. Pankratz[2], N. Lake[3], M. Lek[4], Y. Huang[5], D. Bodnar[5], E. Tsai[5], J. Shi[5], D. Arking[6]; [1]Johns Hopkins Univ., Baltimore, MD, [2]Univ. of Minnesota, Minneapolis, MN, [3]Yale Univ., New Haven, CT, [4]Yale Univ. Sch. of Med., New Haven, CT, [5]Biogen, Cambridge, MA, [6]Johns Hopkins Univ Sch. of Med., Baltimore, MD

Abstract:

Background Mitochondrial heteroplasmy, the coexistence of mutated and wild type mitochondrial DNA within a cell, is associated with overall mortality and aging-related phenotypes. Mitochondrial mutations can be inherited maternally or acquired somatically, with previous work estimating that about 70% of heteroplasmies are somatic and increase in an age-dependent manner.
Methods To study the relationship between nuclear encoded variation and mitochondrial heteroplasmy, we conducted genome-wide association studies on heteroplasmy count as well as a functional approximation of the deleteriousness of the heteroplasmic mutations, using the mitochondrial local constraint score sum (MSS). MSS is a constraint-based metric that accounts for class of mutation and constraint at the nucleotide level. Whole-genome sequencing data from whole blood DNA samples were used to call heteroplasmy through the MitoHPC pipeline. The GWAS was run on REGENIE using 391,028 participants of European Ancestry from the UK Biobank. 11,392,286 SNPs (MAF >0.005) were included after imputation. We performed fine-mapping using SuSiE and colocalization analysis with COLOC, utilizing whole-blood eQTL expression data from GTEx. Genes were identified through a combination of positional mapping and colocalization analysis.
Results Three independent loci identified in the GWAS on heteroplasmy count were

mapped to *TERT*, *MDFIC*, *OXA1L* and *SLC7A7*. Comparatively, four independent loci were associated with MSS, mapping to *THRB*, *TERT*, *ITPR2*, and *TCL1A*. Among these hits, rs2887399 (β=0.004; $p=2.6x10^{-21}$) co-localized with reduced *TCL1A* expression.
Both *TCL1A* and *TERT* are associated with clonal hematopoiesis (CH). Given this relationship and our previous work that has shown that CH correlates with heteroplasmy, we investigated whether CH may be mediating the observed associations. After accounting for CH in our GWAS models, we observed no statistically significant difference in our results, indicating the signal is not mediated through CH. We also did not find the signal was mediated by leukocyte telomere length despite telomere length's association with *TERT* and heteroplasmy. These findings suggest a potential independent role for these genes in the development of heteroplasmy. In particular, *TERT* has been reported to serve a protective role against mitochondrial DNA from damage and reduce the production of reactive oxygen species. *IPTR2* also plays a role in mitochondria-ER mediated contact, metabolic stress and cellular senescence.
Conclusion Future directions involve carrying out a rare variant analysis to identify additional genes that may contribute to heteroplasmy.

A phenome-wide association study of the structural variants in 467,152 UK Biobank genomes identifies non-coding structural variants associated with human diseases

Authors: S. Atanur[1], F. Hu[1], Z. Zou[1], S. ODell[1], M. Feistner[1], Q. Wang[2], K. Carss[1], O. Burren[1], K. R. Smith[1], S. Petrovski[1]; [1]AstraZeneca Plc, Cambridge, United Kingdom, [2]AstraZeneca Plc, Waltham, MA

Abstract:

In Humans, structural variants (SV), defined as genomic rearrangements of 50 or more base pairs (bp), affect more bases of the genome than single nucleotide variants (SNVs)[1]. Even though SVs have been implicated in human disease, genetic association studies have been largely focused on SNVs due to the technical challenges of reliably identifying and genotyping SVs in the absence of whole genome sequencing (WGS) data. As such, the contribution of SVs to human health and disease remains to be fully explored. To address this, we carried out a phenome-wide association study (pheWAS) using 1.64 million SVs identified in 467,152 UK Biobank (UKB) WGS, by performing both single SV (MAC>=5) as well as a rare (AF<=0.01) SV collapsing analysis using 23,386 binary and 4,995 quantitative phenotypes that included plasma protein abundances measured on the Olink platform.We identified 3,509 statistically significant (*p*<=1e-8) SV - binary and 50,497 SV - quantitative

trait associations of which 562 and 7,393 involved rare SV (AF<=0.01). The majority (~98%) of the significant associations were located in the non-coding regions of the genome, including 11% overlapping enhancer regions. Rare SVs showed stronger effect sizes than common SVs. Most protein truncating (deletions and partial duplications), promoter and UTR *cis*-SVs were associated with a reduction in plasma protein (PP) levels in contrast to *cis*-SVs resulting in full gene duplications that were associated with increased PP abundance. Enhancer SVs showed mixed effect on PP levels.By integrating our results with SNV genetic association data, we were able to identify putatively causal SVs providing insights into disease biology. For example, a 5,235bp deletion (chr2:203,034,348-203,039,583), overlapping a cardiac pericyte and vascular smooth muscle enhancer, was associated with an increased risk of atherosclerotic heart disease (p=5.524 x 10^{-16}, OR=1.125). This SV was in strong LD with a previously reported coronary artery disease risk variant (rs148812085)[2]. Similarly, a 3' UTR deletion (chr19:47,869,503-47,871,044), of the gene *SULT2A1*, a bile salt sulfotransferase, was a protective against cholelithiasis (p=4.8x10^{-27}, OR=0.819). This SV was associated with reduced SULT2A1 plasma protein abundance and was in strong LD with protective cholelithiasis GWAS variant (rs62129966)[3].In this study, we have performed the largest pheWAS to date that encompasses the full allele frequency spectrum of SVs. This rich catalogue of robust SV associations provides additional insights into disease biology in both coding and non-coding portions of the genome, that in turn might inform therapeutic development.

Detecting large complex structural variants from human genome assemblies

Authors: P. Audano[1], D. Porubsky[2], The Human Genome Structural Variation Consortium, E. Eichler[3], C. Beck[4]; [1]The Jackson Lab., Farmington, CT, [2]Univ. of Washington, Seattle, WA, [3]Univ of Washington, Seattle, WA, [4]UConn Hlth.and The Jackson Lab., Farmington, CT

Abstract:

Human structural variants (SVs) (50+ bp) contribute to human diversity, can duplicate and delete whole genes, and may contribute to signatures of positive selection or disease associations. While most SVs are simple variants, complex SVs (CSVs) affect multiple fragments of DNA in a single rearrangement. Because CSVs contain more than one breakpoint and are often mediated by repeats, they have been difficult to identify with short-read technologies. While long-read sequencing can now resolve many CSVs to basepair accuracy, variant discovery tools still identify them as simple SVs that partially or incorrectly represent the complex event, or CSV sites are omitted altogether. As part of a global collaboration, the Human Genome Structural Variation Consortium (HGSVC), we constructed near-T2T phased assemblies for 130 phased haplotypes derived from 65

diverse individuals and created a variant call resource including SVs, indels, and SNPs. Within the assemblies, we observed that large CSVs leave distinct signatures when aligned to a reference genome, although all 10 SV callers we applied were not sensitive to them. We find that PAV's algorithm for trimming redundantly-aligned contig bases in the repeats mediating CSVs is critical for refining breakpoints. With these data, we developed an assembly-guided method for tracing CSVs through local and distal template switches and complex repeat structures. Excluding centromeres and satellite repeats, initial analyses identify 45 CSVs per haplotype ranging from 2.5 kbp to 1.2 Mbp and affecting 11 Mbp of sequence with most mediated by large segmental duplications. We find 115 genes on one haplotype that are altered by CSVs including *BAZ1B*, *GPRIN3*, *NF1*, *NOTCH2NLR*, and *TRIM50*. With high-fidelity assemblies, we can now investigate breakpoint features for signatures of the mutational mechanisms driving large CSVs. With thousands of long-read genomes now in production, high-throughput methods for accurately identifying CSVs, their breakpoints, and duplicated segments are needed to access a dynamic landscape of structural variation that was once inaccessible.

Session 70: Exploring the Genetic Spectrum of Obesity

Location: Room 501

Session Time: Friday, November 8, 2024, 10:15 am - 11:45 am

An abdominal obesity missense variant in the transcription factor and thermogenesis gene *TBX15* shows signals of adaptation to cold in Finns and affects downstream adipocyte expression in *trans*

Authors: M. Deal[1], A. Kar[1], S. T. Lee[1], M. Alvarez[1], S. Rajkumar[1], U. Arasu[2], D. Kaminska[2], V. Männistö[2], S. Heinonen[3], B. W. van der Kolk[3], U. Säiläkivi[3], T. Saarinen[3], A. Juuti[3], J. Pihlajamäki[2], M. U. Kaikkonen[2], M. Laakso[2], K. H. Pietiläinen[3], P. Pajukanta[1]; [1]Univ. of California Los Angeles, Los Angeles, CA, [2]Univ. of Eastern Finland, Kuopio, Finland, [3]Univ. of Helsinki, Helsinki, Finland

Abstract:

Abdominal obesity is predisposing to multiple cardiometabolic diseases, and it is important to elucidate biological mechanisms underlying abdominal obesity GWAS variants. We leveraged subcutaneous adipose tissue (SAT) single-nucleus RNA-sequencing (snRNA-seq) data of two independent Finnish cohorts from individuals with obesity undergoing bariatric surgery to determine cell-type-specific mechanisms of abdominal obesity GWAS variants. We first identified adipocytes to be the cell-type enriched for heritability of abdominal obesity. We then focused on an adipocyte marker gene and transcription factor (TF), *TBX15*, containing one missense abdominal obesity GWAS variant, rs10494217, predicted to be deleterious. *TBX15* is known to regulate a bulk SAT gene expression network and has previously been implicated in thermogenesis in mice. We found a large north-south decreasing allele frequency gradient and significant F_{ST} values for rs10494217 between Finns, a population with a history of genetic isolation, and 17 different global populations. The higher allele frequency of the abdominal obesity risk allele T in Finns may be an adaptation to cold due to the role of *TBX15* in thermogenesis. To further confirm this result, we performed selection analyses and found that the abdominal obesity risk allele is significantly selected for in the Finns using extended haplotype homozygosity (EHH). Because *TBX15* is a TF and rs10494217 a potentially deleterious missense variant, we searched for downstream effects of the abdominal obesity risk allele carrier status on adipocyte gene expression in *trans*. We found 13 unique adipocyte marker genes, the adipocyte expression of which is consistently affected by the risk allele carrier status in *trans* using SAT snRNA-seq data from the two Finnish bariatric surgery cohorts. Two of these *trans* genes have previously been linked to thermogenesis, which further supports

the potential role of *TBX15* in thermogenesis. Additionally, we discovered that SAT expression of one of the trans genes, a long non-coding RNA (lncRNA), *AC002066.1*, is strongly associated with adipocyte diameter, implicating it in metabolically unhealthy adipocyte hypertrophy. These results highlight a unique case of selection of an abdominal obesity GWAS variant in a TF gene in a genetically isolated population and identify downstream *trans* genes affected by the risk allele carrier status, the expression of which is related to thermogenesis and adipocyte hypertrophy.

Functional characterization of 14 obesity-associated genes using CRISPR in human white adipose tissue implicates *SLTM* as a novel lipid accumulation gene

Authors: N. Baya[1,2], &. Sur-Erdem[3,4], S. Venkatesh[1,2], S. Reibe[2], P. D. Charles[2,4], M. Claussnitzer[5], D. S. Palmer[2], C. M. Lindgren[2,1,3]; [1]Ctr. for Human Genetics, Univ. of Oxford, Oxford, United Kingdom, [2]Big Data Inst., Univ. of Oxford, Oxford, United Kingdom, [3]Nuffield Dept. of Women's and Reproductive Hlth., Univ. of Oxford, Oxford, United Kingdom, [4]Target Discovery Inst., Univ. of Oxford, Oxford, United Kingdom, [5]Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

Obesity is a common disease that significantly increases the risk of comorbidities and mortality. Developing drugs that prevent or treat obesity can reduce its disease burden. This study aimed to discover novel therapeutic targets for obesity by performing the largest-ever systematic CRISPR interrogation in human white adipose tissue (hWAT) of obesity-associated genes prioritized from a hypothesis-free exome-wide scan.
Using UK Biobank whole-exome sequencing data from 401,059 individuals, we performed rare (MAF < 0.01), damaging variant burden tests, conditioned on GWAS loci, in 18,787 genes for nine obesity-related traits. We identified 76 significant genes ($P < 4.37\times10^{-5}$, FDR < 1%). These genes had higher (one-sided *t*-test $P = 7.52\times10^{-26}$) common-variant prioritization scores compared to non-significant genes, highlighting the convergence of common and rare variant evidence.
Significant genes were prioritized for CRISPR knockout (KO) by rigorously evaluating further lines of genetic, transcriptomic, and proteomic evidence for each gene:

1. Allelic series, where a monotonic gene dosage-phenotype relationship is predictive of successful therapeutic targets.

2. Sufficient gene expression in hWAT, to justify KO as a viable experimental approach.

3. Protein-phenotype associations, to confirm that gene associations translate to the proteome.

We selected 14 genes for single-gene KO, including 11 target genes (*ABCA1, COL5A3, DENND5B, EXOC7, HERC1, MFAP5, MLXIPL, PCSK1, SLTM, TRIP10, UBR2*) and three control genes (*INSR, PLIN1, PPARG*) previously characterized for adipocyte lipid metabolism.
We measured the effect of single-gene KO on lipid accumulation in hWAT adipocytes, followed by RNA sequencing to confirm the KOs and assess differential gene expression. Single-gene KOs of six genes showed significant differences (two-sided *t*-test $P < 0.05$) in lipid accumulation: Two (*PPARG, SLTM*) showed reduced accumulation, and four (*COL5A3, EXOC7, PCSK1, TRIP10*) showed increased accumulation.
SLTM KOs, previously uncharacterized in hWAT, showed reduced lipid accumulation (fold change = 0.51, $P = 1.91\times10^{-4}$). In a longitudinal analysis, we found that carriers of predicted loss-of-function *SLTM* variants had earlier obesity onset (Cox model $P = 1.10\times10^{-3}$). These results implicate *SLTM* as a novel regulator of lipid accumulation, with gene dysfunction leading to earlier obesity onset.
Our findings underscore the importance of functional genomics in elucidating the genetic basis of obesity. Future research should investigate the druggability of these genes and the impact of their perturbation on systemic metabolic pathways.

Prioritization of effector genes within body mass index loci yields molecular insight into the biology of body weight regulation

Authors: V. Diez-Obrero[1,2], J. D. Arias[3], X. Yin[4], R. A. J. Smit[1,2,5,6], E. Marouli[7], E. P. Wilson[8], A. E. Justice[9], K. L. Young[10], M. Graff[10], M. Guindo-Martinez[1,5], Y. V. Sun[11], P. Deloukas[7], K. E. North[10], K. L. Mohlke[8], 23andMe Research Team, S. I. Berndt[3], J. N. Hirschhorn[12,13,14], R. J. F. Loos[1,2,5,6], The Million Veteran Program, The Genetic Investigation of ANthropometric Traits(GIANT) Consortium; [1]The Novo Nordisk Fndn. Ctr. for Basic Metabolic Res., Faculty of Hlth.and Med. Sci., Univ. of Copenhagen, Copenhagen, Denmark, [2]The Novo Nordisk Fndn. Ctr. for Genomic Mechanisms of Disease, Broad Inst. of MIT and Harvard, Cambridge, MA, [3]Div. of Cancer Epidemiology and Genetics, Natl. Cancer Inst., Bethesda, MD, [4]Dept. of Biostatistics and Ctr. for Statistical Genetics, Univ. of Michigan Sch. of Publ. Hlth., Ann Arbor, MI, [5]The Charles Bronfman Inst. for Personalized Med., Icahn Sch. of Med. at Mount Sinai, New York, NY, [6]The Dept. of Environmental Med. and Climate Sci., Icahn Sch. of Med. at Mount Sinai, New York, NY, [7]William Harvey Res. Inst., Barts and The London Sch. of Med. and Dentistry, Queen Mary Univ. of London, London, United Kingdom, [8]Dept. of Genetics, Univ. of North Carolina at Chapel Hill, Chapel Hill,

NC, [9]Population Hlth.Sci., Geisinger Hlth., Danville, PA, [10]Dept. of Epidemiology, Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, [11]Dept. of Epidemiology, Emory Univ. Rollins Sch. of Publ. Hlth., Atlanta, GA, [12]Div. of Endocrinology and Ctr. for Basic and Translational Obesity Res., Boston Children's Hosp., Boston, MA, [13]Program in Med. and Population Genetics, Broad Inst., Cambridge, MA, [14]Dept.s of Genetics and Pediatrics, Harvard Med. Sch., Boston, MA

Abstract:

Background and aim: The GIANT Consortium has performed a genome-wide association study (GWAS) for body mass index (BMI) including up to 5,611,938 individuals of diverse populations (based on a combination of self-identified ethnicity and genetic similarity, predominantly: European (73.0%), admixed American (13.2%), East Asian (7.7%), African (4.5%), South Asian (1.6%) and Middle Eastern (<1%)). A total of 7,141 genome-wide significant (p-value ≦ 5x10−9) signals were identified in the overall, multi-population meta-analysis. Signals were identified by conditional and joint analysis. Here, we aim to prioritize effector genes within these loci to ultimately provide new biological insight into obesity pathogenesis. **Methods:** We implemented an effector gene prioritization strategy based on the colocalization approach. We integrated summary statistics of the BMI GWAS with gene expression QTLs (eQTLs) from 47 tissues and 17 cell types (9 from the brain and 8 immune cells). Also, we generated three plasma protein QTL (pQTL) meta-analyses based on distinct proteomic platforms: ‘Olink’, ‘SomaLogic’ and ‘cross-platform’ (Olink & SomaLogic) and included brain cortex pQTLs based on mass spectrometry. We considered genes in cis (1 Mb each side) of the lead BMI SNP of each locus. For each locus, we included QTL signals having at least a SNP correlated (LD R2 ≧ 0.8) with the lead BMI SNP, and within a 500 Kb window. We determined presence of colocalization in cases where the ‘coloc’ PPH4 was ≧ 0.8. **Results:** A total of 18,513 protein-coding genes are encoded within 1Mb each side of the lead SNP of 6,952 of the 7,141 signals associated with BMI. We prioritized 747 (4%) of the genes in cis, supported by e/pQTL colocalizations in 967 (14%) of the 6,952 BMI signals, in at least one of 46 tissues and 14 cell types. We prioritized just 1 gene in 764 (79%) of the 967 signals. The prioritization of 31 genes was supported by both eQTLs and pQTLs. For example, *SLC25A12* was prioritized in the rs17288588 locus, supported by brain cortex pQTLs, brain caudate eQTLs, and arterial eQTLs. The locus with the highest number of tissues/cell types with colocalized e/pQTLs (i.e. the most ‘tissue pleiotropic’) was rs13336931, where 41 tissues pointed to *EXOSC6*. A large fraction of the prioritized genes has not been described in the context of body weight regulation, which opens the path towards discovery of novel biology. **Conclusion:** We provide a catalog of gene - BMI locus associations pointing to specific tissues/cell types within hundreds of new BMI-associated

loci. This serves as a rich resource for hypothesis generation and further functional investigation for elucidating the biology of body weight regulation.

The phenotypic variability, dose-response, and temporal effects in polygenic prediction of adiposity traits

Authors: O. Dikilitas, M. Naderian, M. Kosel, M. Hamed, D. Schaid, I. Kullo; Mayo Clinic, Rochester, MN

Abstract:

Introduction: Body-mass index (BMI) and waist-to-hip ratio (WHR) are important measures of obesity with significant polygenic heritability. We aimed to evaluate the variability, nonlinear, and temporal associations of polygenic scores (PGS) with these traits.
Methods: We derived PGS for BMI (PGS_{BMI}) and WHR adjusted for BMI ($PGS_{WHRadjBMI}$) using PRS-CS-auto on summary statistics from ~700k individuals of European ancestry (EA) in GIANT consortium and selecting ~1M common variants for each score. We conducted association analyses in the Atherosclerosis Risk in Communities, a prospective cohort including 8,631 EA individuals with anthropometric data collected across seven visits over three decades. PGS were residualized using principal components and standardized. We used ordinary least squares regression with baseline BMI (log transformed) and WHR as outcome variables and measurement age, sex, and the respective PGS as predictors. For repeated measurements, we used generalized least squares regression with serial correlation structures for residuals. Temporal and non-linear effects of PGS were evaluated using interaction terms and restricted cubic splines.
Results: The study cohort (age 54.4±5.7 years, 52.4% women, follow-up 24.5±7.6 years) had a baseline BMI of 26.4 kg/m^2 (median [IQR]; 21.6-29.6) and WHR of 0.94 (0.88-0.98). PGS_{BMI} and $PGS_{WHRadjBMI}$ explained 11% and 4% of the phenotypic variance and exhibited a linear relationship with their corresponding traits (p-value <0.0001). However, there was large phenotypic variability across the entire distribution of PGS, including the extremes (IQR for 1st and 10th deciles of PGS; BMI 19.6-28.3 and 22.6-35.9, WHR 0.79-1.03 and 0.87-1.05). One standard deviation increase in PGS_{BMI} was associated with 5.5 [95% CI; 4.8-6.3], 5.8 [5.2-6.4], 6.0 [5.4-6.5], 5.8 [5.2-6.3], and 5.3 [4.7-6.0] percent increase in BMI for ages 40, 50, 60, 70, and 80 years, respectively (p-interaction 0.008). For $PGS_{WHRadjBMI}$, one standard deviation increase corresponded to 0.018 [0.015-0.021] - 0.011 [0.009-0.014] unit increase in WHR adjusted for BMI across the same age groups, progressively decreasing with age (p-interaction 0.0004).
Conclusions: Although PGS_{BMI} and $PGS_{WHRadjBMI}$ exhibited moderately strong associations

with their respective traits, there is large phenotypic variability across the PGS distribution-consistent with multifactorial etiology of obesity. The associations of both PGS with their phenotypes have modest age-dependent differences, likely of limited clinical significance. These findings highlight limitations for polygenic prediction for obesity in the clinical setting.

Weight loss with semaglutide is influenced by traditional metabolic risk factors and BMI-associated genetic variants

Authors: M. Levy, N. Telis, L. M. McEwen, K. M. Schiabor Barrett, A. Bolze, S. White, N. L. Washington, E. T. Cirulli; Helix, San Mateo, CA

Abstract:

Introduction: Semaglutide, a GLP-1 receptor agonist, was FDA-approved in 2017 for managing type 2 diabetes (Ozempic). In 2021, its approval was extended to chronic weight management (Wegovy). Despite the drug's proven efficacy, there has been significant heterogeneity in the degree of weight loss experienced by users, with contributing factors remaining unclear.
Objective: Leveraging a multi-state network of clinico-genomic cohorts, we aimed to determine if metabolic risk factors and an established polygenic score (PGS) for body mass index (BMI) were associated with weight loss trajectories after initiating injectable semaglutide.
Methods: This study utilized clinical-grade sequencing and electronic health record data from multiple U.S. cohorts (n>100k). The BMI PGS was calculated using 27k variants (PGS001228). Among individuals starting semaglutide, we analyzed BMI measurements over up to 12 months of use. Multivariable linear mixed effects modeling was used to assess whether the rate of BMI change differed across PGS quintiles and metabolic risk factors. The model included interaction terms with time and adjusted for baseline BMI, semaglutide dose, age, sex, ancestry, cohort, calendar time, comorbidities, and concomitant insulin use.
Results: We analyzed 7,728 BMI measurements among 1,370 semaglutide users. Median baseline BMI was 37.0 kg/m^2 (IQR: 32.4-42.4) and median follow-up time was 9.0 months (IQR: 6.2-10.8). Overall, median BMI reductions after 6 and 12 months were 4.6% (IQR: 1.4-8.3) and 6.0% (IQR: 2.2-10.8), respectively. Individuals with the highest semaglutide dose (2.4 mg) had greater rates of weight loss than those with the lowest dose (0.5 mg) (β=-2.08% [per 6 months], p=0.0003). Compared to the lowest PGS quintile (≤20%), the top PGS quintile (≥80%) had a lower rate of BMI loss (β=+1.27%, p=0.002). Traditional

metabolic risk factors associated with lower rates of BMI loss included male sex (β=+1.11%, p=0.0002), type 2 diabetes (β=+0.85%, p=0.007), hypertension (β=+0.64%, p=0.042), and obstructive sleep apnea (β=+0.63%, p=0.035). Overall, 17.3% of the variation in BMI loss on semaglutide could be explained by these risk factors.
Conclusion: Findings underscore the role of traditional metabolic risk factors and BMI-associated genetic variants in modulating the weight loss response to semaglutide. Further research is needed to identify and understand the impact of specific contributing variants. This study has implications for possible advances in precision medicine for obesity.

Session 71: Long-Read Transcriptomes in Health and Disease

Location: Four Seasons Ballroom 4

Session Time: Friday, November 8, 2024, 10:15 am - 11:45 am

POISEN: A Bioinformatics Pipeline to Identify Poison Exons in Long-Read Transcriptomes

Authors: M. Broad, J. Hong, K-M. Lamar, J. Calhoun, G. Carvill; Northwestern Univ., Chicago, IL

Abstract:

Alternative poison exon (PE) splicing is a regulatory mechanism that tightly controls protein abundance with cell-type-specificity. PEs, when included in an mRNA transcript, introduce a premature termination codon, triggering nonsense-mediated mRNA decay (NMD). Previously, we identified rare pathogenic variants near PE splice sites and adjacent intronic regions in *SCN1A* in patients with Dravet syndrome. We showed that these variants cause aberrant PE splicing in mature iPSC patient-derived neurons, resulting in low protein abundance and explaining *SCN1A* haploinsufficiency and loss-of-function. Other studies have identified pathogenic variants near PE splice sites in the genes *SYNGAP1* and *FLNA* in patients with haploinsufficient neurodevelopmental disorders (NDDs), such as developmental epileptic encephalopathy, intellectual disability, autism spectrum disorder, and periventricular nodular heterotopia.
Despite their significance, PEs remain understudied due to several challenges. PE-containing isoforms are rapidly degraded by NMD, complicating the study of their natural biology and pathogenic disruptions. Likewise, short-read RNA sequencing struggles to identify the precise order of splice junctions within a transcript due to the length of the reads, hindering the ability to computationally resolve the location of a PE within a transcript. To address this, we developed POISEN (Poison exOn dIScovery for long-rEad traNscriptomes), a bioinformatics pipeline to identify PEs in long-read transcriptomes. POISEN leverages existing long-read annotation tools with customized approaches to specifically locate PEs in NMD-predicted mRNA transcripts. We sequenced day 20 and day 60 cerebral organoids using PacBio Iso-Seq to investigate PEs relevant to neurodevelopment. Our findings reveal 19,127 NMD-targeted mRNA transcripts and 1,414 PEs shared between our cerebral organoid transcriptomes and fetal human cortex, thereby underlining the relevance and translatability of our model. Additionally, these PEs were significantly enriched in genes involved in mRNA splicing (n=66; adj p-value=9.7e-08).

We will curate the PEs identified by POISEN in an online repository for the scientific community to facilitate further research into the roles of PEs in neurodevelopment and Mendelian disorders. Additionally, antisense oligonucleotides have shown to effectively target aberrant PE splicing to rescue haploinsufficiency, with clinical trials underway for *SCN1A*-related epilepsy. Therefore, our PE database will serve as a promising list of potential therapeutic targets for treating genetic NDDs and present new options for patient care.

The Spatial Atlas of Human Anatomy (SAHA) project: Unveiling cellular landscapes of health and diseases and orchestrating a new paradigm in precision medicine

Authors: J. Park[1], R. De Gregorio[1], B. Robinson[1], E. Hissong[1], S. Patel[1], F. Socciarelli[1], P. Danaher[2], E. Metzger[2], Y. Liang[2], J. Reeves[2], J. Schmid[2], R. Tecotzky[2], S. Darabi[3], H. Clifford[3], S. Kothen-Hill[3], G. VACEK[3], J. Beechem[2], M. Loda[1], O. Elemento[1], A. Alonso[1], S. Houlihan[1], R. Schwartz[1], C. Mason[1]; [1]Weill Cornell Med., New York, NY, [2]NanoString Technologies, Seattle, WA, [3]NVIDIA, Santa Clara, CA

Abstract:

The Spatial Atlas of Human Anatomy (SAHA) maps the cellular and molecular landscapes of 30 human organs from healthy adults at unprecedented spatial resolution. This collaborative project establishes best practices in experimental design, sample processing, data analysis, and data standards for high-content spatial analysis, capturing variability across genders and ancestries and providing a benchmark reference for spatial precision medicine. Initially presented at the 2023 ASHG annual meeting, this year we report the completion of Phase III and showcase the highest-ever subcellular resolution maps of cell types, lineage states, metabolic capacity, cellular neighborhoods, and ligand-receptor interactions.

SAHA encompasses the normal atlas of the immune (bone marrow, lymph node) and gastrointestinal (ileum, appendix, colon, liver, pancreas, stomach) tissues, as well as diseased samples (Crohn's disease, colorectal cancer, hepatoblastoma, pancreatic ductal adenocarcinoma). Integrating various spatial platforms, including Xenium, RNAscope, CosMx™ Spatial Molecular Imager, and GeoMx® Digital Spatial Profiler, with histopathology imaging and sequencing data, we map over 5 million cells across more than 100 individuals. By analyzing transcriptome and proteome at tissue-level and subcellular scales, we provide insights into the diversity of cellular composition and function across the normal and diseased gastrointestinal tract. We highlight distinct immune and epithelial

landscapes within each organ, potentially influenced by functional specializations and microbial interactions, such as ephrin (EFN) interactions between immune cells and epithelial cell types unique to each organ. Comparing multiple cancer samples and spatial platforms demonstrates power of SAHA data in revealing insights into organ development, homeostasis, and cancer progression. This perspective paves the way for early disease detection and personalized therapeutic interventions, laying the groundwork for precision medicine.
To increase the accessibility and usability of our extensive data, we also develop a generative AI model, spatialGPT, trained on SAHA data. It leverages deep learning algorithms to predict cell types and niche labels, infer cross-modal information, and identify novel spatial patterns associated with disease progression. All results, including data and associated models, will be available to the scientific community through the SAHA data portal. Our work provides a rich resource for understanding human anatomy and paves the way for advancements in early disease detection and personalized therapeutic interventions.

Genome-wide profiling of highly similar paralogous genes using HiFi sequencing

Authors: X. Chen[1], D. Baker[1], E. Dolzhenko[1], J. Devaney[2], J. Noya[2], A. Berlyoung[2], R. Brandon[2], K. Hruska[2], L. Lochovsky[2], P. Kruszka[2], S. Newman[2], E. Farrow[3], I. Thiffault[3], T. Pastinen[3], D. Kasperaviciute[4], C. Gilissen[5], L. Vissers[5], A. Hoischen[5], S. Berger[6], E. Vilain[7], E. Delot[6], UCI GREGoR Consortium, M. Eberle[1]; [1]PacBio, Menlo Park, CA, [2]GeneDx, Gaithersburg, MD, [3]Children's Mercy Hosp., Kansas City, MO, [4]Genomics England, London, United Kingdom, [5]Radboud Univ. Med. Ctr., Nijmegen, Netherlands, [6]Children's Natl. Hosp., Washington, DC, [7]Univ. of California, Irvine, Irvine, CA

Abstract:

Many medically relevant genes fall into segmental duplications where variant calling is hindered by the presence of highly similar paralogs, including pseudogenes. Long-read sequencing can resolve many homologous regions, but simple read-based alignment and variant calling cannot resolve extensively long and highly similar segmental duplications. We developed Paraphase, a HiFi-based informatics method that resolves highly similar paralogous genes by phasing all haplotypes of a gene family. This approach bypasses the error-prone process of aligning reads to multiple similar regions and streamlines sequence comparisons between genes within the same family. This gene-family-centered approach also enables accurate genotyping when there is a copy number difference between an individual and the reference genome, as is often the case in segmental duplications.
We applied Paraphase to 160 long (>10 kb) segmental duplication regions across the

human genome with high (>99%) sequence similarity, encoding 316 genes. Analysis of these gene families in five ancestral populations showed highly variable copy numbers, revealing a significant source of genetic diversity hidden from prior studies.
Comparisons between gene copies of the same family allowed us to discover genetic events that occur within segmental duplications. Our analysis of 36 trios identified 7 *de novo* SNPs and 4 *de novo* gene conversion events, 2 of which are non-allelic gene conversions between paralogs. We also identified 23 gene families with exceptionally low within-family diversity, indicating that extensive gene conversion and unequal crossing over have resulted in highly similar gene copies.
Finally, we highlighted 9 medically relevant gene families in segmental duplications that are considered challenging to genotype due to high sequence similarity. These include *CYP21A2* (21-Hydroxylase-Deficient Congenital Adrenal Hyperplasia), *PMS2* (Lynch Syndrome), *OPN1LW/OPN1MW* (red-green color vision deficiencies), *STRC* (hereditary hearing loss), *HBA1/HBA2* (alpha thalassemia), *SMN1/SMN2* (spinal muscular atrophy), *IKBKG* (Incontinentia Pigmenti), *GBA* (Gaucher and Parkinson's disease) and the *CFH* gene cluster (age-related macular degeneration). We demonstrated the accuracy of Paraphase by validating calls in 30 pathogenic alleles with confirmatory testing and summarized the extensive genetic diversity in these genes.
Paraphase, combined with HiFi long reads, provides a framework for resolving highly similar paralogous genes, enabling accurate testing in medically relevant genes as well as population-wide studies of previously inaccessible and less studied genes.

Combining spatial transcriptomic with snRNA-seq data enhances differential gene expression analyses

Authors: S. Tang[1], A. Buchman[2], D. Bennett[2], P. De Jager[3], J. Hu[1], J. Yang[1]; [1]Emory Univ., Atlanta, GA, [2]Rush Univ. Med. Ctr., Chicago, IL, [3]Columbia Univ Med Ctr, New York, NY

Abstract:

Background: Spatial transcriptomics (**ST**) data provide spatially-informed gene expression for studying complex diseases such as Alzheimer's disease (**AD**). Existing studies using ST data to identify genes with spatial differential gene expression (**DGE**) have limited power due to small sample sizes. Conversely, single-nucleus RNA sequencing (**snRNA-seq**) data offer larger sample sizes but lack spatial information. In this study, we combined ST and snRNA-seq data to enhance the power of spatially-informed DGE analysis of AD.
Method: We utilized the recently developed deep learning tool **CelEry** (Zhang Q. et al, Nat. Commun. 2023) to infer the spatial location of ~1.5M cells from sn-RNAseq data profiled

from the dorsolateral prefrontal cortex (**DLPFC**). Using LIBD reference ST data of brain samples, we inferred spatial locations for ~1.5M cells from the DLPFC sn-RNAseq data of 436 postmortem brains in the ROS/MAP cohorts, in six cortical layers with distinct anatomical structures and biological functions. We then conducted cortical-layer specific (**CLS**) and cell-type specific (**CTS**) DGE analyses for three quantitative AD-related phenotypes - Beta-Amyloid (**Amyloid**), Tangle density (**Tangle**), and Cognitive decline rate (**Cogdec**). We employed linear mixed models to conduct CLS and CTS pseudo-bulk DGE analyses with calibrated false positive rates, for each cell type in each cortical layer.
Results: We identified 450 CLS and CTS significant genes with p-values<10^{-4}, including 258 for Amyloid, 122 for Tangles, and 127 for Cogdec. Eight genes were shared by all three phenotypes. The majority of these findings, including genes related to AD, cannot be detected without considering spatial information. For example, *KCNIP3*, significant in oligodendrocytes-Layer-5 for all three phenotypes, encodes Calsenilin, a known AD risk factor. The well-known AD-related gene *APOE* was found significant in Layers 5 and 6 for Amyloid and in Layer 2 for Tangles in Microglia. These findings were not obtainned using traditional CTS analyses without spatial information. Interestingly, for Amyloid, Gene Set Enrichment Analyses with DGE results of the cell types and layers containing the most significant genes respectively identified 10 and 12 AD-related pathways in the Microglia of Layer-5 and Layer-6, which include 6 shared pathways.
Conclusion: Incorporating spatial information together with snRNA-seq data detected genes and pathways not identified with traditional CTS DGE analyses. This combined analysis can enhance our understanding of the biology underlying AD.

Benchmarking detection of technically challenging pathogenic variants with long-read sequencing and a head-to-head comparison with short-read sequencing in a clinical diagnostic laboratory

Authors: J. Devaney[1], J. Chong[2], J. Noya[1], A. S. Berlyoung[1], S. Yusuff[1], S. Lynch[1], R. Brandon[1], K. S. Hruska[1], L. Lochovsky[1], A. B. Stergachis[2], C-L. Wei[2], R. Kueffner[1], M. J. Bamshad[2], P. Kruszka[1]; [1]GeneDx, Gaithersburg, MD, [2]Univ. of Washington, Seattle, WA

Abstract:

Short read DNA sequencing has been widely adopted as the gold-standard for clinical diagnostic testing of individuals with Mendelian conditions. However, even with short read genome sequencing (srGS), a substantial proportion of individuals suspected of having a Mendelian condition lack a precise genetic diagnosis, in part due to pathogenic variants difficult to detect (VDDs) via short read sequencing. Long read genome sequencing (lrGS)

approaches have higher sensitivity for VDDs as shown in benchmarking and small patient cohort studies. In the present work, we systematically evaluated lrGS sensitivity to detect the full spectrum of pathogenic variants based on our clinical diagnostic pipeline and assessed its real-world performance in comparison to srGS. First, we measured lrGS sensitivity in a clinical setting. Using PacBio Revio lrGS, we analyzed 176 samples with 407 unique VDDs previously detected via Sanger, multiplex ligation-dependent probe amplification, exome/panels, triplet-primed PCR, microarrays, and/or srGS clinical assays. VDDs represented one or more of the following categories: technically challenging genes/exons, mtDNA changes, repeat expansions, mosaicism, chromosomal copy number changes including trisomies, translocations, inversions, uniparental disomy, mobile element insertions, and epigenetic changes. lrGS calls matched the clinically reported VDDs for 403/407 samples (99.0%) of the tested VDDs. VDDs missed were a mosaic trisomy 18 (~30%), two multi-exon deletions (PKD1 exons 31-33, MCOLN1 exons 1-7), and a homozygous deletion of GH1. Second, we compared diagnostic yield and implementation of srGS and lrGS. We performed Revio lrGS on >100 families with critically ill newborns prospectively ascertained as part of the SeqFirst program that had undergone clinical srGS. In a blinded study design, analysts used equivalent clinical protocols and pipelines for interpretation of srGS and lrGS. During preliminary analysis, lrGS confirmed all positive srGS findings, and spurious findings did not come up in reporting. Using more expansive analysis criteria, we prioritized new candidate variants of uncertain significance in ~1/3 of probands with srGS negative reports. We present one of the first large-scale, real-world studies to detect the full spectrum of pathogenic variants in a lrGS pipeline. Revio lrGS streamlines detection for most categories of pathogenic VDDs and can be implemented with high sensitivity and specificity. Use of lrGS exhibits promise for improving diagnostic yield, and we anticipate further gains from refining laboratory protocols and integrating improved variant callers.

A variety of molecular mechanisms cause copy number gains at 17p11.2 locus causing Potocki-Lupski syndrome: understanding patients with CNVs that do not include *RAI1 gene*

Authors: S. Pande[1], C. Grochowski[1], P. Kaur[1], H. Du[1], Z. Dardas[1], S. Jhangiani[2], L. Potocki[3], P. Hastings[1], J. Posey[1], D. Pehlivan[1], A. Lindstrand[4], C. Carvalho[5], J. Lupski[1]; [1]Baylor Coll. of Med., Houston, TX, [2]Baylor Coll. Med., Houston, TX, [3]Baylor Col Med/TX Child Hosp, Houston, TX, [4]Karolinska Inst.t, Stockholm, Sweden, [5]Pacific Northwest Res. Inst., Seattle, WA

Abstract:

Introduction: Copy number gains at the 17p11.2 locus involving the dosage sensitive *RAI1* gene cause Potocki-Lupski syndrome (PTLS) (MIM: 610883). PTLS is clinically defined by a spectrum of neurodevelopmental phenotypes and congenital anomalies. Two-thirds of individuals with PTLS carry an ~3.6 Mb recurrent duplication with an underlying mechanism of non-allelic homologous recombination (NAHR) between repeat gene clusters at 17p11.2 while the remaining 1/3 carry a non-recurrent duplication; both spanning *RAI1*. When copy number gains are detected surrounding *RAI1* but not overlapping with it, they are sometimes clinically diagnosed with PTLS without a molecular cause. **Methods:** We ascertained 14 probands with copy number gains (confirmed by array) at 17p11.2 not spanning *RAI1* with a clinical diagnosis of PTLS. We performed a combination of high-resolution array CGH (n=14), short-read whole-genome sequencing (sr-GS; n=7), long-read WGS (lr-GS; ONT; n=7 and PacBio HiFi; n=7) and optical genome mapping (n=1) on this subset to better understand 1) mechanisms, 2) refine complexities and define breakpoints and 3) understand the possible (or lack of) impact on *RAI1*. **Results:** Developmental delay/intellectual disability is a major cohort phenotype (12/14; 85.71%), other features include infantile hypotonia (5/14; 36%), epilepsy and regression of developmental milestones (2/14; 14%), behavioral difficulties (6/14; 43%), autism (3/14; 21%), intrauterine growth restriction (1/14; 7%), ophthalmic issues (6/14; 43%) and non-specific dysmorphism (4/14; 29%). Genomic complexities identified in this cohort include DUP-NML-DUP-NML-DUP (n=1), DEL-DUP (n=1), higher order amplifications (a 4X and a 6x amplification, n=2), marker chromosomes (n=3) and simple copy number gains (n=7). The higher order amplifications appear to involve identical overlapping copy number gains with distal breakpoint mapping to the polymorphic neoplasia associated isodicentric 17q breakpoint cluster, two include a region mapping to Birt-Hogge-Dube [BHD1; MIM: 135150] locus. **Conclusion:** The copy number gains at the 17p11.2 locus excluding *RAI1* provide an insight to further explore (i) the mechanisms of structural variant mutagenesis, (ii) mechanisms interfering with *RAI1* transcription and function without spanning the gene and (iii) the role of genes other than the ‘driver *RAI1* gene’ as potentially contributing to PTLS-associated phenotypes. These PTLS patients receive a provisional diagnostic designation, requiring more rigorous and systematic clinical evaluations excluding other non-17p11.2-related disorders with similarly non-specific features.

Session 72: Pharmacogenomics: DNA and Drugs

Location: Room 505

Session Time: Friday, November 8, 2024, 10:15 am - 11:45 am

Incorporation of Local Ancestry (LA) in a GWAS of warfarin dose requirement in African Americans (AAs) identifies a novel CYP2C19 Splice QTL ⭐

Authors: A. Singh[1], C. Alarcon[1], E. Nutescu[2], T. O'Brien[3], M. Tuck[4], L. Gong[5], T. E. Klein[5], D. O. Meltzer[6], J. A. Johnson[7], L. H. Cavallari[8], M. Perera[1]; [1]Northwestern Univ., Chicago, IL, [2]Univ. of Illinois Chicago, Chicago, IL, [3]George Washington Univ., Washington, DC, [4]Veterans Admin. DC Med. Ctr., Washington, DC, [5]Stanford Univ., Palo Alto, CA, [6]Univ. of Chicago, Chicago, IL, [7]The Ohio State Univ., Columbus, OH, [8]Univ. of Florida, Gainesville, FL

Abstract:

AAs have been underrepresented in genetic studies which has led to a large gap in knowledge for AAs compared to other populations. As an admixed population, AAs can inherit specific loci from either their African or European ancestor which is known as LA. Several studies have incorporated LA into GWAS to identify novel results only found when considering the patient's LA. These findings indicate that LA may be important in identifying variants that regulate phenotypes in AAs. Previously, Perera et al (2013) found a SNP located near the gene VKORC1 and SNPs near the CYP2C locus that modulate warfarin dose requirement for AAs. However, LA was not considered in that study. This analysis expands on the previous findings to help identify variants whose associations with warfarin dose requirement are modulated by LA.

A cohort of 340 AA individuals on a stable dose of warfarin, determined by INR at the time of dose, was used. Patients were genotyped using Illumina 610 Quad BeadChip and the genotypes were imputed and phased using Beagle. FLARE was used to infer LA for each sample at each SNP assuming two-way admixture. We then incorporated LA into the GWAS analysis using TRACTOR. A meta-analysis using METAL was done to combine the African-specific and European-specific estimates to get a LA-adjusted estimate for each SNP. We replicated the analysis in an independent cohort of AAs (ACCOuNT, N=309) to confirm top associations. To validate the functional role of top hits, we performed long-read RNA-sequencing from AA hepatocytes carrying each genotype for expression of CYP2C9 and CYP2C19 (>3 per genotype). We also quantified the protein expression of CYP2C9 and CYP2C19 in the hepatocytes.

We identified 6 genome-wide significant SNPs (P<5E-8) in the CYP2C locus and 50 near

genome-wide significant SNPs (P<5E-6) that were not found previously. Our top association was rs7906871 (P=3.2E-8), located between CYP2C9 and CYP2C19. None of the genome-wide significant SNPs are in LD with the top hits previously found. All significant genome-wide associations were replicated in the ACCOuNT cohort (P = 2.7-2.8E-5) with the same direction of effect. Using FUMA, we found 3 independent signals in the CYP2C locus (rs7906871, rs9332172, and rs12770901). rs7906871 was an sQTL for CYP2C19 in liver tissue in GTEx (P=6.1E-7). rs12770901 was in high LD to the previously published association, rs12777823. Using RNA-seq data in AA hepatocytes, splicing changes between exons 6 and 7 in CYP2C19 were observed for AAs who carried rs7906871. In conclusion we have found and replicated a novel CYP2C19 variant with association to warfarin dose requirement and potential functional consequences to this important enzyme.

Genome-wide association study on ACE-inhibitor switching identifies missense variants in *NTSR1* and *CACNA1H*

Authors: F. Vaura[1], T. Kiiskinen[1], J. Rämö[1,2], M. Tamlander[1], FinnGen, S. Ripatti[1,3,4]; [1]Inst. for Molecular Med. Finland (FIMM), Helsinki Inst. of Life Sci. (HiLIFE), Univ. of Helsinki, Helsinki, Finland, [2]Cardiovascular Disease Initiative, Broad Inst. of MIT and Harvard, Cambridge, MA, [3]Massachusetts Gen. Hosp. & Broad Inst., Cambridge, MA, [4]Faculty of Med., Univ. of Helsinki, Helsinki, Finland

Abstract:

Background: Persistent dry cough is the most common side effect of angiotensin-converting enzyme inhibitors (ACEi), a first-line treatment for hypertension and heart failure. Treatment guidelines advise resolving ACEi-associated cough by switching to an angiotensin receptor blocker (ARB). We aimed to identify genetic predictors for ACEi-associated cough.
Methods: We conducted a genome-wide association study (GWAS) in 84,759 first-time ACEi users (mean age 58.1 years, 44.7% female) from the FinnGen study, with medication purchase register follow-up since 1995. We defined cases (N=15,145) as ACEi users who switched to an ARB within 1 year of treatment initiation and controls (N=69,614) as those who continued using an ACEi for at least 1 year. Downstream analyses after GWAS included fine-mapping with SuSiE.
Results: We identified 11 genome-wide ($P<5\times10^{-8}$) significant loci at *KCNA2*, *NRXN1*, *OSBPL10*, *KCNIP4*, *PREP*, *MAPKAP1*, *HELLPAR*, *CACNA1H*, *L3MBTL4*, *SLCO4A1/NTSR1*, and *NTSR1*. Fine-mapping pointed to 17 credible sets (CS), including four

sets not previously reported. Notably, lead CS variants included two missense variants rs148569146-G-C (p.G301R, *NTSR1*, OR=0.39, AF=1.4%) and rs3751664-C-T (p.R788C, *CACNA1H*, OR=0.88, AF=11%). The *NTSR1* missense variant ($P=4.0\times10^{-45}$) showed 28-fold enrichment in Finns compared to non-Finnish Europeans. Lead CS variants were not associated with prevalent hypertension or any other trait in FinnGen. The relative risk for a cough diagnosis (ICD-10 or ICPC-2) within 1 year of ACEi treatment initiation was 3.6 ($P<0.001$) in cases compared to the GWAS controls. The *NTSR1* missense variant conferred adjusted OR=0.46 ($P=2.0\times10^{-4}$) for a cough diagnosis in first-time ACEi users within 1 year of treatment initiation and OR=0.90 ($P=8.8\times10^{-5}$) for a cough diagnosis in ACEi non-users (N=316,253) over lifetime follow-up.
Discussion: We identified four new loci (*NRXN1*, *OSBPL10*, *HELLPAR*, and *CACNA1H*) and two new protein-coding variants associated with a reduced likelihood of ACEi-to-ARB switching. The missense variant in *NTSR1* is particularly interesting because it is 28-fold Finnish-enriched and, to our knowledge, has not been previously reported. *NTSR1* has been linked to bronchoconstriction and airway inflammation, while *CACNA1H* contributes to nociception and itching via the Cav3.2 calcium channel. Our results provide new evidence for the neurological hypothesis of ACEi-associated cough. The protective missense variants in *NTSR1* and *CACNA1H* could potentially guide the development of cough-suppressing therapies. Further work is needed to clarify the possible overlap between our findings and the genetics of chronic cough.

Epigenetic patient stratification via contrastive machine learning refines hallmark biomarkers in minoritized children with asthma

Authors: A. Gorla[1], J. Witonsky[2], J. Elhawary[2], J. Mefford[1], J. Chen[1], J. P. García[3], S. Huntsman[2], D. Hu[2], C. Eng[2], E. Ziv[2], S. Sankararaman[1], J. Flint[1], N. Zaitlen[1], E. G. Burchard[2], E. Rahmani[1]; [1]UCLA, Los Angeles, CA, [2]UCSF, San Francisco, CA, [3]7Univ. of La Laguna, Santa Cruz de Tenerife, Spain

Abstract:

Refining patient stratification is crucial for advancing precision medicine in asthma. Several blood hallmarks, including total immunoglobulin E (IgE) levels and peripheral blood eosinophil count (BEC), are routinely used in asthma clinical practice for endotype classification and predicting drug response. However, these biomarkers appear ineffective in predicting treatment outcomes in some patients. For example, the state-of-the-art biologic drug Dupilumab only halves the risk of disease exacerbations compared to placebo over a one-year treatment period, indicating it is effective for only about half of the

potential exacerbation cases in the target population defined by these biomarkers. Furthermore, these biomarkers differ in distribution between populations, potentially compromising medical care and hindering health equity due to biases in drug eligibility. Here, we propose constructing an unbiased patient stratification score based on DNA methylation (DNAm) and utilizing it to refine the efficacy of hallmark biomarkers for predicting drug response in asthma. We developed a novel contrastive machine-learning method for patient stratification. Leveraging whole-blood DNAm from Latino (discovery; n=1016) and African American (validation; n=756) pediatric asthma case-control cohorts, we applied our method to refine the prediction of bronchodilator response (BDR) to the short-acting β2-agonist albuterol, the most commonly used drug to treat acute bronchospasm worldwide. While IgE and BEC correlate with BDR in the general patient population, remarkably, our score based on 7,662 CpGs renders these biomarkers predictive of drug response only in patients with high DNAm scores. IgE correlates with BDR in patients with above-median DNAm scores (OR for response 1.38; 95% CI [1.19, 1.60]; P=1.9e-5) but not in patients with below-median scores (OR 1.05; 95% CI [0.91, 1.20]; P=0.53); and BEC correlates with BDR in upper-quartile (OR 1.17; 95% CI [1.08, 1.28]; P=2.8e-4) but not in lower-quartile patients (OR 1.07; 95% CI [0.94, 1.22]; P=0.29). These results hold even within existing endotype classifications, including T-helper (Th)2 high asthma, confirming the DNAm score does not merely recapitulate known disease classifications. The top-weighted CpGs in the DNAm score include genes involved in asthma (*STAT3*, *RASSF1A*, *MEOX1*), BDR (*DDX54*), and lung function (*ALDH2*). Our findings suggest that the traditional asthma biomarkers IgE and BEC may be unjustified for millions. Revisiting drug eligibility criteria relying on these biomarkers could improve precision and equity in asthma treatment.

Understanding the impact of drug perturbations on disease-specific protein networks

Authors: S. Moix, M. C. Sadler, Z. Kutalik; Univ. of Lausanne, Lausanne, Switzerland

Abstract:

Understanding the interplay between diseases, drugs, and their protein targets is essential for drug design and repurposing. While current proteomics data only capture a limited portion of the human proteome, they offer valuable insights into these interactions. Leveraging plasma protein measurements, phenotypic, and medical data from electronic health records (EHR) of 54,219 UK Biobank participants, we constructed comprehensive drug-disease-protein triplets. To estimate bidirectional causal relationships between 2,045

proteins and key health markers (e.g., serum lipids, glycemia values, and blood pressure), we used inverse-variance weighted (IVW) Mendelian randomization (MR). Compared to simple disease-protein correlation, MR allowed us to identify the direction and magnitude of the effects, with discrepancies highlighting unmeasured confounding in the correlation estimates. Focusing on low-density lipoprotein (LDL) as a benchmark, we found that 42 proteins affected LDL levels (e.g., PCSK9: α_{ivw} = 0.303; p = 7.3e-90 vs β = 0.099; p = 2.4e-74), with 6 of these links (e.g., A2AP-LDL) being bidirectional. Next, to obtain causal estimates between drugs and diseases, we used longitudinal EHR data. By comparing pre-treatment and post-treatment measures in patients, we determined the effect of drugs on diseases (e.g., average reduction of 1.5 mmol/L LDL for statins). Reversely, to estimate the disease effect on treatment initiation, we compared pre-treatment averages of treated individuals to drug-naive individuals (e.g., LDL average difference of 0.5 mmol/L). Finally, to estimate the effect of drugs on protein levels, we compared protein levels between treated and drug-naive individuals. As pre-treatment differences in protein levels may confound results, we compared our proteomic findings with cell line perturbation experiments. For statins, Connectivity Map perturbation data showed a concordant direction of effects for genes reported in the curated Comparative Toxicogenomics Database (r = 0.341; p = 4.2e-9), but only a weak correlation (r = 0.037; p = 0.035) with protein-drug associations. While discrepancies may reflect biological differences, we will correct drug-protein correlational estimates by determining the confounding effect of the disease. Identifying the drug-to-disease mediation via proteins, we find, e.g., that statins indirectly reduce LDL by 0.024 mmol/L (p = 7.8e-19) through ANGPTL3. Overall, our study leverages state-of-the-art drug response, proteomics, and disease genetics data to provide a comprehensive understanding of the complex interplay between drugs and disease-specific protein networks.

Prioritization of icosapent ethyl for the potential reversal of metabolic dysfunction associated fatty liver disease using a genetically informed drug repurposing pipeline ⭐

Authors: H. Seagle[1,2], N. K. Khankari[3], A. P. Akerele[3,4], J. N. Hellwege[3,5], M. M. Shuey[3], M. Levin[6], K. Lee[7], J. S. Lee[8], K. Heberer[9], D. R. Miller[10,11], P. Reaven[12], K-M. Chang[6,13], J. A. Lynch[7], T. L. Edwards[3,5], M. Vujkovic[6]; [1]Vanderbilt Univ., Nashville, TN, [2]Atlanta VA Med. Ctr., Atlanta, GA, [3]Vanderbilt Univ. Med. Ctr., Nashville, TN, [4]Meharray Med. Coll., Nashville, TN, [5]VA Tennessee Valley Hlth.care System, Nashville, TN, [6]Univ. of Pennsylvania, Philadelphia, PA, [7]Salt Lake City VA Med. Ctr., Salt Lake City, UT, [8]Stanford Univ., Stanford, CA, [9]Palo Alto VA Med. Ctr., Palo Alto, CA, [10]Univ. of Massachusetts, Lowell, MA, [11]Dept. of

VA Chicago, Chicago, IL, [12]Phoenix VA Hlth.Care System, Phoenix, AZ, [13]Corporal Michael J. Crescenz VA Med. Ctr., Philadelphia, PA

Abstract:

Metabolic dysfunction-associated steatotic liver disease (MASLD) affects over 100 million American adults. While weight loss can help reduce liver fat content, it may not resolve fibrosis caused by metabolic dysfunction-associated steatohepatitis (MASH). The recent FDA approval of Rezdiffra, a thyroid hormone receptor beta agonist, marks the first drug approved to treat MASH, invigorating drug development in this area. Drug targets with support from genome-wide association studies (GWAS) are greater than two times as likely to be approved in clinical trials than those without. Here we present a genetically-informed multi-stage drug-repurposing pipeline designed to prioritize existing drug targets for potential MASLD reversal. Our initial dataset included 90,408 MASLD cases and 128,187 controls from a large multi-ancestry Million Veteran Program GWAS. For each gene, genetically predicted gene expression (GPGE) profiles were generated from 47 GTEx v7 tissues using S-PrediXcan. Drug-gene pairs were identified using directional gene-drug mapping in publicly available databases (Open Targets, Drug Gene Interaction Database). Mendelian randomization (MR) analysis was used to estimate the genetic effect of candidate genes on alanine aminotransferase (ALT) levels to proxy therapeutic effects. We identified 212 significant genes, 81 in the MASLD GWAS and an additional 131 using GPGE. Of these genes, 13 encoded protein targets for 81 drugs and we estimated genetic effects consistent with beneficial pharmacological treatment. For these 81 drugs, we obtained GWAS summary statistics for 19 primary indications. Ultimately, 14 were excluded due to non-significance, yielding seven final drug targets with five primary indications. The effect of proxied icosapent ethyl (IPE) on ALT level was striking, a 1.2-1.4 units/L reduction. The triglyceride-lowering effect of IPE was proxied using one standard deviation change (either increase or decrease) gene expression at three different gene targets within the FADS locus. However, statistically significant effects were only observed when the triglyceride-lowering effect of IPE was proxied via increased *FADS1* expression and decreased *FADS2* expression ($p = 6.8ex10^{-160}$, IVW beta = -1.26 and $p = 3.4x10^{-127}$, IVW beta = -1.39, respectively). Using genetic information from the largest MASLD GWAS study to date, we identified IPE as a possible genetically supported drug-repurposing target. Our analysis suggests that reducing triglycerides via IPE as proxied by genetically-predicted *FADS1* and *FADS2* expression, may help lower ALT with reduced liver injury. Further exploration of IPE may be warranted.

Combining genetics with real-world patient data enables ancestry-specific target identification and drug discovery

Authors: F. Cheng[1], Y. Hou[1], N. Lorincz-Comi[2], J. Leverenz[1], J. Haines[3], J. Cummings[4]; [1]Cleveland Clinic, Cleveland, OH, [2]Cleveland Clinic Fndn., Cleveland, OH, [3]Case Western Reserve Univ., Cleveland, OH, [4]Univ. of Nevada Las Vegas, Las Vegas, NV

Abstract:

Background: Although high-throughput DNA/RNA sequencing technologies have generated massive genetic and genomic data in human disease, translation of multi-omics findings into new patient treatment has not materialized by lack of effective approaches.
Method: To address this problem, we have used Mendelian randomization (MR), fine-mapping, and colocalization analysis of large patient's genetic and real-world patient data to evaluate druggable targets using Alzheimer's disease (AD) as a prototypical example. We utilized the genomic instruments from 9 expression quantitative trait loci (eQTL) and 3 protein quantitative trait loci (pQTL) datasets across five human brain regions from three biobanks. We tested the outcome of Mendelian randomization across genome-wide association studies (GWAS) datasets of European-American (EA) and African-American (AA) ancestries, with 275,540 AD cases and 1.55 million controls. We searched EA- and AA-specific repurposable drugs using propensity-scored matching drugome-wide association studies from ~80 million electronic health records.
Result: We identified 25 drug targets in EAs and 6 new drug targets in AAs. Among 6 AA-specific targets, TRPV3 is a potent drug target and replicated in AA-specific eQTL data from the Metabrain cohort. We pinpointed that an anti-inflammatory AD target of epoxide hydrolase 2 (EPHX2): (1) a pQTL lead SNP rs2741342 ($P_{GWAS} = 5.72\times10^{-13}$; $P_{pQTL} = 1.19\times10^{-16}$) located in an enhancer of *EPHX2*; (2) a protein-coding variant of rs751141 (p.Arg287Gln) was associated with reduced EPHX2 protein expression ($P_{pQTL} = 5.50\times10^{-16}$) from the AD knowledge portal, and (3) we experimentally validated that p.Arg287Gln on EPHX2 is associated with reduced p-Tau expression in patient iPSC-derived neurons. We demonstrated that TPPU blocked deterioration in hippocampal-dependent cognitive ability in a TgF344-AD rat model and EC5026 improves cognition in a 5xFAD mouse model. We identified 23 candidate drugs associated with reduced risk of AD in mild cognitive impairment patients. We found that usage of either apixaban (hazard ratio [HR] = 0.74, 95% confidence interval [CI] 0.69-0.80) and amlodipine (HR = 0.91, 95%CI 0.88-0.94) were significantly associated with reduced progression to AD.
Conclusion: Combining genetics and real-world patient data identifies ancestry-specific therapeutic targets and medicines for AD and other human diseases if broadly applied.

Further functional and clinical validation of candidate targets and drugs in ethnically diverse population are warranted.

Session 73: Stats Just Wanna Have Fun: New Methods in Statistical Genetics

Location: Four Seasons Ballroom 2&3

Session Time: Friday, November 8, 2024, 10:15 am - 11:45 am

Integrative Statistical Framework for Detecting Divergent Selection and Linking to Disease

Authors: C. Sottolano[1], G. Shaw[1], T. Mosbruger[1], Y. Duan[2], A. Allen[2], D. S. Monos[1,3], T. J. Hayeck[1,3]; [1]Children's Hosp. of Philadelphia, Philadelphia, PA, [2]Duke Univ., Durham, NC, [3]Univ. of Pennsylvania, Philadelphia, PA

Abstract:

Over the course of evolution, environmental pressures (like exposure to pathogens) have left discernible patterns of genetic variation across different human populations. Existing methods of detecting these patterns excel in settings of hard sweeps, where selective pressure is strong enough to drive specific haplotypes to fixation. However, they are less effective in detecting soft sweeps, which are thought to be frequent and potentially associated with many diverse phenotypes. Soft sweeps may result in more subtle genomic signatures, which can be difficult to detect with existing methods. We developed a new test statistic for detecting divergent balancing and positive selection, primarily focused in settings of soft sweeps. The test detects regional differentiation in the site frequency spectrum, linkage disequilibrium, and density of polymorphisms across populations to characterize signatures of soft selective sweeps. We used a forward time demographic model to simulate the out-of-Africa migration to carry out a sensitivity analysis. Our method showed a performance increase of 18% in AUC and 16% in precision at a false discovery rate of 0.05 in certain settings compared to the leading methods.
This study leverages the latest release of 1000 Genomes Project (1KGP), high coverage whole genome sequencing with hundreds of complete trios for improved phasing, to characterize genetic diversity and evolutionary selection across and within world populations. We find strong evidence of population specific differentiation was mapped throughout the 1KGP both at the super and sub-populations levels. We linked selection signal in 1KGP to summary statistics from large-scale genome-wide association studies (GWAS), matching ancestral group in both. This ties potential divergent evolution to corresponding genes and diseases in specific populations. Divergent selection in and around the *BTNL2* gene was seen in East Asians at variants tied to both liver and lung function. Whereas in Africans strong divergent selection signal is observed in and

around *ALG10B* at variants related to cardiomyopathy, along with signal in *HLA-C* and *SPOCK3* related to immune response. Interestingly, consistent patterns enrichment of selection signal across populations were observed in both unprocessed pseudogenes and polymorphic pseudogenes pointing to potential biological functions favoring diversification and development of novel genes. Our results reveal both unique and shared patterns of evolutionary selection across global populations, highlighting their implications of genetic variation on disease phenotypes.

Genome-Wide Assessment of Pleiotropy Across >1000 Traits Among >1.5 Million Participants of Diverse Biobanks ⭐

Authors: M. Levin[1], S. Koyama[2], R. Bhukar[2], D. Zhang[3], J. Woerner[1], B. Truong[4], A. Rodriguez[5], R. Madduri[6], B. Voight[1], S. Damrauer[1], P. Natarajan[7]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Broad Inst., Cambridge, MA, [3]Perelman Sch. of Med., Philadelphia, PA, [4]Harvard / Broad Inst., Cambridge, MA, [5]Argonne Natl. Lab., Lincolnwood, IL, [6]Argonne Natl. Lab., Naperville, IL, [7]Massachusetts Gen. Hosp., Boston, MA

Abstract:

Large-scale genetic association studies have identified thousands of trait-associated risk loci, establishing the polygenic basis for complex traits and diseases. While prior studies suggest that the majority of trait-associated loci are pleiotropic, the extent to which this pleiotropy reflects shared causal variants or confounding by linkage disequilibrium remains poorly characterized. To define a set of candidate loci with potentially pleiotropic associations, we performed genome-wide association study (GWAS) meta-analyses of up to 1,121 traits/diseases across diverse populations similar to Admixed American (AMR, N_{Max} = 51,326), African (AFR, N_{Max} = 128,012), East Asian (EAS, N_{Max} = 295,011), and European (EUR, N_{Max} = 1,289,246) reference populations across among up to 1,632,720 participants of the VA Million Veteran Program (MVP), UK Biobank (UKB), FinnGen, Biobank Japan (BBJ), Tohoku Medical Megabank (ToMMO), and Korean Genome and Epidemiology Study (KoGES). We identified 26,946 genome-wide significant loci (1MB region with $P_{GWAMA} < 5 \times 10^{-8}$) in intra-population analysis and 25,865 in multi-population analysis ($P_{MR\text{-}MEGA} < 5 \times 10^{-8}$). Among these, 12% (n = 3,114) of loci in population-wise and 9% (n = 2,324) in multi-population analyses did not reach genome-wide significance in any of the original summary statistics. In aggregate, the genome-wide significant loci fell within 2,824 non-overlapping genomic windows on average ~600kb in size, each containing genome-wide significant signals for a median of 6 traits (IQR 2 to 15), including 2,235 (79%, 95% CI 78 to 81%) pleiotropic regions associated with >1 trait. Multi-trait colocalization was performed

using HyPrColoc to identify clusters of traits sharing putative causal variants at these loci, identifying 1,535 loci (64%, 95% CI 63 to 66%) with high-confidence (posterior probability > 0.7) evidence of a shared causal variant across 2 or more traits. Pleiotropic causal variants were enriched (OR 2.9, 95% CI 2.2 to 3.7, $p = 3 \times 10^{-12}$) in the most constrained non-coding regions of the genome (Gnocchi score >= 4), but compared to lead GWAS variants, were not more enriched for protein-altering genetic variation (OR 0.98, 95% CI 0.87 to 1.11). These results provide a contemporary map of genetic pleiotropy across the spectrum of human traits/diseases and diverse genetic backgrounds.

Pleiotropic heritability quantifies the shared genetic variance of common diseases

Authors: Y. Zhao[1], A. L. Price[2], X. Jiang[3]; [1]Fudan Univ., Shanghai, China, [2]Harvard Sch Pub Hlth., Boston, MA, [3]Cambridge Univ., Cambridge, United Kingdom

Abstract:

Common diseases are highly pleiotropic (Watanabe et al. 2019 *Nat Genet*), but the overall contribution of pleiotropy to disease architectures is unknown, as most studies focus on genetic correlations with each auxiliary disease in turn (Lee et al. 2013 *Nat Genet*, Bulik-Sullivan et al. 2015b *Nat Genet*). Here, we propose to quantify the genetic variance of a target disease that is shared with a specific set of auxiliary diseases. We define pleiotropic heritability (h^2_{pleio}) as the proportion of target disease variance explained by any linear combination of the genetic values of the auxiliary diseases. We estimate h^2_{pleio} from the genetic covariance matrix of the target and auxiliary diseases, accounting for noise in covariance estimates via genomic block-jackknife. We compare h^2_{pleio} to h^2, restricting all computations to the genetic variance explained by SNPs. Our method produces unbiased estimates and requires only summary statistics as input. We estimated h^2_{pleio} for 15 highly heritable target diseases from the UK Biobank (avg N=230K), with respect to the same 15 auxiliary diseases; in all analyses, we excluded auxiliary diseases with high genetic correlation to the target disease ($r_g^2 > 0.5$). We reached 3 main conclusions. First, h^2_{pleio}/h^2 is generally large (avg = 47%, jackknife s.e.(avg) = 6%, s.d. = 24% across diseases); several estimates were close to 100%, e.g. 81% (s.e. 10%) for depression (MDD) and 74% (s.e. 15%) for type 2 diabetes (T2D). Estimates were little changed when also including 16 quantitative auxiliary traits (avg = 53%). Second, h^2_{pleio}/h^2 is broadly distributed across disease systems, defined according to 7 Phecode categories. Due to shared pleiotropy across disease systems, h^2_{pleio}/h^2 decreased only slightly when removing all auxiliary diseases from same category as the target disease (avg = 43%; 79% for MDD (removing mental) and 72% for T2D

(removing endocrine)), and even when further removing one other category whose removal had the greatest impact (avg = 31%; 67% for MDD (removing mental and endocrine) and 57% for T2D (removing endocrine and respiratory)); other disease systems contributing substantial h^2_{pleio}/h^2 (> 20% when considering each disease system individually) included digestive for MDD, and mental and circulatory for T2D, consistent with known biology (Eijsbouts et al. 2021 *Nat Genet*, Suzuki et al. 2024 *Nature*). Third, h^2_{pleio}/h^2 is 1.3 larger than the analogous quantity V^2_{pleio}/V^2 quantifying pleiotropic total variance (avg = 38%, s.e.(avg) = 10%, s.d. = 17%). We conclude that roughly half of common disease heritability is pleiotropic, primarily with auxiliary diseases from a broad range of unrelated disease systems.

ENCODE cCRE-based WGS analysis of 100 traits in UK Biobank identifies 1,987 associations driven by rare-variants

Authors: J. Flanagan, S. Lee; Seoul Natl. Univ., Seoul, Korea, Republic of

Abstract:

The release of large scale whole-genome sequencing (WGS) data can provide novel insights into the role of non-coding rare variants in complex traits. Previous association analyses of WGS data have utilised a sliding-window approach to test sets of rare-variants. However, this non-targeted approach may overlook distinct biological features within the genome removing contextual information unique to specific regions. Our study employed approximately 1 million candidate Cis-Regulatory Elements (cCREs) identified in the ENCODE Screen Registry v3, representing 7.9% of the GRC38 genome, to define rare-variant sets. The sizes of cCRES range from 200 to 300 bp, and SAIGE-GENE+ was used for region-based testing. A total of 100 traits (60 Continuous and 40 binary traits) were analysed. Analysis of the 150K White-British individuals of the WGS discovery set identified 1,987 associations reaching genome-wide significance ($P = 5.0 \times 10^{-8}$), implicating 1,334 unique cCREs. Subsequently, 88% of these associations were successfully replicated in a separate cohort of the remaining 250K White-British WGS samples of the 500K WGS release. The associations were enriched to promoter-like signals (PLS) and proximal enhancer-like signals (pELS). PLS and pELS comprised 18.2% and 21.3% of associations, respectively, while they represent 3.8% and 15.3% of total cCREs. Additionally, we conducted comprehensive conditional analyses for 230 cCREs across five traits to identify rare-variant associations independent of common variant associations previously identified in UK Biobank imputed datasets, resulting in 68 genome-wide significance associations.

By leveraging data on enhancer/promoter-gene interactions and key epigenetic markers across over 1,500 cell types, our analysis provides a better understanding of the role of rare variants in complex traits. Our leukaemia case study illustrates the method's potential, highlighting key loci including cCREs EH38E3244983 ($P = 7.7 \times 10^{-20}$), a promoter-like signal for *SRSF2*, and EH38E2266769 ($P = 1.61 \times 10^{-16}$), a proximal enhancer-like signal for *BCL6*. These observations demonstrate that cCRE-based analysis is not only effective at identifying associations driven by rare variants but also facilitates deeper biological insights.

Enhancing regulatory variant prioritization via long-range DNA sequences and multi-task learning

Authors: Y. Takahashi[1], **Q. Wang**[2], Y. Okada[3], Japan COVID-19 Task Force; [1]Osaka Univ., Suita, Japan, [2]The Univ. of Tokyo, Bunkyo-ku, Japan, [3]The Univ. Tokyo / Osaka Univ. / RIKEN, Tokyo, Japan

Abstract:

Identifying causal variants from associated regions obtained through GWAS is crucial for understanding disease mechanisms and developing new therapeutic strategies. However, prioritizing causal variants in a regulatory region of the genome remains challenging, due to intensive linkage disequilibrium.
Here, we developed the Expression Modifier Score (EMS) v2, a deep learning model designed to enhance the prioritization of regulatory variants with high precision. EMSv2 is built using GTEx v8 dataset and predicts the effects of variants on gene expression across 49 tissues. EMSv2 achieves higher prediction performance compared to alternative methods (Chen, KM. et al., 2022, Wang QS. et al., 2021), especially in minor tissues such as brains (e.g., AUPRC<0.001 to >0.05 in brain spinal cords). The power gain is attributed to implementation of (1) features accounting for long-range DNA sequence interaction using Enformer (Avsec, Ž. et al., 2021), (2) customization of loss-function in model training, and (3) multi-task learning framework.
We further validated the efficacy of EMSv2 utilizing multiple datasets, including the UK Biobank and the Japan COVID-19 Taskforce (JCTF; extended from Wang QS. et al., 2022). First, we showed that EMSv2 presents high prediction performance (AUPRC=0.71, AUROC=0.91) in prioritizing putative causal whole blood eQTLs identified from JCTF, highlighting its utility across genetically diverse populations. Second, we demonstrated that EMSv2 is a powerful resource for complex trait-causal variant prioritization. Combining EMSv2 with the gene-level Polygenic Priority Score (PoPS [Weeks EM et al., 2023]) resulted

in over 3x enrichment of causal variants compared to using EMSv2 or PoPS alone. Finally, we provide rich examples where EMSv2 prioritizes likely-causal variants in gene expression and complex trait associated loci, allowing for downstream functional interpretation (e.g., rs61835060 on *SFMBT2* and rs73050371 on *EML2*).
Overall, our study presents a powerful resource for regulatory variant prioritization in diverse populations and contributes to the elucidation of the mechanisms underlying the impact of noncoding variants on gene regulation. We also release EMSv2 through an interactive browser (https://japan-omics.jp/).

Trans-modeling of large-scale proteomics data uncovers enriched protein-protein interactions and drug targets

Authors: Z. Zhang[1], J. Liu[2,1], L. Wu[3], B. Zhao[4], C. Wu[1]; [1]UT MD Anderson Cancer Ctr., Houston, TX, [2]Rice Univ., Houston, TX, [3]Univ. of Hawaii Cancer Ctr., Honolulu, HI, [4]Univ. of Pennsylvania, Philadelphia, PA

Abstract:

Proteome-wide association studies (PWAS) have emerged as a powerful approach to identify putative causal proteins for complex traits and diseases. However, traditional PWAS predominantly focus on *cis*-acting elements and ignore in-depth biological knowledge, limiting their ability to identify putative causal proteins for complex traits and diseases. To address these limitations, we developed a novel framework to train PWAS imputation models that integrates summary-level protein quantitative trait loci (pQTL) data from both *cis*- and *trans*-loci across the genome and leverages protein-protein interaction networks. Applying our approach to extensive pQTL data from the UK Biobank Pharma Proteomics Project (UKB-PPP; n = 46,218 for European and n = 931 for African ancestries) and deCODE genetics (n = 35,892 for European ancestry), we trained large-scale PWAS models for both ancestries and prioritized biologically relevant protein networks for 620 and 642 proteins, respectively. Compared to classic *cis*-only models, the resulting models showed significantly improved predictive performance (1,796 versus 1,267, 42% more models with estimated predictive $R^2 \geq 0.01$). Applying our models to GWAS summary statistics from the FinnGen, IEU OpenGWAS, GBMI, and MVP projects, we conducted a systematic multi-ancestry PWAS analysis for over 700 phenotypes. Compared to classic *cis*-only PWAS models, the resulting models showed significantly improved predictive performance and much higher power in sequential association studies (7,270 versus 1,650, 341% more associations found), enabling the identification of numerous novel protein-trait associations. Notably, using an external dataset of 6,690 FDA-approved

drugs, we demonstrated that associations identified by our method are 2.4 times and 1.3 times more likely more likely to be validated for drug targets than those identified using Mendelian Randomization and cis-only PWAS, respectively. Our PWAS framework and multi-ancestry models are freely available on our website, facilitating the discovery and characterization of the proteomic architecture of complex traits and diseases.

Session 74: The Non-coding Genome: From Nucleotide to Protein

Location: Four Seasons Ballroom 1

Session Time: Friday, November 8, 2024, 10:15 am - 11:45 am

Interpreting Regulatory Differences between Species in Terms of Potential Cis- and Trans- Mechanisms

Authors: Y. Duan, S. Li, Z. Mielko, R. Gordan, G. Wray, G. Crawford, A. Allen; Duke Univ., Durham, NC

Abstract:

Non-coding regulatory regions are essential to human traits, development and health. Analyzing variation in these regions can help us understand the regulatory mechanisms and identify regions and variants associated with disease. Comparative multi-omic analyses across near-human primates are particularly effective in understanding the variation due to the large number of variants observed across species while remaining highly comparable to humans.

In this study, we focused on functionally characterizing how cis and trans non-coding variation affects chromatin accessibility. We jointly analyzed DNase-seq/ATAC-seq, whole genome sequencing (WGS), and RNA-seq across five primate species: human, chimpanzee, gorilla, orangutan, and macaque. Unlike studies using hybrids or fused cells to dissect cis and trans effects, our approach relied on computational analyses. These studies using allele-specific expression (ASE) to partition cis and trans regulatory components do not provide mechanistic interpretation. In contrast, our approach explains the cis vs trans mechanisms leading to regulatory effects via transcription factor (TF) binding effects and abundance.

To minimize bias while maximizing data retention, we developed a cross-species analysis pipeline. This pipeline includes cross-species and cross-replicate alignment, and functionality estimates for SNVs and indels. We also developed methods to partition the variation in chromatin accessibility into cis and trans effects. Our pipeline along with our variance partitioning techniques, is adaptable for other multi-omic datasets from various species.

For cis effects, we focused on the TF binding changes due to variation (including indels). We identified TFs positively associated with accessibility, such as JUN and FOS, and those negatively correlated. For trans effects, we analyzed TF abundance using RNA-seq data. Principal Component Analysis (PCA) was used to simplify the analysis of both effects. This

was followed by canonical correlation analysis (CCA), orthogonalization, and linear modeling with precision weighting to partition the variation into cis-only, trans-only, and shared effects.
We identified species-specific differential regions and ubiquitously open regions. These regions were intersected with the evolutionary constraint, conservation scores and with cis and trans effects partitioning. Human specific open regions are more constrained than conserved and show increased explainability in terms of both cis and trans effects, suggesting recent gain-of-function regions might tend to have a less complex regulatory architecture.

Nanopore sequencing of chromatin accessibility

Authors: P. James[1], H. Jeffery[1], G. Dodd[1], A. Rand[1], M. Stoiber[1], P. Rescheneder[1], S. Juul[2]; [1]Oxford Nanopore Technologies, Oxford, United Kingdom, [2]Oxford Nanopore Technologies, New York, NY

Abstract:

Chromatin packaging is a key factor determining gene expression patterns via the regulation of genome accessibility to elements such as transcription factors and other DNA binding proteins. Chromatin structure dysregulation can affect gene expression and has been associated with oncogenesis as well as diseases such as ATR-X syndrome. Identifying regions of open chromatin in disease systems helps identify the epigenetic drivers, differentiators, and progression of different disease phenotypes. Several methods, previously described, aim to identify regions of open chromatin, such as ATAC-seq, DNAse-seq, MNase-seq, FAIRE-seq, or more recently SMAC-seq and Fibre-seq. Nanopore sequencing of long native DNA reads provides an opportunity to identify regions of open chromatin, in a non-destructive manner, by enzymatically marking these areas with exogenous DNA modifications prior to detection during base calling.
Here we present a footprinting approach using a commercially available non-specific 6mA methyltransferase (EcoGII), to label open chromatin. By treating isolated nuclei with EcoGII, the chromatin structure is preserved without the need for cross-linking. Meta-analysis of biological sites, including CTCF binding sites, transcription start sites, and DNaseI hypersensitive sites shows characteristic chromatin structures. Meanwhile, the detection of SNPs, indels, SVs, and 5mC is unaffected by the addition of the 6mA and the performance is comparable with a standard nanopore sequencing run. In summary, we can add an extra layer of information, chromatin structure, to an already information-rich nanopore sequencing dataset. Furthermore, combining native 5mC methylation and haplotyping by germline SNP analysis on the same dataset gives haplotype resolved

measures of chromatin accessibility end epigenetic silencing, without the need for a priori information of regulatory element location.

BRAIN-MAGNET: A novel functional genomics atlas coupled with convolutional neural networks facilitates clinical interpretation of disease relevant variants in non-coding regulatory elements

Authors: S. Barakat[1], R. Deng[1], E. Perenthaler[1], A. Nikoncuk[1], K. Lanko[2], M. Maresca[1]; [1]Erasmus MC - Univ. Med. Ctr., Rotterdam, Netherlands, [2]Erasmus MC Univ. Med. Ctr., Rotterdam, Netherlands

Abstract:

Genome-wide assessment of genetic variation is becoming a routine in human genetics, but functional interpretation of non-coding variants both in common and rare diseases remains extremely challenging. Here, we employed the massively parallel reporter assay ChIP-STARR-seq to functionally annotate activity of >140 thousand non-coding regulatory elements (NCREs) in human neural stem cells (NSCs) as a model for early brain development. Highly active NCREs show an increasing sequence constraint and harbor *de novo* variants in individuals affected by neurodevelopmental disorders. They are enriched for transcription factor (TF) motifs including YY1 and p53 family members and for the presence of primate-specific transposable elements, providing insights on gene regulatory mechanisms in NSCs. Examining episomal NCRE activity of the same sequences in human embryonic stem cells (ESCs) identified cell type differential activity and primed NCREs, accompanied by a rewiring of the epigenome landscape. Leveraging on the experimentally measured NCRE activity and nucleotide composition of the assessed sequences, we build BRAIN-MAGNET, a convolutional neural network that allows the prediction of NCRE activity based on DNA sequence composition, and which identifies functionally relevant nucleotides and TF motifs within each NCRE that are required for NCRE function. The application of BRAIN-MAGNET including its functional validation allows fine-mapping of GWAS loci identified for common neurological traits, and prioritization of possible disease causing rare non-coding variants in currently genetically unexplained individuals with neurogenetic disorders, including those from the Genomics England 100,000 Genomes project. This includes the discovery of novel enhanceropathies caused by noncoding single nucleotide variants, as validated using induced pluripotent stem cell based disease modelling. We foresee that this NCRE atlas and BRAIN-MAGNET will help reducing missing heritability in human genetics, by limiting the search space for functional relevant non-coding genetic variation.

Variable number tandem repeats (VNTRs) regulate epigenome and transcriptome in human prefrontal cortex

Authors: C. Dillard[1], K. Girdhar[2], B. Zeng[2], J. Bendl[1], J. Fullard[2], P. Garg[1], M. Shadrina[1], A. Sharp[1], G. Hoffman[1], P. Roussos[1]; [1]Icahn Sch. of Med. at Mount Sinai, New York, NY, [2]Mount Sinai, New York, NY

Abstract:

To date, most studies have investigated only the impact of single-nucleotide polymorphisms (SNPs) on molecular phenotypes, such as gene expression generated from human postmortem brain tissue to identify quantitative trait loci (QTLs). Additionally, genome-wide association studies (GWAS) have predominantly focused on SNPs, which do not capture the impact of other, complex structural genetic variants. To address this gap, our research specifically examines the impact of variation in length of variable number tandem repeats (VNTRs) on gene expression and chromatin accessibility phenotypes generated from the human postmortem brain. To identify expression VNTRs (eVNTRs) and chromatin accessibility VNTRs (caVNTRs), we first identified VNTRs (motif size ≥ 10 bp and width ≥ 100 bp) using large-scale whole-genome sequencing data from post-mortem brain tissue of 989 donors, resulting in the identification of approximately 90,000 VNTRs. From this dataset, we identified approximately 2,000 eVNTRs using gene expression data from a subset of 648 donors in two cortical regions (prefrontal cortex and anterior cingulate cortex). Additionally, using chromatin accessibility data from the same cohort of donors and cortical regions, we identified approximately 4,000 caVNTRs in neuronal and non-neuronal cells. Notably, only 20%(9%) of eVNTRs(caVNTRs) were located in promoter regions while the remaining significant proportion was located in non-coding regions. Integration of expression and chromatin accessibility association statistics of neighboring SNPs with eVNTRs(caVNTRs) summary statistics in fine-mapping test identified ~ 200(200) eVNTRs(caVNTRs) as causal variants (pip >0.3) highlighting the relevance of VNTRs over SNPs in these regions. To further assess the disease relevance of these eVNTRs/caVNTRs, we imputed genome-wide risk VNTRs for various neurological and psychiatric traits by leveraging the correlation between VNTR copy numbers and neighboring SNPs. This analysis identified dozens of genome-wide significant VNTRs associated with neurological traits. Further colocalization of eVNTRs/caVNTRs with the combined VNTR and SNP GWAS summary statistics, found that many of these VNTRs colocalized with neurological and psychiatric disease associated risk variants. For example, in Alzheimer's disease, we found ~20 VNTRs colocalized with risk eVNTRs that are proximal to the chromosome 17 *MAPT*-inversion locus. Finally, our results provide further evidence that the association of VNTR

copy numbers have functional effects in the genome specifically in the case of neurological diseases.

CRISPRi perturbation screens and eQTLs capture different target genes for non-coding GWAS variants

Authors: S. Ghatan[1], W. Oliveros[1,2], N. E. Sanjana[1,3], J. Morris[4], T. Lappalainen[5,1]; [1]New York Genome Ctr., New York, NY, [2]Barcelona Supercomputing Ctr., Barcelona, Spain, [3]New York Univ., New York, NY, [4]Univ. of Toronto, Toronto, ON, Canada, [5]SciLifeLab & NY Genome Ctr., New York, NY

Abstract:

Most genetic variants associated with complex traits are found in non-coding regions of the genome. Expression quantitative trait loci (eQTLs) have traditionally linked genotypes, phenotypes, and genes to elucidate the impacts of trait-associated variants. Recently, large-scale CRISPRi perturbation screens with single-cell transcriptomic outputs have emerged as a novel method to identify causal genes for non-coding variants. Nonetheless, systematic differences and similarities between these two methodologies remain poorly understood.

In this study, we aimed to systematically compare the *cis* and *trans* target genes identified by eQTLs and CRISPRi perturbations for genome-wide association study (GWAS) variants for blood cell traits from UK Biobank. First, we characterized target genes for 136 GWAS variants from our previously published data of CRISPRi screens within K562 cells in addition to additional published CRISPRi data sets. For these loci, we identified 168 significantly differentially expressed *cis*-genes within 500kb. Next, we aggregated whole blood and single-cell eQTL data from multiple studies. Among 29 blood cell trait GWAS, altogether 4,061 (26%) independent putative causal variants colocalized (PP.H4 > 0.8) with *cis*-eQTLs linked to 1,244 genes in whole blood or single-cell immune cells. Of the 136 variants with target genes detected with CRISPRi, only 42 had a significantly colocalized *cis*-eQTLs, and 15 of these variants had the same CRISPRi and blood *cis*-eQTL target gene. A GWAS variant colocalizing with an eQTL variant was not associated with an increased odds of differentially expressed *cis*-genes in CRISPRi perturbation data (Wald test p-value = 0.31). Three of the 15 variants (*IKZF1*, *NFE2*, *GFI1B*) had *trans*-regulatory networks in CRISPRi data, with significant but modest overlap and low correlation (Pearson's r = 0.19) between *trans*-eQTL targets genes. One reason for the limited overlap both in *cis* and *trans* may be cell-type specificity of the effects, indicated by cell-type variation in chromatin states and in eQTL effect sizes. In summary, CRISPRi screens and

eQTL analyses show modest overlap and often point to different putatively causal genes of GWAS loci, both in *cis* and *trans*-genes. While recent work has highlighted challenges in eQTL studies, our work highlights distinct gene regulatory mechanisms captured by each method. These findings indicate that both methodologies can offer complementary evidence regarding the target genes of GWAS variants.

Connecting rare variation to extremes of plasma protein levels

Authors: X. Xie[1], T. Li[2], C. Benner[1], J. Irudayanathan[1], H. A. Hejase[1], B. Van De Geijn[1], R. Pendergrass[1], A. Mahajan[1], A. Battle[3], M. McCarthy[1]; [1]Genentech, South San Francisco, CA, [2]Johns Hopkins Sch. of Med., Baltimore, MD, [3]Johns Hopkins Univ., Baltimore, MD

Abstract:

The emergence of large-scale plasma proteomic data has facilitated the integrative analysis of the relationships between genetic variation and plasma protein concentrations, unveiling insights into disease mechanisms and potential biomarkers. While association studies have identified common variants affecting protein expression, and whole exome sequencing has revealed functional rare coding mutations, the effects of rare non-coding variants are still largely unexplored. Here, we analyzed whole genome sequencing (WGS) data and Olink-measured plasma protein levels for 2,923 proteins related to 10,765 genetically unrelated individuals of European ancestry in the UK Biobank. We found that rare variants (MAF < 0.01) are significantly enriched both near proteomic outliers (10kb) and in trans near their regulatory partners. Using the Bayesian statistical model RIVER to integrate variant genomic annotations with observed protein expression, we identified 26,323 functional non-coding rare variants (FRVs) in cis (1MB) of 2,916 proteins across nearly 80 million variants, including 982 genes without previously detected cis-pQTLs. The FRVs exhibit significant ($p < 0.001$) putative effect on DNA accessibility in plasma proteome related cell types such as pericytes and lymphocytes, and are strongly enriched in regulatory elements (AUC=0.82) including super-enhancers of peripheral blood immune cells (AUC > 0.95). Importantly, our FRVs are enriched in pathogenic variants in ClinVar ($p <$ 1e-3), and among 5,146 ClinVar variants with conflicting annotations, we found 240 (4.7%) and 397 (7.7%) with evidence of high and low pathogenicity respectively. Consequently, FRVs demonstrate higher effect sizes across 2,408 complex traits in FinnGen ($p <$ 1e-3). In total, using ICD10 summary diagnosis for 26 diseases and 1461 UKBB phecodes, we identified 461 protein-disease associations where protein expressions predicted by FRVs significantly correlate with disease outcomes ($p <$ 1.7e-5 with Bonferroni correction) in the 400k UKBB European cohort. These associations include pairs with previously established potential links (e.g. COMT with melanoma, KLK3 with prostate cancer) and also pairs with

less understood relevance. Notably, integrating rare variant scores significantly improves risk stratification over polygenic risk scores. Our results therefore represent a comprehensive survey of the genetic landscape of rare non-coding variation underlying human plasma proteome, advances our understanding of the genetic basis of complex traits, highlighting the critical role of rare non-coding variants in disease predisposition.

Session 75: Tick-Tock: The Aging Genome

Location: Room 405

Session Time: Friday, November 8, 2024, 10:15 am - 11:45 am

Cell-type-specific effects of aging on the human prefrontal cortex transcriptome across the lifespan

Authors: K. Girdhar[1], H. Yang[1], T. Clarence[1], M. R. Scott[2], D. Lee[3], J. Bendl[4], P. Fnu[1], C. A. McClung[2], J. Fullard[5], G. Hoffman[3], P. Roussos[3]; [1]Mount Sinai, New York, NY, [2]Univ. of Pittsburgh, Pittsburgh, PA, [3]Icahn Sch. of Med. at Mount Sinai, New York, NY, [4]Icahn Sch. of Med. at Mount Sinai, New York City, NY, [5]Mount Sinai Sch. of medicine, New York, NY

Abstract:

Understanding the molecular mechanisms of human brain development, maintenance, and age-related decline is crucial for interpreting the genetic architecture of developmental and late age related neurodegenerative diseases. While previous studies have successfully investigated age-related transcriptomic changes within childhood (<20 yo) or adulthoods (>=20-100 yo), no research to date has conducted a harmonized study spanning the entire lifespan to systematically create a cell-specific reference aging trajectories of the human brain transcriptome.

To address these limitations, we conducted single-nucleus RNA sequencing on the prefrontal cortex (PFC) region of 284 human postmortem controls spanning an age range of 0-97 years. Our analytical approach was to track age specific changes across four distinct groups: childhood (0-20 yo), young (21-40 yo), middle (41-60 yo) and late adulthood (>60 yo) across 26 major subclasses of diverse cells, including neuronal, glia, endothelial, and mural cells in the PFC, and integrated these changes with lifespan aging trajectories of every gene.

Our analyses revealed several important insights. First, the majority of age-related changes in the PFC transcriptome occurred during childhood and late adulthood, with almost no changes during young and middle adulthood. Second, significant variability in the effect sizes of these age-related genes across neurons and glial cells was observed during childhood, particularly pronounced in excitatory neurons. Conversely, age-related effect sizes during late adulthood showed significant concordance, indicating a presence of a common aging mechanism (CAM) across neurons and glial cells. Interestingly, DNA repair, cell cycle maintenance and immune activation were observed as CAM in neuronal and glia cells respectively. Next, we identified ten distinct age-related characteristic curves of PFC transcriptome from ~330K genes across 26 subclasses by combining the nuclei from four

age groups into a single continuous lifespan dataset. These characteristic curves can be categorized broadly into the following: (1) “housekeeping”, (2) “developmental”, (3) “aging reversal” genes that reversed their aging direction during late adulthood and (4) “dynamic” genes with continuous expression changes across the lifespan. Broadly, the aging transcriptome in the PFC showed a peak expression age of ~14 years for 50% of genes across all subclasses. The resulting aging atlas provides: (a) a resource of the lifespan aging trajectory of every gene in 26 diverse subclasses of PFC and (b) valuable insights into the manifestation of common aging mechanisms across diverse subclasses.

Characterization of de novo retrotransposition events in the aging germline

Authors: S. Li[1], P. Sudmant[2]; [1]Univ. of California, Berkeley, Berkeley, CA, [2]Univ. of California Berkeley, Berkeley, CA

Abstract:

Aging is an emergent phenomenon hallmarked by the deterioration of physiological processes over time. De novo germline mutations are directly transmissible to offspring: increased de novo mutation frequency in the male germline in aging poses significant risk to reproductive success. In particular, de novo structural variants (dnSVs) affect large genomic regions and are known contributors to congenital neurodevelopmental disorders. Notably, autism spectrum disorder is associated with both an elevated dnSV burden and advanced paternal age at conception. Consequently, characterization of dnSV burden in the male germline is critical for understanding increased risk of congenital disease in aging. Advances in highly accurate single-molecule long-read sequencing now enable direct characterization of dnSVs without relying on proxy measures (i.e. read depth). This approach captures native methylation context and improves accuracy by identifying mutations in phase context.We applied PacBio HiFi long-read sequencing to identify dnSVs in bulk sperm samples from twenty donors aged 27 to 62 at an average of 45X coverage. We created highly contiguous phased assemblies (average N50 = 140Mb) for each donor and used personal genome alignment to identify clonal (shared amongst sperm descended from a common progenitor) and unique (private to <4 gametes generated during meiosis) variants. In the first phase of our project, we characterized the frequency and distribution of de novo retrotransposition events, identified by well-defined consensus sequences captured within the span of a single read. The majority of de novo events stem from AluYb8, AluYa5, and L1HS activity, ranging between 2.1 to 10.2 events per 100 individual cells. We observe an increase in de novo AluY subfamily events in individuals of advanced paternal age (40+ years). Remarkably, approximately 50% of events were found in genic and lncRNA regions expressed in the testis per cross-analysis with the GTEx dataset. Finally, we identify

multiple complex clonal dnSVs occurring in regions enriched with ancient L1 sequences, suggesting microhomology-mediated mechanisms. These results suggest that long-read sequencing is a promising method to evaluate the prevalence and spectra of dnSVs in the aging germline.

Linking Rare Non-Coding Variants Associated with Human Longevity to Cellular Senescence via Integrative Functional Genomic Approaches

Authors: J. Yang[1], J-R. Lin[2], Z. Zhang[2], S. Milman[2], N. Barzilai[2], Y. Suh[1]; [1]Columbia Univ. Med. Ctr., New York, NY, [2]Albert Einstein Coll. of Med., Bronx, NY

Abstract:

Centenarians, despite representing only a tiny proportion of the global population, hold the key to access longevity. By decoding the genomes in a large cohort of Ashkenazi Jewish (AJ) centenarians, we have detected rare coding variants that protect against age-related diseases, along with numerous genetic variations in non-coding regions with unknown functions. Non-coding variants, once considered “Junk DNA”, are now known to be significantly enriched in cis-regulatory elements (CREs) that regulate transcriptional activity. However, the functional roles of non-coding variants are difficult to predict due to incomplete knowledge of non-coding regulatory elements, their mechanisms of action, and the cellular states and processes in which they function, let alone the identification of truly causal variants and their target genes. To partially address this challenge, we employed *in vitro* phenotype-based CRISPR screens to discover longevity-associated variant-residing CREs capable of modulating cellular senescence. We prioritized 594 rare (AF<1%) non-coding variants identified in our AJ centenarian cohort by mapping non-coding variants in linkage disequilibrium (LD) to potential CREs annotated by Cis-element Atlas (CATlas). Pooled activation (CRISPRa) or inhibition (CRISPRi) of these longevity-associated CREs using CRE-targeting sgRNAs alleviated cellular senescence in human mesenchymal stromal cells compared to non-targeting sgRNAs. Sequencing-based sgRNA enrichment analysis in endpoint cells identified putative senescence-modulating CREs. Surprisingly, almost all of these CREs were located in intergenic or intronic non-promoter regions. To further elucidate the role of these senescence-modulating CREs in transcriptional regulation, we conducted transcriptome-based single-cell CRISPR interference screens to identify cis-regulated causal genes and their trans-effect networks, leading to the discovery of novel genes driving cellular senescence and potential targets for extending human healthspan and lifespan.

Longitudinal Proteomic Aging Index Construction using Functional Principal Component Analysis

Authors: Z. Rao; Univ. Of Minnesota, Minneapolis, MN

Abstract:

Biological age measures an individual's physiological and functional state through biomarkers, offering a more accurate reflection of the body's aging compared to chronological age. Proteins are crucial in physiological functions. Published and our previous studies have constructed proteomic aging clocks (PACs) that are positively associated with mortality and cancer incidence risk. However, these earlier aging clocks were based solely on cross-sectional data with proteins measured at one time point. Thus, we constructed a longitudinal proteomic aging index (LPAI) using data from the Atherosclerosis Risk in Communities (ARIC) Study, which includes 4955 plasma proteins measured by SomaScan in 11,761 participants at Visit 2 (1990-92), Visit 3, and Visit 5 (5,183 persons, 2011-13), with 4221 participants attending all three visits.Our two-step approach employed functional principal component analysis (FPCA) followed by elastic net penalized Cox regression to construct the LPAI. FPCA captured both the variation in overall protein levels over time and their rate of change. Elastic net regression then facilitated variable selection to identify proteins and their functional PCs pertinent to aging. We constructed the LPAI using this method in a training set of 2954 participants who attended all three visits and validated the LPAI in a test set of the remaining 1267 participants. In the test set, we used Cox proportional hazards regression to estimate the association between LPAI and death from all causes (N=291), cardiovascular disease (N=76), and cancer (N=70) from Visit 5 through 2019, and incident cancer (N=58) ascertained from state cancer registries supplemented with medical records) from Visit 5 through 2015. Covariates adjusted for included chronological age, smoking status, drinking status, body mass index (BMI), hypertension, diabetes, and other relevant risk factors. The LPAI was positively associated with all-cause (HR= 3.40, 95% CI: [2.63,4.38]), cardiovascular disease (HR=3.33, 95% CI: [2.21, 5.01]), and cancer mortality (HR=2.27, 95% CI: [1.38, 3.72]) in the test set. LPAI was also positively associated with cancer incidence (HR= 1.73, 95% CI: [1.07, 2.79]).To our knowledge, the biological aging index we developed is the first to leverage longitudinal omics data. LPAI has the potential to provide more comprehensive and accurate insights into biological aging changes and trends over time. Funding: NHLBI, NCI, NPCR.

Epigenetic age acceleration across chronological age groups and its modifiable lifestyle risk factors in middle-aged and elderly adults

Authors: S. Yang[1], K. Oh[1], M. Yuk[1], J. Youn[1], Q. Dong[2], Z. Wang[2], N. Song[1]; [1]Chungbuk Natl. Univ., Cheongju, Korea, Republic of, [2]St. Jude Children's Res. Hosp., Memphis, TN

Abstract:

Background Epigenetic age acceleration (EAA), which is defined as the difference between chronological age and biological age measured by epigenetic clocks, can be used as an aging biomarker. As the aging population is growing, the burden of aging-related diseases is increasing and strategies promoting healthy aging are needed. In this study, to gain insights into preventive approaches to aging in middle-aged and elderly adults, we estimated epigenetic age (EA) and EAA across chronological age groups and sex and assessed the association of EAA with potentially modifiable factors in a Korean population. **Method** From the Korean genome and epidemiology study (KoGES), a total of 2,747 participants had genome-wide DNA methylation profiles generated by the Korea National Biobank and were included in the current analysis. EA was calculated using multiple epigenetic clocks, but we primarily used Zhang's Elastic Net clock after assessing their performances. The correlation between EA and chronological age was estimated by a simple linear regression. Adjusted least square means (ALSMs) of EAA were compared by chronological age groups (40s, 50s, 60s, ≥70s years) and sex. Associations were estimated between EAA and potentially modifiable factors including the history of 13 common diseases, lifestyle factors (e.g., smoking, drinking, exercise), and body mass index by a linear regression model. **Result** The annual change rate of EA was 0.80 years in middle-aged and elderly adults (≥40 years) and decreased with age (40s=0.86, 50s=0.84, 60s=0.77, 70s=0.68). ALSM of EAA was higher in 50s (ALSM=0.16 years) and 60s (ALSM=0.35 years) and lower in 40s (ALSM=-0.57 years) and ≥70s (ALSM=-0.54 years). In addition, ALSM of EAA was greater in males (ALSM=0.24) than in females (ALSM=-0.30, P<0.01)[ZW1] . EAA was significantly associated with history of myocardial infarction (β=0.15, $P=3.47\times10^{-04}$) and lifestyle factors, specifically, smoking (β=0.38, $P=2.44\times10^{-03}$) and drinking (β=0.32, $P=0.01$). **Conclusion** Our study showed greater EAA among adults in 60s (vs. other ages) and males (vs. females) and that EAA was associated with the history of diseases and lifestyle factors among middle-aged and elderly adults, suggesting that lifestyle and behavioral modification may mitigate accelerated aging.

Session 76: Translating Genetics into Screening Programs

Location: Room 401

Session Time: Friday, November 8, 2024, 10:15 am - 11:45 am

Identification of actionable genetic variants in 4,198 volunteers from the Viking Genes research cohort and implementation of return of results

Authors: J. Wilson[1], L. Klaric[2], M. D. Muckian[1,3], K. Johnston[1], C. Drake[1], M. Halachev[1], E. Cowan[4,5], L. Snadden[5,4], J. Dean[5], S. L. Zheng[6,7], P. K. Thami[6], J. S. Ware[6,7], G. Tzoneva[8], A. R. Shuldiner[8], Z. Miedzybrodzka[5], S. M. Kerr[1]; [1]Univ Edinburgh, Edinburgh, United Kingdom, [2]Univ. of Edinburgh, Edinburgh, United Kingdom, [3]London Sch. of Hygiene and Tropical Med., London, United Kingdom, [4]NHS Grampian, Aberdeen, United Kingdom, [5]Univ. of Aberdeen, Aberdeen, United Kingdom, [6]Imperial Coll. London, London, United Kingdom, [7]Guy's and St. Thomas' NHS Fndn. Trust, London, United Kingdom, [8]Regeneron Genetics Ctr., Tarrytown, NY

Abstract:

Notwithstanding variable penetrance, the benefits of returning clinically actionable results to participants in research cohorts are accruing, yet such a genome-first approach is challenging where this is not standard practice. Here, we describe the return of such results in two founder populations from Scotland. Between 2005 and 2015, we recruited >4,000 adults with grandparents from Orkney and Shetland into the Viking Genes research cohort. Return of genetic data was not offered at baseline, but in 2023 we sent invitations for consent to return of actionable genetic findings to living participants. We generated exome sequence data from 4,198 participants, and used the ACMG v3.2 list of 81 genes, together with ClinVar review and pathogenicity status, plus manual curation, to develop a pipeline to identify potentially actionable genetic variants. We identified 104 individuals (2.5%) carrying 110 actionable genotypes at 39 variants in 23 genes, and validated these by a separate sequence analysis method. Ten actionable variants across seven genes (*BRCA1, BRCA2, ATP7B, TTN, KCNH2, MUTYH, GAA*) have risen 50 to >3,000-fold in frequency through the action of genetic drift and are therefore of potential utility for future screening. Working closely with the NHS clinical genetics service, which provided genetic counselling and validation of the research results in a clinical setting, we notified 66 consenting participants (or their next of kin) of their actionable genotypes. Viking Genes is one of the first UK research cohorts to provide the opportunity to its volunteers to benefit directly from their participation in genetic research and thus provides an ethical and logistical exemplar of implementation of return of results.

Early Check Genome Sequencing of Newborns to Detect Genetic Risk of Type 1 Diabetes

Authors: J. Carter[1], K. Kucera[2], A. Gwaltney[1], H. Cope[2], G. Page[3], L. Gehtland[1], A. Forsythe[1], N. Gaddis[2], M. Schu[1], H. Peay[2]; [1]RTI Intl., Research Triangle Park, NC, [2]RTI Intl., Durham, NC, [3]RTI Intl., Atlanta, GA

Abstract:

Background: Early Check is a research program that provides expanded newborn screening (NBS) using genome sequencing (GS) across the state of North Carolina. Parents use an online portal to consent to screen their newborns for large panels of monogenic conditions and genetic risk for type 1 diabetes (T1D). While array data have typically been used to calculate genetic risk scores (GRS), GS allows for multiplexed screening of hundreds of monogenic conditions and the generation of GRS using one test, which is highly valued by NBS. We are assessing the acceptability and feasibility of recruiting, screening, and reporting results for T1D risk in newborns, including the feasibility of calculating GRS in a population-based sample using GS data. **Methods:** A commercial laboratory performed GS using residual dried blood spots after standard state NBS. We employed self-reported race/ethnicity for the first 2000 samples while we finalize and validate our ancestry calculation pipeline. After monogenic reporting was completed, the genome variant calling format (gVCF) data were used to calculate T1D genetic risk using the previously published and validated T1DGRS2 (n=67 biomarkers). Samples were excluded if missing ≥2 human leukocyte antigen (HLA) markers due to impact on T1D GRS accuracy. Results are returned in an online portal and reported as one of three categories of lifetime risk (LR) to develop T1D: Low Concern (LC) LR <2%; Moderate concern (MC) LR 2-5%; and Higher concern (HC) >5% LR. Follow-up differs by risk category.**Results:** After 8 months of screening, 1801 newborns have been sequenced and ~80% of parents have selected T1D screening for the newborn. White parents were significantly more likely to select T1D screening than Black (*padj* < 0.0001) or Hispanic parents (*padj* < 0.05). To date, 1314 GRSs were calculated; 87.5% were LC, 7.2% were MC, and 3.8% were HC. Twenty (1.5%) were unable to be calculated due to marker missingness. The GRS mean was 10.45, SD = 2.3, and ranged from 2.32-17.44. Mean GRS scores were significantly different based on self-reported race/ethnicity (p < .0001).**Conclusions:** Our study demonstrates the feasibility of using GS data to multiplex screening for both monogenic and complex disorders. While a low percentage of MC and HC newborns will develop T1D, NBS provides an opportunity to monitor for early disease with validated diagnostic tests (autoantibodies). We anticipate transitioning to genetic ancestry calculation in summer 2024. Our

preliminary data, as supported by other studies, indicates the need for pan ethnic GRS and ancestry-specific thresholds.

Combining gene genealogies and pedigrees to inform genetic screening programs

Authors: A. Mejia Garcia[1], G. Sillon[1], D. D'Agostino[2], L. Baret[1], A-L. Chong[2], G. Chong[3], N. Hamel[2], k. Sin Lo[4], A. Diaz-Papkovich[1], W. Foulkes[2], G. Lettre[5], A. Shapiro[2], S. Gravel[1]; [1]Human genetics Dept., McGill Univ., Montreal, QC, Canada, [2]McGill Univ. Hlth.Ctr., Montreal, QC, Canada, [3]Jewish Gen. Hosp, Montreal, QC, Canada, [4]Montréal Heart Inst., Montreal, QC, Canada, [5]Université de Montréal, Montreal, QC, Canada

Abstract:

Introduction: Gene genealogies represent the ancestry of a sample and are often encoded as ancestral recombination graphs (ARG). It has recently become possible to infer these gene genealogies from sequencing or genotyping data and use them for evolutionary and statistical genetics. Unfortunately, inferred gene genealogies can be noisy and subject to biases, making their applications more challenging. The goal of this project is to study the application of ARG methods to systematically impute and trace the transmission of all disease variants in founder populations where long shared haplotypes allow for accurate timing of relatedness. We apply the methods to the population of Quebec, where multiple founder events led to uneven distribution of pathogenic variants across regions, and where extensive population pedigrees are available. **Methods:** Using the CARTaGENE cohort data, we reconstruct the ARG using genotype data of 30k genotyped individuals. We identified carriers of known and novel pathogenic mutations using whole genome sequencing data of a subset of individuals (WGS, n=2173). Using the ARG, we estimated mutation age and whether each variant was introduced once or multiple times in the population. We developed an ARG-based imputation strategy to infer carrier rates within the genotype data and used ISGen to impute the variants in unsampled individuals via the genealogical record, allowing for estimation of regional frequency estimates for many pathogenic variants. **Results:** We applied our method to 8 known founder mutations in Saguenay, and we estimated their mutation age. We found a positive correlation between GnomAD carrier frequency in non-Finnish Europeans and mutation age (R=0.61). Moreover, we estimated carrier frequency in Saguenay, and we did not find significant differences in carrier frequency against the literature for *CYP27B1, CTNS* and *GNPTAB*. In addition, we validated our imputation approach by genotyping the mutation *PMS2 C.2117DEL* in 4000 CARTaGENE individuals and we confirmed 5/6 imputed carriers (99,97% accuracy, 100%

sensitivity, and 90,90% Kappa statistic) . On the other hand, we imputed the *TTC25 C.245DEL* variant, and found a 99,97% accuracy, 87,50% sensitivity and 92,55% Kappa statistic against TOPMED imputation. Moreover, our regional frequencies estimation showed a higher carrier frequency in Bas-Saint-Laurent (8.6/1000) and Cote-Du-Sud (7/1000).**Conclusions:** We demonstrated that ARG-based imputation is useful for the study of rare mutations and allows posterior regional frequency estimation.

Rate and profile of secondary findings in 381 participants in the DDD-Africa from the DR Congo ⭐

Authors: A. Lumaka Zola[1], G. Mubungu[1], P. Makay[2], N. Louw[3,4], P. Mpangase[5], H. Firth[6,7], M. Hurles[8], N. Carstens[9], A. Krause[3,4], Z. Lombard[10,11], P. Lukusa Tshilobo[12,13,14], K. Devriendt[14]; [1]Univ. of Kinshasa, Kinshasa, Congo, Democratic Republic of the, [2]Université de Kinshasa, Kinshasa, Congo, Democratic Republic of the, [3]Div. of Human Genetics, Natl. Hlth.Lab. Service, Johannesburg, South Africa, [4]Sch. of Pathology, Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa, [5]Sydney Brenner Inst. for Molecular BioSci.. Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa, [6]Human Genetics Programme, Wellcome Sanger Inst., Wellcome Genome Campus, Hinxton, United Kingdom, [7]East Anglian Med. Genetics Service, Cambridge Univ. Hosp. NHS Fndn. TrustC, Cambrige, United Kingdom, [8]Wellcome Sanger Inst., Cambridge, United Kingdom, [9]Genomics Ctr., South African Med. Res. Council, Tygerberg, Cape Town, South Africa, [10]Div. of Human Genetics, Natl. Hlth.Lab. Service & Sch. of Pathology, Johannesburg, South Africa, [11]Dept. of Internal Med., Sch. of Clinical Med., Faculty of Hlth.Sci., Univ. of the Witwatersrand, Johannesburg, South Africa, [12]Ctr. for Human Genetics, Faculty of Med., Univ. of Kinshasa, Kinshasa, Congo, Democratic Republic of the, [13]Dept. of Pediatrics, Faculty of Med., Univ. of Kinshasa, Kinshasa, Congo, Democratic Republic of the, [14]Ctr. for Human Genetics, Univ. Hosp., Univ. of Leuven, Leuven, Belgium

Abstract:

Background: Access to Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) is increasing in Africa. However, little is known about Secondary Findings (SFs) in Sub-Saharan Africa in genes from the American College of Medical Genetics and Genomics (ACMG) list as well as in non-ACMG genes relevant to Africa. We aimed to determine the rate and profile of SFs in a dataset of WES from individuals from Central Africa. **Methods**: Pathogenic and likely pathogenic (P/LP) variants in 97 genes of the ACMG v3.2 list were searched in clinical exome sequencing data of 381 individuals from 147 families from the DR Congo participating in the DDD-Africa study. The pathogenic Glu7Val

variant in the *HBB* gene and the A haplotype in the *G6PD* gene were also filtered. P/LP Classification was extracted from ClinVar. Heterozygosity for the A haplotype was not considered actionable and, thus not retained as a secondary finding. **Results**: We identified 12 P/LP variants in eleven ACMG SF v3.2 genes (*RPE65, TTN, SCN5A, PMS2, MYBPC3, RB1, ATP7B, BRCA1, GAA, TTR,* and *LDLR)* across 34 individuals from 28 families. The overall ACMG SF rate is 8.92%. After excluding 3 individuals with heterozygous variants in *ATP7B* and *GAA* as recommended by the ACMG, the adjusted SF rate was 8.13% for ACMG v3.2 genes. Two SFs in *TTR* were the most frequent (15 individuals from 10 families), followed by an ultra-rare known splice acceptor variant in the *MYBPC3* 5 individuals from 4 families) and an ultra-rare known pathogenic intronic variant in *RB1* (2 individuals from 2 families).The pathogenic *HBB* Glu7Val variant was present in 3 homozygous and 89 heterozygous (reportable in 23.4% of participants). The A haplotype of the *G6PD* was present in reportable status in 43 hemizygous and 9 homozygous (reportable in 13.7% of participants). **Conclusion**: We observed a much higher SF rate in ACMG genes than previously reported (2.3% in West Africa or about 1.2% in African American). The higher prevalence of *TTR* variants in our region may explain this discrepancy. Some ultrarare variants in Western datasets might be of interest in Africa. Systematic reporting of P/LP variants in *HBB*, *G6PD* could significantly benefit Africa's public health.

An association study without genotype sharing for uncovering germline susceptibilities in pediatric cancers

Authors: M. Artomov[1,2,3], A. Loboda[1,3], C. E. Cottrell[1,2], Y. Akkari[1,2], K. Tsuchiya[1,2], S. Reshmi[1,2], G. Lammi[1,2], M. Daly[4,3], R. Wilson[1,2], E. Mardis[1,2]; [1]Nationwide Children's Hosp., Columbus, OH, [2]The Ohio State Univ. Coll. of Med., Columbus, OH, [3]Broad Inst., Cambridge, MA, [4]Massachusetts Gen. Hosp., Boston, MA

Abstract:

Genetic data are subject to strict data sharing regulations. This is especially applicable to cohorts assembled based on shared molecular diagnostics procedures. As a result, analysis of inherited susceptibilities in such cohorts is often performed in a format of a “case study” rather than a “case-control study” limiting the ability for novel disease gene discovery. Patients characterized by molecular diagnostics as a component of clinical care often present severe symptoms and are expected to carry significant inherited disease burden. Therefore, *we aimed to illustrate how technological advances in secure data analysis enable introduction of genetic cohorts with strictly regulated access into case-control studies to power disease gene discovery*.

The need for advanced data analysis and novel susceptibility gene identification is especially high for pediatric patients. The Molecular Characterization Initiative (MCI) is a part of the National Cancer Institute's Childhood Cancer Data Initiative project and, in collaboration with sites affiliated with the Children's Oncology Group, aims to collect, analyze and report clinical and molecular data to support clinicians in choosing the best treatment for each child through precision diagnosis. *We have used enhanced coverage exome sequencing data from 2,529 pediatric cancers from MCI as a pan-cancer case group.*
Our recent solution - SVD-based control repository (SCoRe - http://dnascore.net) provides unlimited access to a database of 40,000 exome controls and a complimentary R-package to be used by end users to generate shareable summary data from case genotypes that will be uploaded to the web-platform. *We used SCoRe to remotely select 5,312 ancestry-matched controls and obtained summary allele counts.* We showed that among the known risk genes for dominant cancer syndromes there was a 3.4-fold increase in rare protein-truncating variant burden ($p=5.3x10^{-6}$) in pediatric patients, while rare synonymous variants showed no difference (p=0.86) confirming the absence of bias in the analysis. We also have performed exome-wide gene-based association study and replication for the top gene candidates across multiple ancestries.
Finally, we developed the computational strategy for scaling remotely accessed control databases. We validated our approach using ~800,000 exomes that are a part of the gnomAD dataset. Through the usage of advanced clustering algorithms, we show that it is feasible to achieve reasonable computational and storage costs. As a result, this approach enables practical usage of gnomAD for ancestry-matched control subject selection.

Biobank-scale genotype-to-phenotype analyses reveal the challenges in using exome sequencing for population screening

Authors: D. Blair, N. Risch; Univ. of California San Francisco, San Francisco, CA

Abstract:

There is growing interest in applying exome and genome sequencing to asymptomatic patients, with the goal of identifying at-risk individuals prior to disease onset. The utility of sequencing for screening depends on the penetrance of the identified variants and the validity of their characterization. Here, we evaluate the prognostic accuracy of exome sequencing for the simplest class of disease-causing variants: predicted loss-of-function mutations (pLOFs) in haploinsufficient disease genes. Using ClinGen, we identified 91 haploinsufficient diseases linked to 133 genes. We then uniformly processed the genomic

and phenotypic data from two biobanks: the UK Biobank (UKBB) and the All of Us (AoU) Research Program. Using statistical modeling to account for differences in data coverage and disease onset, we identified thousands of pLOF carriers for which we could confidently determine the presence or absence of the target phenotypes using both disease diagnoses and symptom-based clustering. pLOF penetrance varied across the diseases (range: 0.8-63%), but the target genotype-to-phenotype associations were consistent and strong. We then assessed the performance of exome sequencing for predicting disease expression. If all pLOFs were reported, the positive predictive value (PPV) of the test was low (4% in UKBB, 8% in AoU). Restricting the reported variants to those with Pathogenic/Likely Pathogenic (P/LP) annotations in ClinVar improved the PPV (6% in UKBB, 12% in AoU), but many pLOFs in both datasets (34% in UKBB, 45% in AoU) were unannotated, limiting sensitivity. Unannotated variants were less predictive of disease, but their penetrance was significantly higher than pLOFs classified as Benign/Likely Benign in ClinVar (4% vs 1% in UKBB, 6% vs 3% in AoU, P-value < 0.05). To improve sensitivity, we built machine learning (ML) models that used variant-specific features to predict pLOF expression independent of the ClinVar annotations, training them in the UKBB and evaluating them in AoU. Compared to P/LP annotations, the top performing ML model had significantly increased PPV (25% vs 12%; P-value < 10^{-4}) but lower sensitivity (45% vs 54%; P-value < 10^{-3}). Combining the annotations with the ML model reduced the PPV (12%) but led to a substantial increase in sensitivity (65%; P-value < 10^{-4}), capturing 25% of the symptomatic carriers of unannotated pLOFs. These results highlight both the reduced penetrance and imprecise characterization of pLOFs, including those with P/LP annotations. We suggest caution when providing genetic counseling to asymptomatic pLOF carriers, as outcomes remain uncertain even for variants with strong evidence for pathogenicity.

Session 77: Exploring Omics: From Genomes to Microbiomes

Location: Mile High Ballroom 2&3

Session Time: Friday, November 8, 2024, 1:15 pm - 2:15 pm

Institution-wide access to a scalable, clinical grade genomic sequencing platform advanced rare disease research and improved clinical outcomes in a pediatric setting

Authors: C. French[1], W. Shao[1], A. Sharma[1], A. Beggs[1,2], J. Chou[1,2], A. Poduri[1,2], S. Rockowitz[1,2], P. Sliz[1,2], Children's Rare Disease Cohorts Initiative; [1]Boston Children's Hosp., Boston, MA, [2]Harvard Med. Sch., Boston, MA

Abstract:

Genomic testing for rare disease patients can improve clinical care and facilitate research. Boston Children’s Hospital has established a genomic sequencing and analysis research initiative to expand access and benefit patients and investigators across a pediatric tertiary healthcare setting. Through the Children's Rare Diseases Cohorts (CRDC) initiative, the hospital offers CLIA-grade sequencing to selected patients enrolled in specialized rare disease research studies. The resulting data, which are consented for broad research use, are harmonized and analyzed with a wide range of CRDC-supported variant interpretation tools. Since the initiative launched, 66 investigators representing 26 divisions/programs and 45 phenotype-based cohorts joined the CRDC. These studies enrolled 4,653 families under a broad-use research consent and 35% of those analyzed so far had a genetic finding either clinically confirmed and returned or pursued with further investigation. The hospital’s genomics data analysis platform has also expanded to support additional institutional data collections, both research and clinical, and now encompasses over 13,800 patients and their families. This combined genomic database and associated phenotypic information have fostered numerous new research projects and collaborations involving trainees and established investigators. Establishment of an accessible and harmonized genomics platform resulted in increased numbers of genetic diagnoses and accelerated innovative research via integration of genomics research with clinical care.

Project Baby Lion - Introducing ultra-rapid genome sequencing in German neonatal and pediatric ICUs

Authors: B. Auber[1], G. Schmidt[1], D. Chen[1], H. Wallaschek[1], J. Ronez[1], A. von Gise[2,3], B. Bohnhorst[4], M. Sasse[2], H. Köditz[2], W. Hofmann[1], M. Losch[1], B. Haermeyer[1], B. Schnur[1], F. Kaisen[1], I. Klefenz[1], D. Steinemann[1], N. Di Donato[1], S. Von Hardenberg[1]; [1]Dept. of Human Genetics, Hannover Med. Sch., Hannover, Germany, [2]Dept. of Pediatric Cardiology and Intensive Care Med., Hannover Med. Sch., Hannover, Germany, [3]Dept. of Neonatology, Children's and Youth's Hosp. Auf der Bult, Hannover, Germany, [4]Dept. of Pediatric Pulmonology and Neonatology, Hannover Med. Sch., Hannover, Germany

Abstract:

Background: Rare genetic diseases are a significant cause of critical illness in children. Ultra-rapid whole genome sequencing (urWGS, <5 days) has diagnostic rates exceeding 30% and has demonstrated clinical benefits for patients requiring intensive care. However, the integration of this technology into routine diagnostics, especially in non-university hospitals, remains a challenge. Here, we present the first German multicenter urWGS study (DRKS00025163), which was conducted over a period of 23 months.
Methods: A total of 130 children (0-14 years, mean age 1.3 years) were enrolled in the study. These patients were admitted to neonatal or pediatric intensive care units (NICUs/PICUs) across 15 hospitals in northern Germany. Eligible patients were enrolled following a multidisciplinary case conference, if no evidence for a non-genetic cause was identified. urWGS was conducted preferentially in a trio-setting. Clinical utility was measured using the C-GUIDE™ survey 14 days after the final report, and parental perception was monitored separately. Optical Genome Mapping (OGM) was conducted for orthogonal structural variant detection in non-diagnostic cases.
Results: 60 patients (46%) received molecular diagnosis with urWGS and a median turn-around time of 2.3 days. The highest diagnostic yield was observed in patients with endocrine/metabolic disorders (70%), epilepsy (69%), and syndromic conditions (55%). Notably, 9% of the causative variants would not have been detectable with exome sequencing such as one *DMPK* repeat expansion or intronic/non-coding variants. OGM did not identify any additional causative variants in non-diagnostic urWGS cases. C-GUIDE™ survey revealed a significantly higher clinical utility for diagnostic compared to non-diagnostic findings. A change in management was observed in 42% of the diagnostic cases.
Conclusion: The successful multicenter patient recruitment in this study demonstrates the feasablity of rapid turnaround time and high diagnostic yield using urWGS. The incorporation of online multidisciplinary meetings proved particularly beneficial for

hospitals without a genetics department, enabling them to provide access to genomic testing for eligible patients. In turn, this approach facilitated the treatment of critically ill children with precision medicine.

Assessing HiFi genome sequencing as first-tier test in rare disease genetics

Authors: A. Hoischen[1], L. Snijders Blok[1], L. Sagath[1], W. Hops[1], B. van der Sanden[1], A. den Ouden[1], R. Derks[1], S. van den Heuvel[1], M. Kwint[1], J. van Reeuwijk[1], R. Timmermans[1], J. Corominas Galbany[1], T. Hofste[1], a. van den Wijngaard[2], H. IJntema[1], J. Weiss[1], C. Gilissen[1], L. Vissers[1]; [1]Dept. of Human Genetics, Radboud Univ. Med. Ctr., Nijmegen, Netherlands, [2]Maastricht Univ. Med. Ctr.+ (MUMC+), Maastricht, Netherlands

Abstract:

Long-read sequencing (LRS) technologies, such as PacBio HiFi genome sequencing on Revio, allow to assess the full human genome for the first time and have the potential to revolutionize human genetics. This technology could offer a comprehensive first-tier test for clinical genetics and rare disease research. To determine the clinical utility of LRS, we performed Revio HiFi genome sequencing on 500 human genomes with ~30-fold coverage, including 100 mutation-positive controls that are impossible or challenging to identify by standard short-read genome sequencing, 100 severe sporadic rare disease cases as patient-parent trios, and 100 severe rare disease cases as singletons who remained undiagnosed after standard-of-care testing. The results of this study demonstrate that LRS can identify 95% of known mutations, including karyotype abnormalities, structural variants, SNVs, InDels (in homopolymer stretches), mutations in homologous sequences, short-tandem repeat expansions, and methylation alterations such as (partial) uniparental disomies. LRS also provides more context to already known variants, such as base-pair resolution breakpoints of SVs, more accurate estimates of repeat expansion size and sequence context, phasing of variants, or methylation-alterations in cis. For variants not readily called from the LRS data, a subset was visible but not (yet) called. So far, the only systematically failed or challenging variant types were repeat expansion disorders with very long AG-rich repeats (*RFC1, FXN*), likely due to the most challenging sequence context for which HiFi quality was not achieved. LRS of undiagnosed cases showed that *de novo* mutations can be readily identified on all variant levels (SNV, InDel, SV), and first cases have been diagnosed with previously hidden or missed (*de novo*) mutations. Preliminary data suggest >10% extra yield in undiagnosed trios and singletons, however systematic interpretation is currently ongoing. Overall, the results of the first year using Revio LRS highlight the comprehensive nature of latest genome sequencing methods. Further research is needed to fully assess the clinical utility of LRS, but the results of this

study support our enthusiasm for LRS as a first-tier test for clinical genetics and rare disease research in the near future.

Expanding the human gut microbiome atlas of Africa

Authors: L. Olubayo; Univ. of Witwatersrand, Johannesburg, South Africa

Abstract:

Population studies are crucial in understanding the complex interplay between the gut microbiome and geographical, lifestyle, genetic, and environmental factors. However, populations from low- and middle-income countries, which represent ~84% of the world population, have been excluded from large-scale gut microbiome research. Here, we present the AWI-Gen 2 Microbiome Project, a cross-sectional gut microbiome study sampling 1,803 women from Burkina Faso, Ghana, Kenya, and South Africa. By intensively engaging with communities that range from rural and horticultural to urban informal settlements and post-industrial, we capture population diversity that represents a far greater breadth of the world's population. Using shotgun metagenomic sequencing, we find that study site explains substantially more microbial variation than disease status. We identify taxa with strong geographic and lifestyle associations, including loss of Treponema and Cryptobacteroides species and gain of Bifidobacterium species in urban populations. We uncover a wealth of prokaryotic and viral novelty, including 1,005 new bacterial metagenome-assembled genomes, and identify phylogeography signatures in Treponema succinifaciens. Finally, we find a microbiome signature of HIV infection that is defined by several taxa not previously associated with HIV, including Dysosmobacter welbionis and Enterocloster sp. This study represents the largest population-representative survey of gut metagenomes of African individuals to date, and paired with extensive clinical biomarkers, demographic data, and lifestyle information, provides extensive opportunity for microbiome-related discovery and research.

Session 78: Genetics of Human Brain: Regulation, Disease Risk, and Assortative Mating

Location: Room 401

Session Time: Friday, November 8, 2024, 1:15 pm - 2:15 pm

Establishing the Molecular Foundation of Brain Anatomy in Living Individuals

Authors: A. Lund[1], L. Liharska[2], E. Vornholt[2], R. Thompson[3], Y. Luo[1], E. Cheng[2], Y. Park[1], B. Fennessy[2], LBP Team, E. Schadt[4], B. H. Kopell[1], A. Charney[2], N. Beckmann[2]; [1]Icahn Sch. of Med., New York, NY, [2]Icahn Sch. of Med. at Mount Sinai, New York, NY, [3]Icahn Sch. of Med. at Mount Sinai, Mount Kisco, NY, [4]Rosetta Inpharmatics, Seattle, WA

Abstract:

The structure of the human brain is critical to brain function, but beyond genetic studies, there are no large-scale efforts to characterize the relationship between anatomical components and brain molecular traits in living humans. Here, we present the first characterization of these relationships, by leveraging neuroimaging-linked molecular data generated from 320 biopsies from the dorsolateral prefrontal cortex (dlPFC) from 202 older living individuals with and without Parkinson's disease (mean age = 61.8 years old). Gene expression (single-nuclei [snRNAseq], N=31 and bulk RNA-seq, N=289) as well as protein expression (liquid chromatography-mass spectrometry, N=155) data generated as part of the Living Brain Project, was processed, and integrated with anatomical features (N=197 features, cortical thickness, volume, and area) measured in the same individuals. Associations of molecular traits to anatomical features were tested for each imaging feature using linear mixed models. Associations replicated between -omics and annotated to specific cell type effects using snRNA-seq data. Molecular signatures of brain anatomical features in living individuals were identified, with multiple features associated with bulk gene expression (19 features with differentially expressed genes at FDR ≤ 0.05). These associations replicated in protein expression (median absolute Spearman rho=0.13). SnRNAseq replicated gene and protein expression signatures, and annotated them to specific cell types, providing further characterization of these signatures and their pathways (e.g.,for medial orbitofrontal thickness, snRNAseq differential expression (DE) signature in excitatory neurons is correlated to gene [rho=0.22] and protein [rho=0.24] DE signatures). By associating molecular traits with imaging anatomical features generated from the same individuals, we provide a novel in-depth characterization of the molecular foundation of brain anatomy in living individuals.

Mapping the regulatory effects of rare non-coding variants across cellular and developmental contexts in the brain

Authors: A. Marderstein, S. Kundu, A. Kundaje, S. Montgomery; Stanford Univ., Stanford, CA

Abstract:

Background: Rare variants are abundant in the human population, and selective constraint prevents many disease-causing and fitness-reducing variants from reaching common frequencies. Thus, identifying rare variants that are important for disease remains a significant challenge, while the relationship of constraint to cell-type-specificity and developmental timing remains unclear. Regulatory deep learning models trained on DNA sequences present a unique opportunity to study variant effects across the frequency spectrum in a diversity of cellular contexts.
Methods: We used a model called *chromBPnet* to predict the cell-type-specific effects on chromatin accessibility from scATAC-seq. Since chromBPnet can be applied to any base pair to predict a variant's impact on accessibility and transcription factor binding for any cell type or context for which we have ATAC data, we used chromBPnet to predict regulatory effects at over 10 million rare variants (<0.1% MAF in 1000G EUR) and 10 million common variants (>5% MAF) from 46 brain cell types in adult and fetal brain. Additionally, we directly inferred selective constraint at rare variants using PhyloP.
Results: As expected, rare variants had larger predicted effects on accessibility than common variants across cellular and developmental contexts. However, we also observed that rare variants had broader effects in fetal brain by impacting a greater number of cell types (on average) compared to common variants. Selective constraint had increased correlation with regulatory effects in the fetal brain compared to adult brain, in terms of effect magnitude and number of cell types impacted. In addition, using data from non-brain tissues, we found that fetal constraint was organ-specific, as cell types from fetal and adult heart exhibited similar constraint to those from adult brain. Overall, we identified 5,872 constrained rare variants with fetal brain effects—such as rs553352185, which is located near a regulatory BMI GWAS locus between the axon guidance-related *ROBO1* and *ROBO2* genes, disrupts a E2F family motif in fetal excitatory neurons, and has a BMI GWAS effect size in the top decile of tested variants.
Conclusion: Our results indicate that variants with large regulatory effects (both in terms of magnitude and number of cell states impacted), particularly in early (fetal) stages of brain development, are more likely to be selected against and rare in the human population. This has implications for human genetic studies of neurodevelopmental

disorders, which undergo stronger selective pressures and for which our results indicate that most of the major genetic influence would be found within rare variation.

The largest to-date exome study of autism spectrum disorder triples the number of autism-associated genes

Authors: F. Satterstrom[1,2], J. M. Fu[1,2], K. McWalter[3], Z. Zhang[3], T. Thomas[1,2], H. Brand[1,2], R. Kueffner[3], E. Shen[1,2], J. Lim[1,2], C. Cusick[1], C. Stevens[1], C. Liao[1,2], L. Wang[1,4], D. J. Cutler[5], K. E. Samocha[1,2], E. B. Robinson[1,2], J. D. Buxbaum[6], B. Devlin[7], K. Roeder[8], P. Kruszka[3], S. J. Sanders[9], M. E. Talkowski[1,2], M. J. Daly[1,2], Autism Sequencing Consortium; [1]Broad Inst., Cambridge, MA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]GeneDx, Gaithersburg, MD, [4]Harvard Med. Sch., Boston, MA, [5]Emory Univ. Sch. of Med., Atlanta, GA, [6]Icahn Sch. of Med. at Mount Sinai, New York, NY, [7]Univ. of Pittsburgh, Pittsburgh, PA, [8]Carnegie Mellon Univ., Pittsburgh, PA, [9]Oxford Univ., Oxford, United Kingdom

Abstract:

Autism spectrum disorder (ASD) genetics is advancing rapidly. In 2022, the Autism Sequencing Consortium (ASC) published an analysis of exome sequences from 20,627 individuals with ASD (and 63,000 individuals total) that identified 72 genes associated with ASD at a false discovery rate (FDR) below 0.001. Here, we report a new analysis of 62,013 individuals with ASD (and 171,000 individuals total). Made possible by combining samples from the ASC, the Simons Simplex Collection, the Simons Powering Autism Research (SPARK) project, and a leading diagnostic laboratory (GeneDx), it is the largest exome study of ASD to date.

For 38,088 of the ASD cases (as well as 9,567 unaffected siblings), we have parental sequences and can identify *de novo* variants. *De novo* protein-truncating variants (PTVs) in constrained genes (the lowest decile of LOEUF) are strongly enriched in cases compared to siblings (0.061/case vs 0.017/sibling, $p < 2.2E-16$), while *de novo* synonymous rates are comparable (0.29/person for both groups, $p = 0.59$). Notably, rates of constrained *de novo* PTVs in GeneDx ASD cases without co-occurring developmental delay (DD) or intellectual disability (ID) align well with research samples ascertained for ASD (0.052/case vs 0.040/case), while GeneDx ASD cases with DD/ID exhibit rates similar to previous studies of DD (0.11/case vs 0.12/case; DD from Kaplanis et al., *Nature* 2020). In addition, among 9,600 male SPARK cases for whom we can calculate an ASD polygenic risk score (PRS), the 10% of cases with the lowest PRSs are significantly more likely to have a constrained *de novo* PTV than the other 90% of cases (0.060/case vs 0.040/case, $p = 0.013$), demonstrating the combined effects of common and rare variation on ASD liability.

Our gene discovery analysis of this large dataset uses a Bayesian model to integrate across genetic variants of different inheritance classes -- *de novo* and inherited variants from the cases with sequenced parents, and rare variants from those without -- as well as variants of different types, including PTVs, missense variants, and copy number variants. Together, these data identify 230 genes associated with ASD at an FDR below 0.001. While the evidence for association for the top genes is primarily driven by *de novo* PTVs, we observe a steadily increasing contribution of other variant types in newly associated genes. We also observe power improvement in gene discovery (~10 genes) by combining AlphaMissense scores and updated MPC scores for missense variant categorization. These results shed light on the etiology of ASD and will help disentangle the shared and distinct genetic architectures of ASD and other neuropsychiatric conditions.

Assortative Mating Across Nine Psychiatric Disorders: Consistency and Persistence Across Cultures and Generations

Authors: C. Fan[1], S. R. Dehkordi[2], R. Border[3], L. Shao[4], R. Loughnan[5], W. Thompson[1], L-Y. Hsu[6], M-C. Lin[7], C. Chi-Fung[8], M-H. Su[9], T. Werge[10], C-S. Wu[8], N. Zaitlen[11], A. Buil Demur[2], S-H. Wang[12]; [1]Laureate Inst. for Brain Res., Tulsa, OK, [2]Inst. of Biological Psychiatry, Roskilde, Denmark, [3]Univ. of California Los Angeles, Los Angeles, CA, [4]Biostatistic Program, La Jolla, CA, [5]Univ. of California, San Diego, La Jolla, CA, [6]Inst. of Epidemiology and Preventive Med., Coll. of Publ. Hlth., Natl. Taiwan Univ., Taipei, Taiwan, [7]Dept. of Publ. Hlth., Coll. of Publ. Hlth., China Med. Univ., Taichung, Taiwan, [8]Natl. Ctr. for Geriatrics and Welfare Res., Natl. Hlth.Res. Inst.s, Zhunan, Taiwan, [9]China Med. Univ., Taipei, Taiwan, [10]Inst. of Biological Psychiatry, Roskilda, Denmark, [11]UCLA, Los Angeles, CA, [12]Natl. Hlth.Res. Inst.s, Zhunan, Taiwan

Abstract:

Importance: Assortative mating (AM) influences the prevalence and comorbidity of psychiatric disorders, biasing genetic architecture estimates. Understanding AM's consistency and persistence across cultures and generations is crucial for accurate genetic studies.**Objective:** To examine AM patterns across nine psychiatric disorders in different cultures (Taiwan, Denmark, and Sweden) and across generations.**Method:** We utilized national registry datasets from Taiwan, Denmark, and Sweden, covering individuals born between the 1930s and 2000s. The analyses included up to 1.4 million mated cases and 6 million matched controls, focusing on all possible pairs across nine psychiatric disorders, i.e. Schizophrenia (SCZ), Attention Deficit - Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), Major Depressive Disorder (MDD), Bipolar Disorder

(BPD), Anxiety Disorders (Anxiety), Obsessive Compulsive Disorder (OCD), Substance Use Disorder (SUD), and Anorexia Nervosa (AN). Spousal correlations were calculated using tetrachoric correlations, and meta-analyses were conducted with random effect models.**Results:** We found positive spousal correlations were observed for all nine psychiatric disorders across Taiwan, Denmark, and Sweden. Despite the population differences, the AM patterns across all 81 possible disease-spousal pairs were highly correlated with the SNP heritability/co-heritability estimated from European based genome-wide association studies (GWAS). Meta-analyses indicated consistent AM patterns across cultures with minimal variations. Generational analysis showed limited change of AM over a quarter of centuries, except increasing AM for Substance Use Disorder (SUD) and decreasing AM for Obsessive-Compulsive Disorder (OCD).**Conclusions:** AM patterns for psychiatric disorders are consistent and persistent across cultures and generations. These patterns significantly impact genetic architecture estimation, suggesting that AM must be considered in genetic studies of psychiatric disorders regardless the study populations.

Session 79: Lessons from Height

Location: Room 405

Session Time: Friday, November 8, 2024, 1:15 pm - 2:15 pm

Leveraging whole-genome sequencing data from 750,000 diverse-ancestry individuals across biobanks to understand the genetic architecture of common anthropometric traits

Authors: H. Wright, G. Hawkes, R. N. Beaumont, K. Chundru, K. A. Patel, L. Jackson, T. M. Frayling, A. Murray, C. F. Wright, A. R. Wood, M. N. Weedon; Univ. of Exeter, Exeter, United Kingdom

Abstract:

Background:
Most association studies for common human phenotypes have focused on common variants, and rare variants that reside in the coding regions of the genome. However, there remains substantial heritability to be explained by rare genetic variation, particularly amongst non-European ancestry individuals. Here, we used whole-genome sequence (WGS) data in 750,000 diverse-ancestry individuals to test the contribution of rare coding and non-coding variants towards the genetic architecture of common anthropometric traits.
Methods:
We performed WGS association analysis of 500,000 individuals in UK Biobank for height, body mass index (BMI) and waist-hip ratio adjusted for BMI (WHRadjBMI). We tested both single variants (minor allele count >= 5; $P < 3e-10$) and rare variant aggregates (minor allele frequency < 0.1%; $P < 9e-9$). We grouped variants for aggregate testing based on overlap with protein coding genes, untranslated regions (UTR), and proximal- or distal-regulatory annotations, as well as measures of conservation and constraint. We replicated our findings in 226,000 individuals of genetically inferred European (N=129,000), African (N=55,000) and Admixed American (N=42,000) ancestries in All of Us.
Results:
We discovered and replicated 123 rare independently associated single variants and 106 aggregates with evidence of association with one of the three traits. For example, we identified two independent rare variants upstream of *IGF2BP2* associated with substantial effect sizes (beta > 0.1 SD) on WHRadjBMI. We also identified a novel coding association between BMI and high-confidence predicted loss-of-function variants in *UBR3* (beta = 2.6 kgm^{-2} 95%CI [1.8, 3.5], $P = 1e-9$, replication $P = 3e-3$), missed by exome sequencing.

Additionally, we identified a novel non-coding association with height in the 5'UTR of *FGF18* (beta = 0.51 cm 95%CI [0.34, 0.68], *P* = 3e-12, replication *P* = 1e-4), a gene previously implicated in osteoblast proliferation in developing bone. *FGF18* is a highly loss-of-function constrained gene with no previous coding variant associations, suggesting that this gene-trait association could only have been identified from non-coding analyses using WGS.

Conclusion:

Our findings demonstrate the power of WGS to identify coding and non-coding associations not detectable by common variant association studies or exome sequencing.

Impact of rare coding variants on height prediction in a diverse set of >1 million individuals

Authors: L. Ganel[1], J. Kosmicki[1], T. Joseph[1], J. Mbatchou[1], A. Ziyatdinov[1], D. Sharma[1], K. Sun[1], J. Torres[2], J. R. Emberson[2], R. Collins[2], J. Berumen[3], J. Alegre-Díaz[3], R. Tapia-Conyer[3], P. Kuri-Morales[4], O. Melander[5], Y-D. I. Chen[6,7], GHS-RGC DiscovEHR Collaboration, Mexico City Prospective Study, Penn Medicine BioBank, MAYO-RGC Project Generation, Colorado Center for Personalized Medicine-RGC Collaboration, UCLA-RGC ATLAS Collaboration, S. Balasubramanian[1], W. Salerno[1], M. Jones[1], J. Reid[1], A. Baras[1], G. Abecasis[1], J. Marchini[1], M. A. R. Ferreira[1], A. Locke[1]; [1]Regeneron Genetics Ctr., Tarrytown, NY, [2]Univ. of Oxford, Oxford, United Kingdom, [3]Univ. Natl. Autónoma de México, Mexico City, Mexico, [4]Tecnológico de Monterrey, Monterrey, Mexico, [5]Dept. of Clinical Sci. Malmö, Lund Univ., Malmö, Sweden, [6]Lundquist Inst. for BioMed. Innovation, Harbor-UCLA Med. Ctr., Torrance, CA, [7]Dept. of Pediatrics, Harbor-UCLA Med. Ctr., Torrance, CA

Abstract:

Reliable prediction of complex traits is a major aspiration for human genetics, with applications ranging from newborn screening to clinical decision making. While prediction is challenging, height is an ideal model trait for developing methodology, as large genome-wide association studies have identified substantial common variant heritability. Through exome sequencing and joint association analysis of height in 826,066 individuals, we identified 206 genes with rare (MAF<1%) nonsynonymous variants associated with height (P<1.75x10^{-9}) after conditioning on 3,034 independently associated common variants. In a replication sample of 242,787 individuals a polygenic score (PGS) of height-associated common variants explained 24.7% of the variation in height, with the top and bottom percentiles corresponding to heights +9.60 cm and -9.12 cm from the mean, respectively.

While rare variants do not appreciably improve prediction at the population level due to their frequency (mean adjusted R^2 0.367 vs 0.378 when adding rare coding variants), we hypothesized that a subset of individuals with height poorly predicted by common variants and demographic covariates will be a) highly enriched for large effect rare variants, and b) exhibit substantially improved prediction when including rare variants. We focused on 913 individuals for which |common variant prediction error| > 2.5 SD. These individuals are 20.2 times more likely ($P = 8.8\times10^{-16}$) than well-predicted individuals (|common variant prediction error| < 0.5 SD) to carry a variant in one of 16 genes with a pLoF singleton burden height association. By comparison, enrichment is 5.56x for individuals whose observed height is > |2.5 SD| from the observed mean. That is, a large difference between observed and common variant predicted height enriches for rare, large-effect carriers better than the phenotypic extremes.The impact of rare variation in trait prediction is particularly evident among carriers of known deleterious variation. Individuals with ClinVar pathogenic variants in *FGFR3* or *PTPN11* (N = 36), causing achondroplasia and Noonan syndrome respectively, see a reduction in predictive error from 14.5 cm to 3.28 cm by adding individually significant rare variants to the PGS; simultaneously, the number of samples with predicted height within 5 cm of the observed value improved by 13 individuals (36% of carriers).
Overall, we demonstrate that large effect rare variants have a tremendous impact on prediction accuracy at the individual level and represent a valuable addition to personalized medicine efforts.

Machine learning reveals 3D regulatory mechanisms for height-associated haplotypes

Authors: W. Gu[1], J. Capra[2], E. Gilbertson[1], R. Salem[3]; [1]Univ. of California, San Francisco, San Francisco, CA, [2]Univ. of California San Francisco, San Francisco, CA, [3]Univ. of California, San Diego, La Jolla, CA

Abstract:

Background: Variants associated with phenotypes in genome-wide association studies (GWAS) are predominantly non-protein-coding and gene-regulatory in nature. There are many ways a variant could disrupt regulation, and one emerging mechanism is by perturbing the 3D genome, thereby affecting the gene expression of target genes. Many phenotypes, including body height, have enriched SNP heritability within topologically associated domain (TAD) boundaries. Although it has been challenging to experimentally evaluate variants for effects on 3D contacts at the genome scale, the 3D genome structure

can be predicted in silico using machine learning models based solely on DNA sequence information. This provides an opportunity to evaluate 3D genome disruption as a mechanism underlying height-associated loci genome-wide.
Methods: We analyzed genetic regions associated with body height from the largest available GWAS. To enable haplotype-aware analyses, we used the NHLBI Trans-Omics for Precision Medicine (TopMed) sequencing data for ~50,000 participants to impute haplotypes for these loci across diverse populations, including Europeans, Africans, East Asians, South Asians, and Admixed/non-admixed Americans. We then predicted alterations in 3D genome contacts for each common haplotype (count ≥ 30 in TopMed).
Results: We evaluated 9917 height-associated haplotypes: 107 regions (top 1%) exhibited substantial divergence, and 17 (top 0.17%) demonstrated extreme disturbance of the 3D genome. The strongest divergence for a height-associated haplotype was near the LCOR gene on chromosome 10. A specific variant at this locus, rs7477274, likely disrupts 3D genome folding by altering the DNA-binding affinity of the CTCF transcription factor. In addition to LCOR, other significant disruptions were observed near the SLC41A2 and FGF2 regions, both of which are CTCF binding sites and expression quantitative trait loci.
Conclusion: We identify several haplotypes that likely influence variation in body height by modifying 3D genome folding. However, this functional mechanism is relatively rare among height GWAS hits. Our results demonstrate how in-silico mutagenesis based on powerful sequence-based machine learning models provides an efficient approach to fine-map GWAS signals and identify potentially functional variants and mechanisms.

GWAS of infant and early childhood height in up to 70 000 children: Genetic influences on the early phases of childhood growth

Authors: N. Fragoso Bargas[1], A. E. Lupu[1], J. H. Sundfjord[1], R. Karimi[1], P. Njolstad[2,1], M. Vaudel[1,3], S. Johansson[1,2]; [1]Univ. of Bergen, Bergen, Norway, [2]Haukeland Univ Hosp., Bergen, Norway, [3]Norwegian Inst. of Publ. Hlth., Oslo, Norway

Abstract:

The genetic basis of adult height has been extensively characterized. However, there is limited knowledge about the genetic background of normal height development in childhood and its potential health implications. We aimed to shed light on the classical models of childhood growth by characterizing the changing genetic effects of height during the first 8 years of life. GWAS was performed on height in the Norwegian Mother, Father and Child Cohort Study (MoBa). The sample size ranged from 72,473 at birth to 27,813 at 8 years of age. The SNP-based heritability increases steadily over time, starting from 0.16 at

birth to 0.40 at 8 years. Notably, the SNP-based heritability of final height (UK Biobank (UKBB), n= ~500k) is 0.39. In MoBa, genetic correlation analysis showed three blocks of data highly correlated (r_g>0.80) within each other: 1) birth to 6 weeks, 2) from three to 12 months, and 3) from one to 8 years. Furthermore, at both 7 and 8 years the correlations with comparative height size at 10-12 years (UKBB, n= ~320k) and adult height were high (r_g>0.88 and >0.74 respectively). These blocks agree well with Karlberg's childhood growth model. We identified 194 independent signals across all MoBa time points. Notably, 47 SNPs were not significant (p=<0.05/194 and consistent direction) in adulthood. Hierarchical clustering from birth to adulthood divided the signals in four clusters; 1) Birth cluster (n=21): contains SNPs relevant at birth with weak effects at later time points, 2) Infancy cluster (n=51): Includes SNPs which associations are stronger between 3 months and 1 year, 3) "life-time growth" cluster (n=43): showed SNPs consistently strong from 1 to 8 years, with sustained effects at 10 years and adulthood, and 4) childhood growth cluster (n=79): with their most pronounced effects from 1 to 8 years, that are attenuated at age 10 and adult life. The SNPs with stronger effects in early life (clusters 1,2 and 4) were selected for enrichment analysis (GLAD4U). The results showed categories (FDR<0.1) linked to musculoskeletal abnormalities, growth disorders, and glucose traits. In these categories we find genes related to growth diseases: *PAX1* (spinal abnormalities), *IGF2* (Beckwith-Wiedemann syndrome), *IGF1R* (growth retardation), *TBX15* (Cousin Syndrome), *XYLT1* (Desbuquois dysplasia type 2), *GNAS* (Albright hereditary osteodystrophy) among others. We demonstrate that there is a time-dependent dynamic in the genetic architecture of height. These results help to understand the growth biology in early life and shed new light on the molecular underpinnings of the three growth phases proposed by Karlberg.

Session 80: Linking Non-coding Variation to Function via Diverse Epigenetic Mechanisms

Location: Four Seasons Ballroom 1

Session Time: Friday, November 8, 2024, 1:15 pm - 2:15 pm

Single cell multi-omics and 3D genome architecture reveals novel pathways and targets of metabolic dysfunction-associated steatohepatitis

Authors: W. Elison[1], S. Corban[1], C. Miciano[1], R. Lancione[1], A. D'Antonio-Chronowska[1], L. Chang[1], Y. Xie[1], Q. Yang[1], S. Sakane[1], L. Tucciarone[1], R. Melton[1], H. Mummey[1], C. McGrail[1], M. Miller[1], B. Ren[1], D. Brenner[1,2], T. Kisseleva[1], A. Wang[1], K. Gaulton[1]; [1]Univ. of California San Diego, San Diego, CA, [2]Sanford Burnham Prebys Med. Discovery Inst., San Diego, CA

Abstract:

Metabolic dysfunction-associated fatty liver disease (MAFLD) affects a quarter of the global population, which can progress to metabolic dysfunction-associated steatohepatitis (MASH) and lead to complications such as cirrhosis, liver failure, and liver cancer. Despite the high prevalence of MAFLD, there are limited treatments. In this study we used large-scale epigenomic and genetic profiling to understand mechanisms underlying MAFLD progression and identify novel therapeutic targets. We performed single cell assays to measure gene expression, chromatin accessibility, histone modifications, and 3D chromatin architecture in 87 liver samples from non-disease, MAFLD, and MASH donors. We generated an integrated map consisting of >400k cells which mapped to 13 liver cell types and sub-types. We observed extensive changes in the abundance and epigenetic profiles of many liver cell types across MAFLD progression as well as across clinical variables such as fibrosis. In MAFLD hepatocytes display altered glucose and cholesterol metabolism and hepatic stellate cells show altered extra cellular matrix activity. In addition, we identified states within liver cell types enriched in MAFLD such as senescent hepatocytes and fibrogenic myofibroblasts. By leveraging the different modalities, we aim to elucidate gene regulatory and cell-cell communication networks and to determine changes in transcriptional networks and signaling in disease. Genetic variants associated with MAFLD risk and related endophenotypes were enriched in regulatory sites active in specific cell types including hepatocytes, cholangiocytes, fibroblasts, and stellate cells. Oue goal is to annotate quantitative trait loci (QTLs) for genomic modalities in liver cell types and colocalizing QTL signals at loci associated with MAFLD and relevant liver endophenotypes to identify QTLs mediating disease. Overall, these findings provide a

comprehensive reference of cell types in the human liver across stages of MAFLD progression that will enable identifying new therapeutic strategies to treat liver disease.

Machine learning identifies chromatin features that predict the sensitivity of regulatory sequences to inhibition of BAF chromatin remodeling activity

Authors: K. Kang, A. Gulka, D. Gorkin; Emory Univ., Atlanta, GA

Abstract:

BRG1/BRM-associated factor (**BAF**) complexes remodel chromatin at regulatory DNA sequences to creates regions of “accessible chromatin” depleted of packaging nucleosome particles. Accessible chromatin enables transcription factors (**TFs**) to bind their target sequence motifs and is a near-universal characteristic of *cis*-regulatory sequence elements (**cREs**). BAF complexes have come under intense study in recent years because mutations in BAF complex subunits have been linked to a variety of cancers and neurodevelopmental conditions.Several studies using genetic and/or pharmacological inhibition of BAF complexes have found that thousands of cREs are dependent on continuous BAF activity to maintain their chromatin accessibility. However, these studies have also shown that not all cREs are dependent on BAF activity to maintain chromatin accessibility. This raises the central question of this study: *What distinguishes cREs that are dependent on BAF for chromatin accessibility from those that are not?*To better understand the features that distinguish BAF-dependent and BAF-independent cREs we turned to the GM12878 cell line which has an available compendium of ChIP-seq data from ENCODE that map the binding profiles of hundreds of TFs and dozens of histone post-translational modifications (**PTMs**). We first treated GM12878 with an allosteric small molecule inhibitor of BAF (BRM014) and performed ATAC-seq to measure the resulting changes in chromatin accessibility genome-wide. We then trained random forest machine learning model to predict significant local chromatin accessibility loss upon BAF inhibition based on hundreds of chromatin features measured by ENCODE.The trained random forest model was able to identify BAF-dependent vs BAF-independent peaks at a 76% accuracy with reproducibility across biological replicates, suggesting that chromatin features are a useful predictor of BAF dependent accessible regions. Furthermore, feature importance analyses were able to identify the most important TFs and PTMs that determine BAF-dependent accessible regions. We found several lineage-specific TFs that were predictive of sensitivity to BAF inhibition, while CTCF binding and promoter-related chromatin were predicted to lack BAF sensitivity. This provides much-needed insights on the molecular

consequence of BAF loss-of-function and demonstrates a powerful approach to further dissect the relationship between TF binding, chromatin state, and BAF activity.

Response sQTLs in primary human chondrocytes identify novel putative osteoarthritis risk genes

Authors: S. Byun[1], P. Coryell[1], N. Kramer[1], D. Susan[1], E. Thulson[1], Y. Sahin[1], S. Chubinskaya[2], R. F. Loeser[1], B. Diekman[3], D. Phanstiel[1]; [1]Univ. of North Carolina at Chapel HIll, Chapel Hill, NC, [2]The Univ. of Texas Med. Branch, Galveston, TX, [3]Univ. of North Carolina at Chapel HIll & Univ. of North Carolina and North Carolina State Univ., Chapel Hill & Raleigh, NC

Abstract:

Osteoarthritis (OA) affects over 500 million people worldwide and is a leading cause of disability. Despite extensive research, current treatments are limited due to the poorly understood molecular mechanisms. Genome-wide association studies (GWAS) have identified over 100 loci associated with OA, but the effects of these variants remain unclear due to linkage disequilibrium and their non-coding nature. Aberrant splicing linked to genetic variation is known in diseases like rheumatoid arthritis, Alzheimer's, and Parkinson's, but has not been extensively explored in OA. This study aims to elucidate the relationships between splicing QTLs (sQTLs) and OA risk variants to improve understanding and identify potential therapeutic targets.

We generated RNA-seq data from primary human chondrocytes isolated from 101 tissue donors after 18 hours of treatment with either PBS (control) or fibronectin fragment (FN-f)-a known OA trigger. We detected and quantified differential splicing events using the LeafCutter algorithm (padj < 0.05, |PSI| > 0.2), identifying 322 genes with differential splicing events corresponding to 335 intron junctions between the PBS and FN-f treated samples. Differentially spliced genes were enriched for OA-related gene ontology terms related to extracellular matrix organization, inflammatory response, and cartilage development. Using QTLtools (padj < 0.05) and a regression model that corrects for sequencing batch, ancestry, age, and multiple principal components, we identified 7,187 sQTLs impacting 3,067 sGenes. We focused on identifying response sQTLs by testing the interaction effects of genotype with the treatment condition. Explicitly testing for genotype/condition interaction effects revealed 491 and 499 sGenes specific to PBS or FN-f, respectively. Notable FN-f response SNP-sGene pairs include rs2834167-A associated with *IFNAR2* (deltaPSI=0.23, MAF=0.23) and rs11605232-T associated with *AMPD3 (deltaPSI = 0.28,* MAF= 0.48*)*. The 54 sQTLs were in linkage disequilibrium ($r^2 > 0.5$) with an

OA risk variant, including *NEK4* and *ITIH1*, which have been previously implicated in OA. One gene of interest was *SNRNP70*. Alternative exon 8 was skipped in both OA tissue and FN-f treatment. CRISPR/Cas9 deletion of exon 8 showed that the modified splicing pattern closely mimics OA-related pathways observed in wild-type versus edited comparisons, suggesting a mechanism for this splicing event in OA.
This study highlights the significant impact of alternative splicing in the pathogenesis of OA and its potential for discovering therapeutic targets. The findings provide a valuable resource for future research and therapy development.

Massively parallel reporter assay highlights the importance of B cell activation in uncovering latent QTLs, especially for eQTLs

Authors: S. Pasula, Y. Fu, D. A. Murphy, J. A. Kelly, R. C. Pelikan, K. L. Tessneer, K. Grundahl, P. M. Gaffney; Oklahoma Med. Res. Fndn., Oklahoma City, OK

Abstract:

The majority of studies to identify causal variants and quantitative trait loci (QTLs) have been performed in the quiescent cell state and, therefore, do not identify effects that regulate mechanisms only in an activated state. To explore this, we activated B cells to identify epigenetic QTLs both in the activated and resting states, including CTCF (ctQTLs), histone (hQTLs), chromatin accessibility (caQTLs) and transcription (eQTLs) modifications. In this study, we used a massively parallel reporter assay (MPRA) to determine which epigenetic QTLs likely contribute to disease mechanisms in B cells. We designed a MPRA oligo library with 24,816 QTL sequences using 200bp of the hg38 genomic sequence flanking the reference and alternate alleles of each variant. The final MPRA library was transfected into six experimental replicates in a lymphoblastoid cell line and performed as described previously (Fu et al., HGG Adv. 2024). We first evaluated the library for significant expression modulating sequences (emSeq). A total of 5147 sequences (20.7%; 2562 variants) produced an emSeq, with the highest proportion found in ctQTLs (22%) and the lowest proportion in eQTLs (16%). EmSeqs were then assessed for allele-specific transactivation potential (emVars). A total of 530 variants (20.7%) were identified as emVars, with caQTLs producing the highest proportion (21.7%), followed by ctQTLs (20.9%), hQTLs (20.0%), and eQTLs (15%). As a whole, B cell epigenetic emVars were significantly associated with several GO biological processes including the regulation of stress-activated MAPK cascade, stress-activated protein kinase signaling cascade, JNK cascade and JUN kinase cascade, consistent with B cell activation pathways. When evaluating QTLs that were originally identified in the activated vs resting states for emVars,

we found an interesting observation among eQTLs, with 19% of activation-dependent eQTLs being emVars compared to only 5% of resting eQTLs being emVars. Such a discrepancy was not observed for emVars of epigenetic QTLs between the activated and resting state (caQTLs 23% vs 21%, hQTLs 22% vs 19% and ctQTLs 18% vs 26%). These results suggest, that within the confines of our experimental design, a substantial proportion of eQTLs are likely to be missed in the analysis of resting cells, thus stressing the need to study activated B cells to expose latent eQTLs.

Session 81: Rare Variants and Admixture Modeling in Diverse Population

Location: Four Seasons Ballroom 2&3

Session Time: Friday, November 8, 2024, 1:15 pm - 2:15 pm

Large-scale admixture mapping in the *All of Us Research Program* improves the characterization of cross-population phenotypic differences

Authors: R. Mandla[1], Z. Shi[1], K. Hou[1], Y. Wang[2], E. Atkinson[3], A. Martin[2], B. Pasaniuc[1]; [1]Univ. of California, Los Angeles, Los Angeles, CA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]Baylor Coll. of Med., Houston, TX

Abstract:

Admixed individuals have largely been understudied in medical research due to their complex genetic ancestry and their underrepresentation in large genetic datasets. However, the consideration of admixture differences can shed light on the genetic underpinnings of cross-population phenotypic variation. To this end, we performed local ancestry inference in the *All of Us Research Program* (AoU; N=245,394) with 5 1000 Genome reference populations and identified genetically-inferred (GIA) two-way admixed African (AFR) and European (EUR) populations (AFR/EUR; N=48,921). We then performed admixture mapping (ADM) with the GIA AFR/EUR individuals for 23 traits.

We identified 59 loci where AFR ancestry was associated with one of the traits ($P<5x10^{-5}$), for a total of 72 associations. Of these loci, 12 (20%) demonstrated evidence of pleiotropy with multiple traits. Additionally, trait genetic correlations across these loci matched known genetic and pathophysiologic relationships between traits, such as observed positive correlation between HbA1c and type 2 diabetes (T2D; pearson R=0.64) ADM effect sizes.

To assess if ADM associations are driven by genome-wide association study (GWAS) signals, we re-ran ADM conditioning on GWAS signals near an ADM-associated loci. In total, 18 (25%) of the ADM associations decreased in significance ($P>1x10^{-4}$) after this adjustment. We also found significant absolute differences in the allele frequencies (AF) of these variants between AoU participants of more homogeneous EUR and AFR ancestry ($P=2.5x10^{-5}$), suggesting ADM is powered to identify loci with AF heterogeneity between populations.

Furthermore, we found a novel positive association between AFR ancestry and end-stage kidney disease at region 9q21.33 (OR=1.40, $P=3.6x10^{-5}$), even after adjusting for T2D status (OR=1.41, $P=3.5x10^{-5}$). This locus contains the gene *NTRK2* and has previously been linked

to estimated glomerular filtration rate differences. *NTRK2* additionally is highly expressed in the thyroid and previous burden analyses using rare missense variants in UK Biobank found a positive association between rare, predicted damaging variants in *NTRK2* and increased kidney failure (OR=5.30, P=0.02).
In summary, we performed the largest admixture mapping effort of admixed AFR/EUR individuals, identifying loci exhibiting ancestry-correlated heterogeneity which may play a role in phenotypic differences between populations. These results further motivate the expansion of both common and rare variant analyses in underrepresented populations to identify novel genetic associations and new, potentially actionable insights into disease biology.

Multi-ancestry GWAS for hypermobile Ehlers-Danlos Syndrome

Authors: W. He[1,2], A. Seth[1,2], R. E. Handsaker[2,3], V. Janoušek[4,5], M. K. Galindo[6], D. Subramanian[7], C. A. Francomano[8,9], W. Gandy[9], 23andMe Research Team, HEDGE Consortium, C. M. Laukaitis[10,11], J. N. Hirschhorn[1,2,3]; [1]Boston Children's Hosp., Boston, MA, [2]Broad Inst. of MIT and Harvard, Cambridge, MA, [3]Harvard Med. Sch., Boston, MA, [4]Biodviser Ltd, Oldham, Manchester, United Kingdom, [5]Charles Univ. and Gen. Univ. Hosp. in Prague, Prague, Czech Republic, [6]Univ. of Arizona, Tucson, AZ, [7]Peter MacCallum Cancer Ctr., Melbourne, Victoria, Australia, [8]Indiana Univ. Sch. of Med., Indianapolis, IN, [9]The Ehlers-Danlos Society, London, UK and, NY, [10]Carle Illinois Coll. of Med., Urbana, IL, [11]Carle Hlth., Urbana, IL

Abstract:

Hypermobile Ehlers-Danlos (hEDS) is a rare connective tissue disorder characterized by joint hypermobility and other comorbidities. Despite familial clustering, the underlying genetic architecture of hEDS remains unknown, and no genome-wide association studies (GWAS) of hEDS have been performed. To begin to elucidate the genetic architecture of hEDS, we used whole genome sequence (WGS) data from 988 hEDS cases from the hEDS Genetic Consortium (HEDGE) study and 4,940 ancestry-matched controls from All Of Us Research Program to perform GWAS and, separately, exome wide burden tests and copy number variant analyses. **Study design and Methods**: We developed MANCS (Multi-Ancestry Nearest Control Selection), a method that selects multiple ancestry-matched controls for each case by identifying the nearest neighbors (weighted Mahalanobis distance) in principal component (PC) space, allowing us to include all HEDGE samples regardless of genetic ancestry. We identified five ancestry-matched controls for each case from 239,460 participants in All of Us with WGS data. After QC to harmonize variant calling, we performed logistic regression for variants with allele frequency (AF) > 1% or allele count

>100 in All of Us and allele count >5 in hEDS cases, adjusting for sex and PCs. Based on the GWAS summary statistic, we used LD score regression to estimate the common variant heritability of hEDS. Finally, we used GWAS summary statistics from 64,143 participants of 23andMe with self-assessed Beighton score, a measure of joint hypermobility, to test whether polygenic scores (PGS) for Beighton score are elevated in European-ancestry HEDGE cases compared with matched controls. **Results**: In a well-controlled GWAS, one locus on chromosome 5 reached the traditional genome-wide significance threshold of $5x10^{-8}$ for association with hEDS (lead variant AF hEDGE vs controls: 0.9% vs 0.1%, OR =8.41, $p=3.41x10^{-8}$). We estimated hEDS to be modestly heritable (~15±8%) with respect to common variants. Finally, hEDS patients had significantly higher Beighton score PGS compared to the controls ($p=2.08\ x10^{-10}$), providing the first evidence of a polygenic basis for hEDS, shared with polygenic risk for increased joint mobility in a large direct-to-consumer sample. **Conclusion**: By matching hEDS cases to All of Us controls, and including all samples regardless of ancestry, we identified a potential novel hEDS-associated locus and observed polygenic contributions to hEDS that are shared with joint hypermobility in a large direct to consumer sample. These findings, combined with our group's hEDS rare variant and CNV analysis, provide valuable insights into the genetic underpinnings of hEDS.

Rare variant associations and fine-scale population structure in the Genes & Health Study of >44,000 British South Asians

Authors: K. Walter[1], C. DeBoever[2], G. Kalantzis[1], T. Heng[1], H. Kim[3], E. Fauman[3], I. Popov[1], V. Iyer[1], S. Bidi[4], K. Catalano[5], K. Hunt[4], B. Jacobs[4], K. Kundu[6], R. Mathur[4], C. Morton[4], S. Mozaffari[7], A. Musolf[8], J. Russell[4], M. Spreckley[4], M. Traylor[9], R. Turner[10], Genes & Health Industry Consortium, Genes & Health Research Team, R. Trembath[11], S. Finer[4], H. C. Martin[1], D. A. van Heel[4]; [1]Wellcome Sanger Inst., Hinxton, United Kingdom, [2]Takeda Dev. Ctr. Americas, San Diego, CA, [3]Pfizer, Cambridge, MA, [4]Queen Mary Univ., London, United Kingdom, [5]Bristol Myers Squibb, Shutesbury, MA, [6]AstraZeneca, Cambridge, United Kingdom, [7]Maze Therapeutics, San Jose, CA, [8]Merck, Cambridge, MA, [9]Novo Nordisk Res. Ctr., Oxford, United Kingdom, [10]GSK, Stevenage, United Kingdom, [11]King s Coll. London, London, United Kingdom

Abstract:

Genes & Health is a large community-based study of more than 60,000 individuals of self-identified Bangladeshi and Pakistani ancestry who were recruited in London, Manchester, and Bradford, in the UK. The study currently comprises 44,026 individuals with whole-

exome sequence data and electronic health record (EHR) data from primary and secondary care provided by the National Health Service (NHS).
We applied exome-wide single-variant and gene-based association tests using REGENIE to 675 binary phenotypes with at least 100 cases in each trait. We set our significance threshold at $p<1x10^{-9}$ which we estimate corresponds to FDR<3.8% for single variant and FDR<5% for gene-based tests, based on tests that only comprise synonymous variants. We found 581 significant single-variant associations between 339 genetic variants and 41 phenotypes, and 118 gene associations between 16 genes and 16 phenotypes using different sets of predicted deleterious variants. Significant gene-based results involve mostly associations of the *HBB* gene with thalassaemia and anaemias, as well as associations of the *LDLR* gene with disorders of lipoprotein metabolism and other lipidaemias, while single-variant associations additionally involve significant associations of variants in the HLA region with various autoimmune conditions.
Previous work has shown that the fine-scale population structure in British Pakistanis is dominated by endogamous *biraderi* groups with varying degrees of bottleneck. We explored this within Genes & Health by clustering unrelated Genes & Health participants together with individuals from reference populations according to shared identity-by-descent (IBD) segments. We identified several clusters with Pakistani ancestry, and we are currently characterising the strength of the bottleneck within those clusters and searching for founder variants associated with phenotypes. The Genes & Health cohort represents a unique opportunity to characterise rare variant associations across the phenotypic spectrum in a diverse population.

Network comparison of ancestry-specific genetically correlated diseases in a meta-analysis of phenome-wide association studies from 1 million individuals

Authors: J. Woerner[1], Y. Nam[1], M. Levin[1], S. Koyama[2], A. Rodriguez[3], B. Voight[1], R. Madduri[3], P. Natarajan[4], S. Damrauer[1], Y. Xu[5], D. Kim[1], A. Verma[1]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Broad Inst., Cambridge, MA, [3]Argonne Natl. Lab., Lemont, IL, [4]Massachusetts Gen. Hosp., Boston, MA, [5]Vanderbilt Univ, Nashville, TN

Abstract:

Biobanks have accelerated our ability to investigate genetic associations across thousands of traits and diseases within given populations. While several studies have conducted genome-wide association studies (GWAS) on numerous traits and diseases, these have primarily focused on European ancestry populations. This limitation restricts our understanding of how the shared genetic architecture of diseases varies between different

populations. We applied a network-based approach to a genome-wide association meta-analysis of >700 phecode traits across diverse populations, including African (AFR, N=128k) and European (EUR, N=871k) reference populations from the UK Biobank and the Million Veteran Program, encompassing 1 million individuals. We calculated pairwise genetic correlation estimates for heritable diseases (h^2 $p<0.05$) using linkage disequilibrium score regression (LDSC) within EUR and AFR individuals separately. Using these genetic correlations, we developed ancestry-specific networks, where 209 nodes represent heritable diseases/traits and edges represent their genetic correlations. Our network analysis incorporates structural equivalence and eigenvector centrality, and 10,000 repetitions of Louvain clustering identified similar substructures in EUR and AFR networks. Despite some differences, clusters were largely similar, with consistent modules grouping almost all 52 cardiometabolic diseases together, and likewise the 25 musculoskeletal diseases. However, 26% of nodes routinely clustered differently across populations, particularly disorders affecting sense organs. Then, we applied the core-periphery model to the networks, distinguishing between central ('core') and marginal ('peripheral') nodes. We observed notable connectivity differences for diseases known to vary genetically across populations, such as glaucoma and hypothyroidism. While most diseases maintained consistent core-periphery structures, heart valve disorders were core in AFR but peripheral in EUR, whereas functional digestive disorders showed the opposite pattern. Overall, this analysis reveals similar graph substructures at the global level, but important differences at the local level. The next step is to leverage these disease networks to generate integrated polygenic and comorbidity risk scores using a network-based approach like netCRS. These findings emphasize the critical need to include diverse research cohorts to study the shared genetic underpinnings of disease. These approaches are clinically relevant for designing studies for multimorbidity and risk prediction tailored to specific population groups.

Session 82: Read All about It: Transcriptomic Insights from New Sequencing Technologies

Location: Room 501

Session Time: Friday, November 8, 2024, 1:15 pm - 2:15 pm

Identifying pathogenic variants that cause Mendelian conditions using long-read transcript sequencing

Authors: Y-H. H. Cheng[1], A. Sedeno Cortes[1], J. Ranchalis[1], K. Munson[1], A. O'Donnell-Luria[2], E. Blue[1], J. X. Chong[1], M. J. Bamshad[1], A. Stergachis[1]; [1]Univ. of Washington, Seattle, WA, [2]The Broad Inst., Boston, MA

Abstract:

Background: Short-read transcriptome sequencing has improved the discovery of pathogenic variants causing rare diseases, but this approach often cannot resolve the underlying molecular mechanism - information necessary for translating genomic discoveries into genomic therapies. Paired sample sequencing with and without nonsense-mediated decay (NMD) inhibition can resolve the molecular mechanism by detecting transcripts undergoing NMD, indicating loss-of-function (LOF) variants. However, because full-length transcripts are challenging to reconstruct with short-reads, it can be difficult to fully capture the impacted isoform using short-read sequencing. To address this, we evaluated a long-read transcript sequencing approach with and without NMD inhibition, aiming to achieve transcriptome-wide resolution of both the genetic and molecular basis in individuals with rare genetic disorders. **Methods:** We benchmarked the utility of long-read transcript data on fibroblast lines from 4 participants with known rare transcript disrupting variants, using data from 23 participants with unsolved rare diseases as controls. For each participant, RNA was isolated from untreated cells and cells treated with the NMD inhibitor cycloheximide (CHX). Bulk full-length transcript sequencing (Kinnex) was performed to achieve ~5 million full-length transcripts per sample. We developed a novel analysis package (NMDSeqR) for processing this data that evaluates transcriptome-wide for isoforms/genes subjected to NMD, isoforms/genes with aberrant expression, or genes harboring novel isoforms. **Results:** Benchmarking against known cases, we successfully identified 4 known aberrant transcript events using our transcriptome-wide approach, such as an NMD-inducing splice donor variant in *HARS1*, a promoter structural variant causing decreased *VTA1* expression, an NMD-inducing branch point variant in *MFN2*, and an exon-skipping cryptic splice donor variant in *SET*. Our long-read data also demonstrated that a putatively pathogenic splicing variant in *SEC31A* actually impacted a transcript isoform

that did not include the *SEC31A* coding sequence, deprioritizing it as a candidate variant. **Conclusion:** Our study provides proof-of-concept that long-read transcript sequencing enables the de novo discovery of rare pathogenic variants associated with novel isoform usage and NMD using a transcriptome-wide approach. Furthermore, by including both an NMD readout and full-length transcript data, we can identify these events transcriptome-wide with lower sequencing coverage, and can directly ascertain the exact aberrantly spliced isoform.

Single-Cell Omics for Transcriptome CHaracterization (SCOTCH): isoform-level characterization of gene expression through long-read single-cell RNA sequencing

Authors: Z. Xu[1], H-Q. Qu[2], J. Chan[3], C. Kao[2], H. Hakonarson[4], K. Wang[5]; [1]Univ. of Pennsylvania, Philadelphia, PA, [2]Children s Hosp. of Philadelphia, Philadelphia, PA, [3]The Children s Hosp. of Philadelphia, Philadelphia, PA, [4]Children's Hosp. of Philadelphia, Philadelphia, PA, [5]Children’s Hosp. of Philadelphia, Philadelphia, PA

Abstract:

Recent development involving long-read single-cell transcriptome sequencing (lr-scRNA-Seq) represents a significant leap forward in single-cell genomics. With the recent introduction of R10 flowcells by Oxford Nanopore, we propose that previous computational methods designed to handle high sequencing error rates are no longer relevant, and that the prevailing approach using short reads to compile "barcode space" (candidate barcode list) to de-multiplex long reads are no longer necessary. Instead, computational methods should now shift focus on harnessing the unique benefits of long reads to analyze transcriptome complexity. In this context, we introduce a comprehensive suite of computational methods named Single-Cell Omics for Transcriptome CHaracterization (SCOTCH). Our method is compatible with the single-cell library preparation platform from both 10X Genomics and Parse Biosciences, facilitating the analysis of special cell populations, such as neurons, hepatocytes and developing cardiomyocytes. We specifically re-formulated the transcript mapping problem with a compatibility matrix and addressed the multiple-mapping issue using probabilistic inference, which allows the discovery of novel isoforms as well as the detection of differential isoform usage between cell populations. We evaluated SCOTCH through analysis of real data across different combinations of single-cell libraries and sequencing technologies (10X + Illumina, Parse + Illumina, 10X + Nanopore_R9, 10X + Nanopore_R10, Parse + Nanopore_R10), and showed its ability to infer novel biological insights on cell type-specific isoform expression. These

datasets enhance the availability of publicly available data for continued development of computational approaches. In summary, SCOTCH allows extraction of more biological insights from the new advancements in single-cell library construction and sequencing technologies, facilitating the examination of transcriptome complexity at the single-cell level.

Applications of long-read RNA sequencing improves the design and interpretability of RNA-based therapeutics

Authors: E. Gustavsson[1,2,3], J. Evans[4,5], A. Fairbrother-Browne[6,2], H. Macpherson[1], J. Brenton[1,2], K. Montgomery[5], M. Grant-Peters[1,2], C. Arber[7], S. Wray[7], Z. Jaunmuktane[7], N. Wood[5], H. Houlden[5], J. Hardy[5,8], S. Gandhi[4,5], M. Ryten[1,2,3,9]; [1]Great Ormond Street Inst. of Child Hlth., Univ. Coll. London, London, United Kingdom, [2]UK Dementia Res. Inst. at The Univ. of Cambridge, Cambridge, United Kingdom, [3]NIHR Great Ormond Street Hosp. BioMed. Res. Ctr., Univ. Coll. London, London, United Kingdom, [4]The Francis Crick Inst., London, United Kingdom, [5]UCL Queen Square Inst. of Neurology, Univ. Coll. London, London, United Kingdom, [6]Great Ormond Street Inst. of Child Hlth., Univ. Coll. London, LONDON, United Kingdom, [7]UCL Queen Square Inst. of Neurology, Univ. Coll. London, LONDON, United Kingdom, [8]UCL Dementia Res. Inst., Univ. Coll. London, London, United Kingdom, [9]Dept. of Clinical NeuroSci.s, The Univ. of Cambridge, Cambridge, United Kingdom

Abstract:

Long-read RNA sequencing enables the analysis of complete transcript structures, thus allowing for granular transcriptome analysis. There is growing interest in RNA isoform diversity as both a cause of disease and a therapeutic target, particularly for diseases of the central nervous system (CNS). It is becoming increasingly clear that there are high cell type biases in transcript use, which could be utilized for targeting therapies. This has occurred simultaneously with advances in antisense oligonucleotide (ASO) technologies, which enable precise targeting of transcripts. Therefore, we implemented long-read RNA sequencing to improve ASO design and evaluation in neurodegenerative diseases. Parkinson's disease (PD) is an exemplar of a prevalent neurodegenerative disorder that is pathophysiologically linked to the *SNCA* gene, which encodes the α-synuclein (αSyn) protein, a key player in PD pathogenesis. We used targeted long-read RNA sequencing in iPSC-derived midbrain dopaminergic neurons, the cell type progressively lost in PD, from patients with *SNCA* mutations. We showed that 75% of expression at the *SNCA* locus originates from transcripts with alternative 5' and 3' untranslated regions (UTRs) and 10%

from previously unannotated open reading frames, detectable in the human postmortem brain. Defining the 3' UTRs enabled the rational design of ASOs targeting 90% of *SNCA* transcripts, effectively reversing PD pathology in *SNCA* mutation neurons, including protein aggregation, mitochondrial dysfunction, and toxicity. Furthermore, we analyzed *SNCA* transcription using long-read RNA sequencing in iPSC-derived glial cell types and employed single-nucleus RNA sequencing (snRNA-seq) of post-mortem brain tissue from patients with PD to assess biases in 3'UTR usage across different cell types. We found significant differences that were leveraged to design ASOs capable of targeting specific cellular subtypes. Understanding gene and transcript structures is crucial for variant interpretation and therapeutic strategies like ASO, and the applications of long-read sequencing provide an accurate method for achieving this.

Combined long- and short-read RNA sequencing of pathogen stimulated primary immune cells identifies the expression of uncharacterized genes and transcripts

Authors: E. Vorsteveld[1], R. Salz[1], C. van der Made[1], S. Kersten[1], M. Stemerdink[1], T. Riepe[2], T-h. Hsieh[1], M. Mhlanga[1], M. Netea[3], P-J. Volders[4], P. Hoen[1], A. Hoischen[3]; [1]Radboud Univ. Med. Ctr., Nijmegen, Netherlands, [2]Radboudumc, Nijmegen, Netherlands, [3]Radboud Univ. Med. Ctr., Nijmegen, Netherlands, [4]Ghent Univ., Ghent, Belgium

Abstract:

Background: Immune responses are shaped by the nature of infections and by inter-individual variability, contributing to differential susceptibility to infections and to various diseases with an inflammatory component. Dynamic transcript and protein expression in a range of cells responsible for the innate immune response is important to shape the first line of defense against a wide variety of pathogens.Methods: Here we perturbed immune-cells with in vitro pathogen exposure. We stimulated PBMCs from 5 healthy donors for 4h or 24h with LPS, S. aureus, Poly(I:C) or C. albicans with RPMI medium as control, resulting in a total of 52 studied samples. We performed short read sequencing using Lexogen QuantSeq in all samples, as well long read sequencing using PacBio IsoSeq in a subset of samples.Results: Short-read sequencing reveals common and distinct genes and pathways expressed during immune responses to different pathogens. Beside well-established genes, we highlight uncharacterized genes *KIAA0040* and *FAM49A*, which show up to 2.3-fold increased expression after pathogen exposure and are co-expressed with modules of established immune genes. Long-read sequencing revealed 47.7% novelty in the transcriptomes. We found widespread isoform switching induced upon pathogen stimulation. We highlight novel transcripts of *NFKB1* and *CASP1* that may indicate novel

immunological mechanisms. Conclusions: Transcriptome profiling of pathogen-stimulated immune cells using paired short- and long-read approaches highlights candidate immune genes and identifies novel transcripts, revealing a more complex transcriptome landscape following pathogen exposure than previously appreciated.

Session 83: Splice Splice Baby: Isoform Expression in Health and Disease

Location: Room 505

Session Time: Friday, November 8, 2024, 1:15 pm - 2:15 pm

An atlas of expressed transcripts in the prenatal and postnatal human cortex ⭐

Authors: R. Bamford[1], S. Leung[1], V. Chundru[1], A. R. Jeffries[1], J. P. Davies[1], A. Franklin[1], X. Chen[2], A. McQuillin[2], N. Bass[1], E. Walker[1], P. O'Neill[1], E. Pishva[1], E. L. Dempster[1], E. Hannon[1], C. F. Wright[1], J. Mill[1]; [1]Univ. of Exeter, Exeter, United Kingdom, [2]Univ Coll. London, London, United Kingdom

Abstract:

Alternative splicing enables multiple RNA isoforms to be produced from a single mRNA precursor, resulting in high levels of transcriptomic and proteomic diversity. Alternative splicing is an important mechanism in the central nervous system and has been widely implicated in brain disorders. Long read sequencing approaches can be used to generate full-length transcript sequences and fully characterise isoform diversity. We used Oxford Nanopore Technologies (ONT) transcriptome sequencing to profile transcript diversity across human brain development and aging profiling human cortex tissue dissected from fetal, neonatal and adult donors (n = 47, aged 6 weeks post-conception to 83 years old). Sequencing data was processed using a bespoke analysis pipeline developed by our group for isoform visualisation and quantification. We identify widespread transcript diversity in the developing cortex and observe many novel transcript isoforms, with 55% of transcripts in protein-coding genes not previously characterized in existing annotations. A large proportion of these previously uncharacterised transcripts have high coding potential and corresponding peptides were detected in independent proteomic data generated on human cortex. We used fluorescence-activated nuclei sorting (FANS) to purify specific cortical cell-types to explore cell-specific transcript expression across development, finding that many novel transcripts were specific to neurons, oligodendrocytes or microglia. Our findings have important implications for genetic studies of human disease. Novel putative coding sequences are highly conserved and overlap *de novo* mutations identified in whole-genome sequencing studies. Within the novel coding regions of genes associated with dominant developmental disorders, we find 15 *de novo* variants from undiagnosed neurodevelopmental disorder trios in the Genomics England 100,000 genomes project including two stop-gained variants in patients with phenotypes consistent with the corresponding single gene disorder. This highlights the potential clinical utility of

long-read transcriptomics. Our findings underscore the potential of novel coding sequences to harbor clinically relevant variants, offering new insights into the genetic architecture of human disease. Our cortical transcript annotations are available as a resource to the research community via an online database.

Uncovering the brain-specific genetic regulation of splicing by mapping splicing quantitative trait loci in 10,887 post-mortem brain RNA-seq samples

Authors: A. Real[1], K. Babupanneerselvam[2], B. Z. Muller[2], W. H. Cuddleston[2], J. Humphrey[2], T. Raj[2], D. Knowles[3]; [1]New York Genome Ctr., New York, NY, [2]Icahn Sch. of Med. at Mount Sinai, New York, NY, [3]New York Genome Ctr. & Columbia Univ., New York, NY

Abstract:

Alternative splicing, a tightly regulated process in cells and tissues, allows genes to produce multiple transcript isoforms and proteins by selecting different exon combinations. In the brain, splicing enhances neuronal diversity and plasticity across regions, and its dysregulation is associated with neurological disorders. However, the underlying mechanisms remain unclear. We investigated splicing patterns in the brain to decipher their genetic regulation across tissues and role in neurodevelopmental and neurodegenerative processes. Leveraging our "BigBrain" resource, we harmonized 10,887 unique RNA-sequencing samples from 12 datasets spanning 7 brain regions. This enabled comprehensive mega-analyses across brain regions and ancestries for both expression and splicing quantitative trait loci (eQTL and sQTL), bolstering statistical power and facilitating precise detection of subtle splicing effects. Using the LeafCutter algorithm, we clustered splicing junctions in a cohort- and tissue-aware manner to detect splicing patterns between samples from different brain cohorts. Principal component analysis highlighted cohorts, brain regions, and library preparation as primary drivers of gene expression and splicing differences across BigBrain tissues. To systematically address this variation and achieve optimal data harmonization, we compared different approaches: 1) global batch correction versus per tissue-cohort correction, 2) fixed-effect, random-effect, and multivariate adaptive shrinkage-based meta-analyses versus mega-analysis across cohort-tissue pairs. Our initial meta-analysis, focusing on a stringent subset of junctions from European individuals shared by at least 10 cohorts, unveiled 48,388 cis-sQTLs for 12,415 genes (FDR 5%) in 4,361 individuals, corresponding to a remarkable 94% discovery rate. Using Benjamini-Hochberg correction, we achieved an 82% replication rate of cis-sQTL meta-analysis in a held-out cohort, with up to 26% for cis-sQTL replication in the meta-analysis, consistent with findings from prior research. We further explored MashR

methods to enhance the accuracy and robustness of our meta-analysis by taking into account tissue correlation and population structure. Finally, we conduct colocalization analysis to aid in the interpretation of neurological GWAS loci. The inclusion of different published transcriptomics datasets within the BigBrain project will shed light on diverse aspects of gene expression, alternative splicing, and the genetic mechanisms governing splicing across distinct brain regions.

A high throughput splicing assay for characterization of rare variants of unknown significance

Authors: R. Sangermano, E. M. Place, S. Patankar, S. Mehrotra, R. M. Huckfeldt, J. Comander, E. A. Pierce, Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR) Consortium, K. M. Bujakowska; Massachusetts Eye and Ear, Boston, MA

Abstract:

Purpose: Understanding the functional consequences of variants of unknown significance (VUSs) is a major challenge in clinical genetics. The lack of definitive interpretation of VUSs limits genetic diagnoses, impairs accuracy of prognoses, and reduces access to the emerging genetically informed treatments. Variants outside of the canonical splice sites may alter pre-mRNA splicing, leading to disease. The purpose of this study was to evaluate if rare VUSs identified in various Mendelian disease cohorts lead to aberrant splicing. **Methods:** We developed a high throughput splicing assay (HTSA), which relies on a split *GFP* minigene. We evaluated a total of 2288 rare variants from different Mendelian disease cohorts, (1037 from inherited retinal degeneration (IRD) genes, 1251 from the GREGoR consortium), and 720 canonical splice site controls. Reference and variant oligos (300bp-long) were cloned as a pool into the splicing assay construct and expressed in landing pad RCA7 HEK293T cells. Exon inclusion led to GFP disruption while exon skipping led to GFP reconstitution, enabling sorting of GFP positive and negative cells by FACS. Sequencing read-based counts of each oligo per GFP cell population enabled calculation of the exon-inclusion ratio (% spliced in) and subsequent comparison of the reference and variant sequences. Other aberrant splicing events were detected by amplicon sequencing of minigene transcripts. **Results:** We detected significant splicing alteration (>30% of aberrant transcripts) for 340 variants, 43% of which were not predicted to alter splicing by the SpliceAI algorithm (SpliceAI score <0.2). Of the splice-altering variants 25% were missense, 9% were synonymous, and 29% were deeper intronic. In the IRD cohort, we finalized molecular diagnosis in 11/380 unsolved cases. Variants in *CEP164* (c.1233+5G>A), *CHM* (c.940G>A), *CNGB3* (c.1208G>A, two

cases), *EYS* (c.1056+3A>C), *FLVCR1* (c.1026T>C), *IFT172* (c.2788-3C>G), and *RPGR* (c.310G>A and c.310+7T>G), led to exon skipping in 62-93% of transcripts. Variants in *CNGB1* (c.2893-7G>A), and *RPGR* (c.1754-3C>G), led to partial intron retention in >60% of transcripts. **Conclusions:** HTSA offers a robust method to study the effects of VUSs on splicing. However, interpretation of the impact of splicing variants requires additional considerations, such as the extent of the splicing defect (percent of aberrant transcripts), the change of the reading frame, protein domains encoded by the skipped in-frame exon, mode of inheritance and pathogenicity of the allele in *trans* for recessive genes, and gene dosage sensitivity.

Defining the landscape of poison exon splicing events in the human brain: implications for neurodevelopmental and neurodegenerative disorders

Authors: P. Pigini[1,2], H. Lindmeier[1], M. C. Silva[1,2], D. Gao[3,4], E. Morini[1,2]; [1]Ctr. for Genomic Med., Massachusetts Gen. Hosp. Res. Inst., Boston, MA, [2]Dept. of Neurology, Massachusetts Gen. Hosp. Res. Inst. and Harvard Med. Sch., Boston, MA, [3]Massachusetts Gen. Hosp., Boston, MA, [4]Program in Med. and Population Genetics and Stanley Ctr. for Psychiatric Res., Broad Inst. of Harvard and MIT, Cambridge, MA

Abstract:

Alternative splicing is a regulatory mechanism that controls transcript localization, translation, and stability, enabling the expression of multiple protein isoforms from a single gene. Splicing is finely controlled in human brain and plays a crucial role in neurogenesis, neuronal migration and structure, and synaptic function. A key role in this regulatory mechanism is played by poison exons (PEs). PEs are highly conserved exon cassettes whose inclusion creates premature termination codons (PTCs) and triggers nonsense-mediated decay (NMD) of the transcript, therefore reducing protein expression. Despite their critical role in gene expression regulation, PEs have not been systematically annotated due to the lack of comprehensive approaches that utilize transcriptome-wide data. This represents a critical barrier to the progress in this field. In this study, we systematically investigated the function of Pes in the human brain and assessed the impact of pathogenic mutations on their regulation. To delineate the full landscape of PEs in the human transcriptome, we included conserved (PhyloP score ≥ 20) alternatively spliced exons with the ability to introduce PTCs positioned at least 150 base pairs upstream of the 3' untranslated region and at least 50 base pairs downstream of the first exon. Our analysis identified 12,014 PEs in the human genome. For each poison exon, we calculated its percent-spliced-in (PSI) value by tissue type and development stage using both GTEx and

BrainSpan data. Overall, we identified 117 PEs uniquely found in the human brain, and 1,214 showing differential splicing compared to other tissues. To identify annotated pathogenic variants affecting PE splicing, we leveraged the ClinVar database and SpliceAI. We discovered 217 annotated pathogenic variants predicted to affect the splicing of 233 PEs in the brain. Notably, many of these variants are known to be associated with neurodevelopmental and neurodegenerative disorders, such as Alzheimer's disease (*CHAT*), epilepsy (*SCN1A* and *SCN2A*), lysosomal dysfunctions (*CLN5* and *CLN6*), Mowat-Wilson syndrome (*ZEB2*), or Batten disease (*CLN3*). We validated 10 of the most impactful candidates using minigene splicing assays and genome editing technologies. Functional analyses of cytotoxicity, differentiation markers, and neurite outgrowth enabled us to characterize the role of these PEs and associated variants in neural cell models. Overall, our work highlights the critical role of PEs in the human brain, and the functional characterization of pathogenic variants affecting their splicing paves the way for novel therapeutic strategies for incurable neurological disorders.

Session 84: Strategies to Interpret Germline Variants in Cancer Predisposition Genes

Location: Four Seasons Ballroom 4

Session Time: Friday, November 8, 2024, 1:15 pm - 2:15 pm

Applying scalable machine-learning approaches to generate evidence to impact variants of uncertain significance in Lynch syndrome genes

Authors: C. Tan[1], Y. Kobayashi[1], J. Reuter[1], T. Manders[2], B. Johnson[3], K. Nykamp[4]; [1]Invitae, San Francisco, CA, [2]Invitae, San Franscisco, CA, [3]Invitae, Fort Lauderdale, FL, [4]Invitae, Berkeley, CA

Abstract:

Intro: Lynch syndrome (LS), the most common cause of hereditary colorectal cancer, is due to pathogenic variants in mismatch repair (MMR) genes and nearly half of identified variants are classified as variants of uncertain significance (VUS) in ClinVar. This highlights the need for new scalable approaches to resolve VUS. We report on the development of machine-learning tools to maximize clinical data and the utilization of data from high-throughput functional assays to empower scientists with new genetic insights to inform variant classification in MMR genes.
Methods: A natural language processing classifier was leveraged to learn features of the provided indication for testing and family history predictive of a molecular diagnosis (moldx) of LS. Combining this info for each patient, a patient score was assigned by comparing the profile of patients with a positive moldx from those with a negative one. Next, a Bayesian inference model was fit to the data and used to sample posterior predictions for the likelihood that variants are pathogenic (variant score). In addition, multiplex assays of variant effect (MAVE) were generated for *MLH1*, *MSH2*, *MSH6* and *PMS2*. Using single cell RNA-sequencing to measure gene expression, machine-learning approaches looked for patterns to distinguish between known benign (B) and pathogenic (P) variants (pathogenicity score). Models with high performance based on AUC (Area Under the Receiver Operator Curve) threshold ≥ 0.8 were integrated into a semi-quantitative variant classification framework based on positive/negative predictive values (PPV/NPV). Classifications were reviewed by clinical genomics experts.
Results: The clinical variant modeling (CVM) strategy was applied to an internal database composed of over 1.5 million patients referred for genetic testing of at least one MMR gene and nearly 10,000 classified variants (variant score AUC=0.98). Across these genes, 273

VUS received high confidence predictions (variant score >0.99 PPV, >95% NPV). High-performance MAVE models were generated: *MLH1* (252 variants, AUC=0.95), *MSH2* (257 variants, AUC=1), *MSH6* (269 variants, AUC=0.99) and *PMS2* (221 variants, AUC=0.95). Across these genes, 104 VUS received high confidence predictions (pathogenicity score >0.95 PPV/NPV). Collectively, LS CVM and MMR MAVE models impacted 331 VUS>B reclassifications and 13 VUS>P reclassifications, potentially impacting 18,122 patient reports and leading to a VUS reduction rate of >24%.
Conclusion: Our approach demonstrates how using previously underutilized clinical data and proactively generating new functional data has the potential to reduce VUS at scale for patients tested for LS.

When clinical meets molecular: why, when and how do *CTNNA1* germline variants cause hereditary diffuse gastric cancer development

Authors: S. Lobo[1,2], A. Dias[1], M. Ferreira[1], J. Herrera-Mullar[3], A. Pedro[1], I. Gullo[1], M. Svrcek[4], R. Hüneburg[5], L. Moreira[6], S. Tinschert[7], L. Boussemart[8], J. L. Davis[9], B. Zäncker[10], L. P. Van Hest[11], K. Schrader[12], M. Baptista[13], J. M. Van Dieren[14], J. Balmaña[15], V. Strong[16], C. Lazaro[17], B. W. Katona[18], C. Colas[19], F. Coulet[20], R. Karam[21], P. S. Pereira[1], P. Benusiglio[22], C. Oliveira[1], CTNNA1 worldwide working group; [1]i3S - Inst. de Investigação e Inovação em Saúde, Univ. of Porto, Porto, Portugal, [2]ICBAS - Sch. of Med. and BioMed. Sci., Univ. of Porto, Porto, Portugal, [3]AmbryGenetic, Aliso Viejo, CA, [4]Sorbonne Université, Equipe Instabilité Des Microsatellites Et Cancer, Ctr. de Recherche Saint Antoine, Paris, France, [5]Natl. Ctr. for Hereditary Tumor Syndromes, Univ. Hosp. Bonn, Bonn, Germany, [6]Hosp. Clínic Barcelona, Dept. of Gastroenterology, Barcelona, Spain, [7]Med. Univ. Innsbruck, Div. of Human Genetics, Innsbruck, Austria, [8]Nantes Université, Univ Angers, CHU Nantes, INSERM, Immunology and New Concepts in ImmunoTherapy, Nantes, France, [9]Natl. Cancer Inst., NIH, Bethesda, MD, [10]Inst. für Klinische Genetik, Univ.sklinikum CarlGustav Carus Dresden, Dresden, Germany, [11]Amsterdam UMC, Vrije Univ.it Amsterdam, Amsterdam, Netherlands, [12]BC Cancer / Univ British Columbia, Vancouver, BC, Canada, [13]Centro Hosp.ar Universitário São João, Porto, Portugal, [14]Netherlands Cancer Inst., Dept. of Gastrointestinal Oncology, Amsterdam, Netherlands, [15]Hereditary Cancer Group, Med. Oncology Dept. Hosp. Vall d'Hebron, Barcelona, Spain, [16]Dept. of Surgery, Surgical Innovations and Outcomes, Mem. Sloan Kettering Cancer Ctr., York Ave, NY, [17]Hereditary Cancer Program, Catalan Inst. of Oncology, Bellvitge Inst. for BioMed. Res., Barcelona, Spain, [18]Univ. of Pennsylvania Perelman Sch. of Med., Div. of Gastroenterology and Hepatology, Philadelphia, PA, [19]Inst. Curie, Univ. Paris Sci. Lettres, Dept. of Genetics, Paris, France, [20]Unitéfonctionnelle d'Onco-angiogénétique et génomique des tumeurs solides,

Département de Génétique médicale, Hôpital Pitié-Salpêtrière, Paris, France, [21]AmbryGenetics, Aliso Viejo, CA, [22]Unité fonctionnelle d'Oncogénétique clinique, Département de Génétique, Groupe Hosp.ier Pitié-Salpêtrière, Paris, France

Abstract:

Introduction: Rare *CTNNA1*/αE-catenin germline truncating variants were found in Hereditary diffuse gastric cancer (HDGC) patients, however, full disease spectrum and variant-type causality are understudied. We aim to explore genotype-phenotype associations in *CTNNA1* variant carriers and molecular pathways causing *CTNNA1*-driven diffuse gastric cancer (DGC). **Methods:** Using a clinical database of 1388 individuals (1577 phenotypes) from 364 *CTNNA1* variant carrier families (Testing cohort: 71 European; Validation cohort: 290/293 American), we analyzed genotype-phenotype associations with multivariable logistic regression. Variants functional impact was assessed with in vitro/in vivo models. Transcriptomic profile of 11 DGC was analyzed. **Results:** From 71 European carrier families (61% ascertained for HDGC), 26 carried truncating variants from which 21 (81%) met HDGC criteria. DGC, occurring on average at 47.3±13.7, was significantly more likely to occur in truncating families than in non-truncating (OR=8.33; 95%CI [3.125-25]; p&lt0.001). While not statistically significant, lobular breast cancer (LBC) followed the same trend in truncating carriers (OR=4.76; 95%CI [0.98-50]; p=0.053). From a validation cohort of 293 families (24% ascertained for HDGC) enriched in truncating variants (271/293), 32 (12%) had HDGC. Here, LBC occurring on average at 55.2±12.3 was more frequent than DGC (40.6±17.0). We created CRISPR/Cas9 edited gastric cancer cells bearing a *CTNNA1* truncating variant with complete *CTNNA1*/αE-catenin loss. Nonsense Mediated mRNA Decay (NMD) blockade increased *CTNNA1* mRNA expression by 13-fold, recovering to wild-type (WT) levels. We created a *Drosophila* α-cat knockout (KO), in which organ development/lethality was rescued with overexpression of human WT/missense αE-catenin, but not with truncated αE-catenin. Paired normal/tumor transcriptomic analysis of DGC from carriers revealed 67 upregulated genes in tumors, including HIF1α and PIK3R3, two cancer therapy targets and drug repurposing candidates in *CTNNA1*-driven DGC. **Conclusion:** We provide a rational to test *CTNNA1* specifically in families meeting HDGC criteria, showing a DGC association with truncating, but not missense variants. Consolidation of LBC association to truncating variants requires larger series. We highlight NMD as a prime mechanism for *CTNNA1* truncated transcripts degradation and created an in vivo model to assess variants functional impact. *CTNNA1* DGC overexpress molecules worth exploring as therapy targets. **Acknowledgements:** FCT PhD fellowship (2020.05773.BD); ERN-GENTURIS (Project ID: No.739547); LEGOH (PTDC/BTM-TEC/6706/2020).

Comparing scalable and automated vs. ACMG/AMP variant interpretation strategies for BRCA1 and BRCA2 in a large clinicogenomic cohort from six US-based health systems

Authors: K. Schiabor Barrett, E. Cirulli, B. Khuder, M. Ferber, N. Washington, A. Bolze; Helix, San Diego, CA

Abstract:

Introduction: Population genetic screening for disease risk is becoming common practice in healthcare. The variant interpretation method used for these screening efforts remains based on the ACMG/AMP rubric. The rubric was developed for case-by-case diagnostic interpretations, incorporates patient-specific information, and is not scalable to large cohorts. We tested if an automated approach to variant interpretation, based solely on well-understood genetic properties of disease-causing variants and clinically established variant calls, could match the accuracy of the results from the more time-consuming ACMG/AMP interpretation process. We re-interpreted small variants in *BRCA1* and *BRCA2* in a clinicogenomic cohort of 80,000 individuals from six health systems who had undergone previous ACMG-style variant interpretation, using an automated approach. We then contextualized the interpretations from each method against breast and ovarian cancer diagnoses in women (individuals of female sex).
Methods: Small variants (SNVs and indels <50bps) were called and vetted for quality via clinical-grade exome pipeline for six health system cohorts. ACMG guided interpretations were performed between 2020-2024 in a clinical testing laboratory. Automated interpretations were made for the entire cohort at once using the following parameters to identify risk variants: 1) non-benign pLOF variants with MAF<0.1% in gnomAD as well as any of the six cohorts or 2) well-accepted pathogenic and likely pathogenic variants from clinical databases (e.g., from the HBOC Expert Panel). Performance for each method was analyzed using Cox proportional hazards regression with age at first diagnosis of either breast or ovarian cancer.
Results: 405 individuals out of 78,926 (0.51%) carried a pathogenic or likely pathogenic variant from the ACMG-style interpretation pipeline. 395/405 (97.5%) of these carriers were also identified via the automated pipeline. Interestingly, 42 additional carriers were found only by the automated approach. In a time to event analysis of women, carriers found only by the automated method showed a similar hazard ratio to those found by both methods, when compared against non-carrier controls (carriers found by both methods (n=272) HR=5.4, $p=2.5e^{-28}$, carriers found only by automated method (n=33) HR=5.9, $p=5.28e^{-6}$).
Conclusion: For *BRCA1* and *BRCA2*, automated interpretation matches and may even

exceed the sensitivity and specificity of ACMG-style sample by sample interpretation strategies, as all variant carriers called by the automated pipeline have breast and ovarian cancer rates matching that of pathogenic carriers called by both methods.

Expanding the reach of paired DNA and RNA sequencing: Results from 450,000 consecutive individuals from a hereditary cancer cohort

Authors: C. Horton, L. Hoang, J. Grzybowski, H. Zimmermann, M. Richardson, J. Ramirez Castano, S. Belhadj, R. Karam; Ambry Genetics, Aliso Viejo, CA

Abstract:

Paired DNA and RNA sequencing has shown promise in improving detection and interpretation of DNA variants identified across a variety of clinical indications. Expansion of genes eligible for RNA sequencing, adjustments to assay design, increases in the number of cases and controls tested, and refinement of guidelines for application of RNA evidence can all contribute to improved utility of RNA sequencing. Here we report outcomes of paired DNA and RNA sequencing in light of these improvements among 450,000 individuals undergoing hereditary cancer multigene panel testing (MGPT). Results of MGPT including concurrent DNA and RNA sequencing performed between April 2019- December 2023 were retrospectively reviewed. The positive rate calculated includes pathogenic and likely pathogenic variants (PVs) and excludes monoallelic variants in genes associated with recessive conditions and moderate risk PVs (*APC* p.I1307K and *CHEK2* p.I157T). Variant classifications were compared before and after application of RNA evidence to calculate impact on positive yield. Medically significant upgrades were defined as those resulting from a newly detected intronic PV and reclassifications from uncertain significance to PV. Genes with no RNA-related reclassifications were further evaluated for rationale.

A total of 455,378 cases were included in this study. The overall positive rate was 9.9%, consisting of 46,027 PVs identified in 45,202 individuals. Medically significant upgrades based on RNA evidence were made to 1,838 variants, resulting in a 4.2% relative increase in positive yield. Among genes with at least 100 PVs reported, the greatest impact on positive yield due to RNA was observed in *LZTR1* (130 of 809 PVs; 19.1% relative increase in yield), *RAD51C* (80 of 691; 13.1%), *CDH1* (21 of 258; 8.9%), *ATM* (388 of 4,865; 8.7%), *APC* (57 of 780; 7.9%), and *PTEN* (19 of 277; 7.4%). Reclassifications based on RNA evidence were made in 53 of the 85 genes evaluated. Reasons for lack of RNA reclassifications included: loss of function via haploinsufficiency was not the established mechanism for disease (27 of 32 genes; 84.4%), extreme rarity of disorder (under 0.01% of

tested individuals) (4 of 32; 12.5%) and technical limitations (1 of 32; 3.1%).
In this study, on average, 1 in 25 pathogenic variants were dependent on RNA evidence. Discovering genes in which RNA sequencing is especially impactful and identifying parameters when RNA evidence is not applicable can provide useful insights to aid providers in risk assessment and test selection. The impact of RNA sequencing may continue to grow as its adoption becomes more widespread and its applications are validated.

Session 87: Framing Heritability for Complex Traits

Location: Room 501

Session Time: Saturday, November 9, 2024, 8:00 am - 9:00 am

Fine-mapped insertions and deletions disproportionately impact 78 diseases and complex traits

Authors: J. Maravall López[1], A. L. Price[2]; [1]Harvard Univ., Cambridge, MA, [2]Harvard Sch Pub Hlth, Boston, MA

Abstract:

Despite increasing evidence that indel and structural variants strongly impact human disease (Mukamel et al. 2021 *Science*), nearly all fine-mapping studies to date have focused on SNP variants. Two recent studies have identified and genotyped millions of insertions and deletions in 1000 Genomes samples (Ebler et al. 2022 *Nat Genet*, Koenig et al. 2024 *Genome Res*), but their impact on human disease is currently unknown.
Here, we assessed the impact of insertions and deletions (of any length) across 78 diseases and complex traits with publicly available summary statistics (average N=302K). We first applied RAISS to impute summary statistics for the Ebler et al. variants from GWAS summary statistics and 1000 Genomes reference LD. We then performed single causal variant fine-mapping, the recommended approach for summary data without in-sample LD.
We identified a total of 453 unique indel variants with posterior inclusion probability (PIP)>0.95 (678 unique indel variants with PIP>0.5). Aggregating across traits, insertions comprised 7.5% of all well-imputed variants but 12.7% of all PIP>0.95 variants, an enrichment of 1.68x (s.e. 0.11x via genomic block-jackknife); deletions comprised 8.3% of all well-imputed variants but 13% of all PIP>0.95 variants, an enrichment of 1.55x (s.e. 0.10x); analyses of Koenig et al. data produced similar findings. Noteworthy examples include an intronic 62bp deletion for height at *COL6A2*, a gene encoding a form of collagen found in most connective tissues; an intronic 24bp deletion for total cholesterol at *ALKBH5*, a gene implicated in lipid metabolism ; and an intronic 1bp deletion for type 2 diabetes at *ETV1*, a transcription factor that regulates insulin secretion. Previous studies have implicated each of these loci as GWAS loci, but have only implicated SNP variation.
We corroborated these findings via heritability enrichment analyses using S-LDSC with the baseline-LD model (Finucane et al. 2015 *Nat Genet*, Gazal et al. 2017 *Nat Genet*), extending the baseline-LD model to include annotations for insertion and deletion variants. Meta-analyzing across traits, we observed an enrichment of 2.54x (s.e. 0.14x) for insertions

and 1.45x (s.e. 0.14x) for deletions ; analyses of Koenig et al. data produced similar findings. This implies that >40% of disease heritability arises from insertion and deletion variants, motivating intense efforts to capture and investigate disease effects of these largely unmodeled variants.

Heritability and effect-size distribution of rare and de novo protein-coding variation

Authors: W. Lu[1], A. Nadig[1], B. Neale[2], E. Robinson[3], K. Karczewski[2], L. O'Connor[4]; [1]Broad Inst., Cambridge, MA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]Harvard Sch. of Publ. Hlth., Boston, MA, [4]Harvard Med. Sch., Boston, MA

Abstract:

Rare protein-coding genetic variants contribute to common diseases and complex traits, and rare variant association studies (RVAS) have identified thousands of genes with significant effects. However, rare variant genetic architecture remains mostly uncharacterized outside of top associations, especially in non-European ancestry groups. New methods are needed in order to quantify the total contribution of rare genetic variation, compare genetic architecture across genetic ancestry groups, and understand the genetic basis of human diseases.
We developed burdenEM, a method to estimate the distribution of gene-wise burden effect sizes. Our method is highly efficient, operating on summary association statistics derived either from a traditional RVAS or from a trio study design, allowing for the characterization of disease-associated inherited and de novo variation, respectively. BurdenEM fits a flexible mixture model for the burden effect size distribution using expectation-maximization and estimates features of that distribution, such as heritability attributable to rare variants. We validated the performance of our method using summary statistics of randomly simulated phenotypes as well as simulated burden summary statistics.
We applied BurdenEM to RVAS data from Genebass.org (N=394,841), and to trio data from the Autism Sequencing Consortium (ASC; N=15,000). For autism, the contribution of de novo LoF variants was 3.1% on an observed scale (0.5% for missense damaging variants and nearly zero for missense benign and synonymous variants). Further, we projected the number of genes that will be discovered with larger sample sizes. As in GWAS, the number of discoveries increases roughly linearly with sample size; however, unlike in GWAS, newly discovered genes often have large effect sizes (similar to those discovered already). Forecasting results for the next tranche of ASC trio sequencing data (N = 30k), we predict that increasing sample size will continue to yield newly associated genes with high

penetrance (i.e. > 50%), suggesting that current gene discovery is primarily limited by mutation rate rather than sample size. Across 4 well-powered Genebass quantitative traits, we stratified contributions to heritability by constraint and estimated the mean burden heritability to be 1.5%, 0.8%, and 0.3% for predicted loss-of-function (pLoF), missense, and synonymous variants, respectively. We also observed consistent trends across multiple ancestry groups in 12 well-powered quantitative traits from the new All by All RVAS data from All of Us (N = 245,394).

Uncovering the contribution of rare variants to the heritability of complex traits: Insights from the UK Biobank whole genome sequencing data

Authors: H. Jung[1], H. Jung[2], J-O. Kang[1], J. Lim[3], B. Oh[4]; [1]Kyung Hee Univ., Seoul, Korea, Republic of, [2]Kyung Hee Univ., Seoul, Korea, Republic of, [3]Kyung Hee Univ, Seoul, Korea, Republic of, [4]Sch Med, Kyung Hee Univ, Seoul, Korea, Republic of

Abstract:

Finding the missing heritability of complex traits remains a major challenge in genetic studies today. The pursuit of missing heritability is driven by the hypothesis that complex traits are influenced not only by common genetic variants but also by rare variants, structural genetic variations, gene-gene interactions, and gene-environment interactions not easily captured by conventional GWAS. Wainschtein et al. (2022) investigated the impact of rare variants (e.g., minor allele frequency below 0.01%) using whole genome sequencing (WGS) data on the heritability of complex traits, focusing on height and body mass index (BMI). Their results showed that rare variants contribute significantly to heritability. However, their study, based on 25,465 samples, showed a large standard error in rare variant heritability estimates. Recently, the UK Biobank released WGS data on 500,000 participants. In this study, we used WGS data from the UK Biobank to estimate rare variant heritability with smaller standard errors. To achieve this, we first selected 110,000 unrelated European samples from the UK Biobank. Next, to select the rare variants, we calculated linkage disequilibrium (LD) values for all single nucleotide variants (SNVs) within 1,000kb from their WGS data, and selected independent 10,707,882 rare variants (minor allele frequency between 0.0001 and 0.01, window size of 1000kb and R^2 threshold of 0.05). Using these 10,707,882 rare variants and 110,000 unrelated European samples, we calculated the genetic relationship matrix (GRM). Then, we estimated rare variant heritability for 55 quantitative traits. Our results showed that the rare variant heritability for 53 quantitative traits was statistically significant with estimates ranging from 6.63% (urea) to 37.17% (height). Specifically, the rare variant heritability of height and BMI were 37.17%

(standard error: 0.02) and 14.85% (standard error: 0.02), respectively. When combining common and rare variant heritability for the 53 traits, the total heritability ranged from 82.25% (height) to 16.76% (peak expiratory flow). Notably, twin studies have reported heritability estimates of 80-90% for height and 50-90% for BMI (Visscher et al., 2012; Elks et al., 2012). Our results showed that for height, we accounted for 82.25% of the 90% twin heritability, while for BMI, we accounted for 39.37% of the 90% twin heritability. This indicates that while most of the missing heritability for height can be explained by rare variants, approximately 30% of the heritability for BMI remains missing. In conclusion, our study demonstrates that rare variants contribute significantly to the heritability of various traits.

Partitioning genetic and non-genetic contributions to epigenetic-defined endotypes of allergic phenotypes in children

Authors: E. Thompson[1], X. Zhong[1], P. Carbonetto[1], A. Morin[1], C. M. Visness[2], G. T. O'Connor[3], L. B. Bacharier[4], M. Kattan[5], R. A. Wood[6], R. Miller[7], C. C. Johnson[8], E. M. Zoratti[8], G. K. Khurana Hershey[9], D. R. Gold[10], C. Seroogy[11], M. C. Altman[12], T. Hartert[13], M. Stephens[1], D. J. Jackson[11], J. E. Gern[11], C. G. McKennan[14], C. Ober[1]; [1]Univ. of Chicago, Chicago, IL, [2]Rho Inc., Federal Res. Operations, Durham, NC, [3]Boston Univ. Sch. of Med., Boston, MA, [4]Monroe Carell Jr Children's Hosp. at Vanderbilt Univ. Med. Ctr., Nashville, TN, [5]Columbia Univ. Med. Ctr., New York, NY, [6]Johns Hopkins Univ., Baltimore, MD, [7]Icahn Sch. of Med. at Mount Sinai, New York, NY, [8]Henry Ford Hlth., Detroit, MI, [9]Univ. of Cincinnati Coll. of Med., Cincinnati, OH, [10]Harvard T.H. Chan Sch. of Publ. Hlth.and Brigham and Women's Hosp., Harvard Med. Sch., Boston, MA, [11]Univ. of Wisconsin Sch. of Med. and Publ. Hlth., Madison, WI, [12]Benaroya Res. Inst. Systems, Seattle, WA, [13]Vanderbilt Univ. Sch. of Med., Nashville, TN, [14]Univ. of Pittsburgh, Pittsburgh, PA

Abstract:

The epigenome harbors vast amounts of variation that have been largely untapped in studies of human health and disease. DNA methylation (DNAm), the most common and most studied epigenetic mark, reflects current and past exposures as well as effects of nearby genetic variants. Thus, variation in DNAm provides a tractable framework for understanding the architecture of complex traits with contributions from genes and environment, such as asthma and allergic diseases. We used the Asthma&Allergy DNAm array (37,256 CpGs) in airway epithelial cells from 284 children (age 11 years) in the Urban Environment and Childhood Asthma (URECA) birth cohort study and an empirical Bayes approach to perform matrix factorization to assign CpGs into 16 distinct patterns, or

signatures (lfsr<0.05). Three DNAm-defined signatures (S5, 6,654 CpGs; S8, 4,067 CpGs; S16, 2,602 CpGs) were associated with at least one of 11 phenotypes measured at age 10 years after adjusting for airway cell type proportions and other nuisance covariates: allergic asthma ($p=9.5x10^{-5}$), allergic rhinitis ($p<0.0041$), allergic sensitization ($p<7.1x10^{-4}$), total IgE ($p<6.1x10^{-5}$), exhaled NO ($p=2.4x10^{-4}$), and blood eosinophils ($p<1.1x10^{-7}$). Next, using RNA-seq from the same cells, we identified genes correlated with each signature (FDR<0.05: S5, 4,305 genes; S8, 4,952 genes; S16, 3,720 genes), revealing independent sets of networks reflecting inhibition of microbial responses (S5), inhibition of epithelial barrier integrity (S8), and activation of T2 immune pathways (S16). The associations between allergic phenotypes and the three signatures were replicated in two independent cohorts. Finally, using genotype data from all three cohorts (N=1,389 individuals), we estimated heritability (h^2) using fine-mapped variants (PIP>0.01) from an allergic rhinitis GWAS, and variants associated with DNAm for the CpGs in the three signatures (meSNPs) or with expression of genes correlated with the three signatures (eSNPs). The estimated h^2 of each was highly significant even using this limited set of variants: F5 h^2 = 17% ($p=2.7x10^{-3}$), F8 h^2 = 30% ($p=9.3x10^{-7}$), and F16 h^2 = 16% ($p=8.9x10^{-7}$). Using an unbiased systems-level, multi-omic approach, we identified three methylation signatures that defined distinct endotypes of asthma and allergic disease that reflected both genetic and non-genetic effects. The significant genetic effects suggest that the three DNAm-defined risk signatures are established in early life.

Session 88: Keeping It REnAL! Genetic Studies of Kidney Disease

Location: Room 405

Session Time: Saturday, November 9, 2024, 8:00 am - 9:00 am

GWAS of multiple renal function biomarkers and kidney multi-omics prioritizes new chronic kidney disease genes

Authors: A. Emmett[1], X. Jiang[1], X. Xu[1], J. Eales[1], S. Hames-Fathi[1], D. Scannali[1], E. Miller-Kasprzak[2], Y. Sun[1], A. Lay[1], P. Bogdanski[2], E. Zukowska- Szczechowska[3], J. Zywiec[3], W. Wystrychowski[3], N. Samani[4], T. Guzik[5], B. Keavney[1], A. Morris[1], Human Kidney Tissue Resource Study Group, F. Charchar[6], M. Tomaszewski[1]; [1]Univ. of Manchester, Manchester, United Kingdom, [2]Poznan Univ. of Med. Sci., Poznan, Poland, [3]Med. Univ. of Silesia, Katowice, Poland, [4]Univ. of Leicester, Leicester, United Kingdom, [5]Univ. of Edinburgh, Edinburgh, United Kingdom, [6]Federation Univ., Ballarat, Australia

Abstract:

Chronic Kidney Disease (CKD) affects 10% of the world's population and is the third fastest growing cause of death globally. Genome wide association studies (GWAS) have identified CKD risk loci by mapping genetic associations with kidney filtration measures, estimated from metabolites. However, there is an unmet need to: 1) separate true kidney function loci from metabolic loci using multiple renal biomarkers, and 2) translate GWAS findings into molecular mechanisms and therapeutic targets. Here we triangulate multiple indices of renal function with human kidney multi-omics, to understand how genetic variants confer CKD risk through DNA methylation, gene expression and protein abundance. We performed GWAS of 4 markers of CKD-defining traits in up to 327,689 UK Biobank participants, and meta-analysed the results with associations from CKD-Gen in up to 567,460 individuals. GWAS identified 421 independent loci associated with estimated glomerular filtration rate calculated from creatinine (eGFRcr), 235 loci for eGFR calculated from cystatin C (eGFRcys), 198 loci for blood urea nitrogen (BUN) and 51 loci for urinary albumin creatinine ratio (UACR). 237 loci were shared between multiple biomarkers. To elucidate molecular underpinnings of CKD loci, we mapped methylation, expression and protein quantitative trait loci in human kidney tissue (mQTL, *n* = 366; eQTL, *n* = 645, pQTL, *n* = 83). Leveraging molecular QTLs for 89,214 CpGs, 8,899 genes and 243 proteins, we performed colocalization analyses to pinpoint shared causal associations with CKD loci. Of 508 unique loci associated with any CKD biomarker, 46% colocalized with DNA methylation (842 CpGs), and 34 colocalized with expression (230 genes). Four CKD loci

colocalized with protein abundance of DPEP1, GSTA1, GCDH and MST1. Enrichment analyses revealed that putative CKD genes (with colocalizing methylation, expression, or protein abundance) are preferentially expressed in the thick ascending limb and proximal tubules of the kidney, and function in renal development and glucose transport. We uncovered molecular colocalization at 49 multi-marker CKD loci, mapping onto 94 genes, including 30 new candidates. These include novel genes of potential therapeutic relevance as targets of existing medications, such as *CA12* targeted by acetazolamide and *SLC9A3* targeted by tenapanor. The remaining multi-marker CKD genes are replicated by prior CKD gene prioritization studies, including *SHROOM3*, *UMOD* and *MUC1*. Altogether, multi-marker CKD GWAS and human kidney multi-omics identified genes and processes underpinning CKD risk, with implications for basic science and translational research.

Characterization of a novel *ASAH2* variant associated with diabetes and kidney failure in Tongan and Samoan patients

Authors: C. Dumaguit[1], R. J. Nicholson[1], C. Simeone[2], J. Taloa[1], K. Lao[1], C. Littlefield[1], J. A. Maschek[1], S. A. Summers[1], M. G. Pezzolesi[1], W. L. Holland[1]; [1]Univ. of Utah, Salt Lake City, UT, [2]Beth Israel Deaconess Med. Ctr., Boston, MA

Abstract:

Sphingolipids, such as ceramides, are bioactive lipids that accumulate with metabolic disease and modulate insulin resistance, apoptosis, and fibrosis. We investigate novel genetic divers of high ceramides and their resulting contribution to heritable metabolic disease pathologies, such as diabetes and kidney disease. Identification and characterization of novel risk alleles linked to ceramide accumulation will aid in the screening and treatment of high-risk patient populations. Native Hawaiians and Pacific Islanders (NHPI) face the highest rates of obesity, diabetes and end-stage kidney disease (ESKD) of any racial or ethnic group in the United States. We have identified a novel coding mutation in neutral ceramidase (*ASAH2*), which is exclusive to patients of Tongan and/or Samoan ancestry and is associated with an elevated risk of diabetes and kidney failure. Three probands were identified, with elevated plasma ceramides and an identical novel coding variant at position chr10:50236064, which was not observed in 125,643 genomes or exomes from the Genome Aggregation Database (gnomAD). This point mutation (c.511C→G) results in a V171L substitution of a highly conserved amino acid within the catalytic domain of ASAH2. Heterozygous carriers of the *ASAH2* mutation have higher circulating ceramides compared to NHPI non-carriers with diabetes and kidney failure.

Mechanistically, the single point mutation occurs at the first base pair of exon 5, where it fails to serve as a splice acceptor. Indeed, sequencing of mRNA from CRISPR knock-in cells revealed that the mutation elicits complete excision of exon 5 (ASAH2^{ex5del}), which encodes crucial amino acids within the ASAH2 catalytic domain; functional studies have confirmed ablation of ceramidase activity in ASAH2^{ex5del} cells. To further characterize metabolic disease risk with ASAH2 loss-of-function, we challenged *Asah2* knockout (*Asah2*$^{-/-}$) mice with a high-fat diet for 16 weeks. Both male and female *Asah2*$^{-/-}$ mice displayed worsened insulin sensitivity compared to *Asah2*$^{+/+}$ littermate controls. Additionally, we probed these mice for altered susceptibility to diabetic kidney disease. Male and female mice were challenged with high-fat diet prior to the additional insult of low-dose streptozotocin. 8-weeks after diabetes induction, female *Asah2*-/- mice had elevated proteinuria, kidney hypertrophy, and increased expression of markers related to fibrosis and kidney damage. Together, these data reveal a novel risk allele for diabetes and kidney disease that is present in 1 in 10 individuals of Tongan or Samoan descent, which could be treated with precision therapies designed to lower ceramide accumulation.

KidneyGenAfrica: Putative novel genetic loci and improved polygenic prediction for kidney function derived from aggregating 10 continental African genome-wide association studies ⭐

Authors: S. Fatumo[1], A. Kamiza[2], J. Fabian[2], J-T. Brandenburg[3], members of KidneyGenAfrica; [1]MRC/UVRI & LSHTM Uganda Res. Unit, Entebbe, Uganda, [2]Univ. of the Witwatersrand, Johannesburg, South Africa, [3]SBIMB, Johannesburg, South Africa

Abstract:

Background: Most genome-wide association studies (GWAS) on estimated glomerular filtration rate (eGFR) have been performed in Europeans. To increase discovery of additional loci in individuals of African ancestry, we assembled 10 GWAS of eGFR across diverse geographical regions in Africa with 28K individuals as part of the newly established KidneyGenAfrica. Additional GWAS of eGFR in 80K African-ancestry individuals in the diaspora were aggregated from the Million Veteran Program (MVP), UK BioBank (UKBB), and chronic kidney disease genetic consortium meta-analysis (CKDGEN). **Methods:** We performed a three-stage meta-analysis: (1) regional meta-analyses within East, West, and South Africa; (2) Continental African meta-analysis of East, West, and South Africa; and (3) Pan-African meta-analysis using data from continental Africa, MVP, UKBB, and CKDGEN. Meta-analyses were performed using GWAMA, METAL and Metasoft, respectively. We also performed fine-mapping, colocalization, and PheWAS. Polygenic risk scores (PRSs) were

developed from GWAS summary statistics using data from continental Africans, Africans of Diaspora, and Pan-African ancestry, and tested and validated in a Malawi cohort. **Results:** We identified novel loci (rs74383679, AVPR1B), (rs73788952, OPRM1), (chr6:112537967), (rs7763270, LAMA4), (rs6670659, HSD3B1), (rs1706775, GATM), and (rs4243062, TBC1D2B) in regional meta-analysis. Pan-African meta-analysis detected an additional 20 independent loci, including six novel loci. These loci were mapped to genes involved in kidney function. Our fine mapping reduced the credible set size and identified eight loci with a posterior probability of causality > 0.99. Colocalization recapitulated known eGFR-related genes, and PheWAS identified 26 loci, in addition to loci associated with cardiometabolic and immunological traits. The pan-African ancestry-derived PRSs performed and predicted better than continental and African diaspora PRSs. **Conclusion:** We identified novel region-specific loci in continental Africa. By incorporating data from continental Africa and the diaspora into PRSs, we enhanced their accuracy, underscoring the critical role of genetic diversity in ensuring the equitable application of PRS across different populations.

SLC6A19 loss of function is associated with improved kidney function and metabolic reprogramming of kidney cells

Authors: S. Mozaffari, A. Joslin, L. Pang, P. Wong, Y. Xi, L. Sanman, J. Ullman, M. Hoek, R. Graham, K. Estrada; Maze Therapeutics, South San Francisco, CA

Abstract:

Chronic kidney disease (CKD) causes progressive loss of kidney function ultimately resulting in kidney failure. The plasma biomarkers creatinine or cystatin C are the basis for calculating estimated Glomerular Filtration Rate (eGFR), a clinically relevant quantitative endophenotype of kidney function. A burden test incorporating predicted loss of function (pLOF) and rare missense variants (MAF <1%) previously identified *SLC6A19* as among the strongest effects in the genome for improved kidney function (UK biobank: plasma creatinine, beta = -0.13, p = 4.58e-35). Utilizing an allelic series of variants that impact *SLC6A19* function or expression levels composed of aggregated pLOF and rare predicted damaging missense variants, a known hypomorphic missense variant (D173N, rs121434346), and an eQTL variant (rs11133665), we performed a meta-analysis for serum creatinine across multiple large-scale datasets (UK Biobank, deCODE, CKDGen: > 600,000 individuals). The hypomorphic variant D173N (rs121434346) is associated with decreased serum creatinine (beta: -0.16, p: 2.16e-27) in deCODE and UK Biobank; Aggregated UK Biobank pLOF variants (beta: -0.15, p: 1.02e-04) and missense variants (MAF<0.001) (beta:

-0.08, p-value: 3e-13) are associated with decreased serum creatinine. Individuals homozygous for D173N had higher eGFR (16.8 ml/min/1.73m^2 increase, p = 1 x 10^{-3}*).* Consistent results were obtained for cystatin C, suggesting loss of function of *SLC6A19* is associated with improved markers of kidney function. The gene *SLC6A19* encodes the protein B_oAT1, a sodium-dependent neutral amino acid transporter involved in the uptake of free amino acids from the diet and minimizing loss of amino acids via the urine. We expanded on the previous finding that loss of *Slc6a19* is protective against kidney injury in mice (PMC9484999). We subjected WT and *Slc6a19*-/- animals to renal injury induced by the proximal tubule toxin aristolochic acid (AAI) and performed bulk RNAseq in the kidneys of these mice to elucidate mechanisms of protection. During the injury and recovery phase, *Slc6a19*-/- animals were protected against proximal tubule damage and exhibited reduced expression of the renal injury markers *Havcr1*, *Lcn2*, and *Vcam1*. We observed a strong reversal of the overall kidney AAI injury transcriptomics signature (p-value: 9.55e-74) in *Slc6a19-/-* animals compared to WT, and possible mechanisms of protection via metabolic reprogramming away from injury induced glycolysis.

Session 89: Long-Read Sequencing Offering New Insights into Neurological Disease

Location: Room 505

Session Time: Saturday, November 9, 2024, 8:00 am - 9:00 am

Long-read sequence and assembly of autism reference genomes

Authors: Y. Sui[1], M. Wu[1], W. T. Harvey[1], M. Noyes[1], Y. Kwon[1], I. Wong[1], K. M. Munson[1], K. Hoekzema[1], G. Garcia[1], J. Knuth[1], J. Kordosky[1], A. P. Lewis[1], E. Eichler[1,2]; [1]Univ. of Washington, Seattle, WA, [2]Howard Hughes Med. Inst., Seattle, WA

Abstract:

Autism spectrum disorders (ASDs) are genetically and phenotypically heterogeneous and the majority of cases still remain genetically unresolved. To better understand potential pathogenic variation and epigenetic signatures contributing to ASD, we constructed phased and near-complete genome assemblies of 25 unsolved ASD families (100 individuals). We generated deep whole-genome sequence data from peripheral blood or cell lines from parents and their offspring using three orthogonal sequencing technologies: PacBio HiFi, ONT and Illumina. We focused on families with female-only probands and where no obvious pathogenic variant was identified by traditional short-read sequencing (SRS) and the polygenic risk score was not exceptional. We constructed highly contiguous phased genome assemblies from each individual (average contig N50=48.4Mbp, QV=54), which allowed us to discover novel variants, assess transmission properties and map recombination events. We developed both read- and assembly-based pipelines to facilitate comprehensive characterizations of *de novo* mutations (DNMs), structural variants (SVs) and DNA methylation profiles. Long-read sequencing (LRS) provides access to ~92% of the human genome and increases DNM discovery by ~32% and SV discovery by 2- to 3-fold when compared to SRS datasets. Filtering SVs from a dataset of 149,000 SVs obtained from HGSVC and HPRC controls, we identified 3820 SVs exclusive to ASD families, with the majority being classified as privately inherited (3800). Probands show a slight trend for more autosomal SVs relative to unaffected siblings (1842 vs. 1782, ns) and more rare X-chromosome SVs relative to their sex-matched unaffected sisters (26 vs. 19, ns). Out of 12 proband *de novo* SVs, 3 map within the intronic regions of potential candidate genes, including a 71 bp insertion within *CNTN3*, a 73 bp insertion overlapping a regulatory element of *CPT1C* and a 5370 bp deletion removing regulatory elements from *ARHGAP26*. An analysis of DNMs identified a stop-gain mutation in the ASD gene *SYNGAP1* that had been filtered in SRS datasets. Preliminary analysis of methylation in 4 sex-matched quads

suggests greater skewing between 2 haplotypes of the X chromosome in affected females than unaffected females at CpG islands. Our results highlight the value of constructing ASD reference genomes using LRS to identify pathogenic variants missed by SRS and discovering epigenetic differences potentially relevant to sex biases. This new resource will serve as a benchmark for systematic genetic and epigenetic characterization of not only ASD but also as a road map for other genetically complex disorders in the future.

Long-read sequencing to diagnose Autosomal recessive Parkinson's disease in diverse populations

Authors: K. Daida[1], H. Yoshino[2], M. Laksh[1], B. Baker[1], M. Ishiguro[2], R. Genner[3], K. Paquette[1], Y. Li[2], K. Nishioka[2], The French Parkinson's disease genetic study group, G. Cogan[4], C. Tesson[4], S. Lesage[5], S. Schaake[6], J. Trinh[7], K. Lohmann[8], C. Sue[9], K. Billingsley[1], M. Funayama[2], C. Klein[10], A. Brice[11], A. Singleton[12], C. Blauwendraat[1], N. Hattori[2]; [1]NIH, Bethesda, MD, [2]Juntendo Univ., Tokyo, Japan, [3]Johns Hopkins Univ. / NIH, Bethesda, MD, [4]Paris Brain Inst., Paris, France, [5]ICM HOPITAL DE LA PITIE-SALPETRIERE, Paris, France, [6]Univ. of Lubeck, Lubeck, Germany, [7]Univ. of Luebeck, Luebeck, Germany, [8]Univ. of Lübeck, Lübeck, Germany, [9]Univ. of New South Wales, Sydney, Australia, [10]Univ. of Lübeck and Univ. Hosp. Schleswig-Holstein, Lubeck, Germany, [11]Hosp Pitie-Salpetriere, Paris, France, [12]NIA, Bethesda, MD

Abstract:

PRKN and *PINK1* are the most common causative genes of autosomal recessive Parkinson's disease (ARPD). While the significance of biallelic variants in these genes as causes of ARPD is well established, the role of heterozygous variants remains conflicting. Notably, some Parkinson's disease (PD) patients exhibit a typical ARPD phenotype with early-onset PD while only one *PRKN* variant is identified. Using long-read sequencing, we identified a 7.4M bp inversion in *PRKN* in a family with early-onset PD that had remained genetically undiagnosed for over 20 years, which was confirmed by optical genome mapping. Motivated by this finding, we expanded long-read sequencing to include early-onset PD patients of Asian, European, and African ancestry who have only one known *PRKN* or *PINK1* pathogenic variant. (n=65) Preliminary analysis revealed that long-read sequencing identified a second variant, undetectable by conventional sequencing methods, in over 20% of PRKN heterozygous carriers.The newly identified variants included complex inversions such as duplication-normal-duplication/inversion and overlapping copy number variants within the same allelic exons. This study highlights the complex nature of *PRKN* variants and demonstrates the utility of long-read sequencing in accurately

diagnosing ARPD across diverse populations.These findings underscore the necessity of employing long-read sequencing to uncover hidden genetic variants, providing a more comprehensive understanding of the genetic basis of ARPD and improving diagnostic accuracy for patients with early-onset PD.

Mapping parent-of-origin methylation pattern during development by long-read 5-base HiFi sequencing reveals novel imprinting motifs and insight into human disease

Authors: E. Grundberg[1], B. Koseva[1], W. Cheung[1], A. Johnson[1], C. Marsch[2], I. Thiffault[1], T. Pastinen[1]; [1]Children's Mercy Kansas City, Kansas City, MO, [2]Univ. of Kansas Med. Ctr., Kansas City, KS

Abstract:

Parent-of-origin specific expression or imprinting, involves differential regulatory element activity and often measured through methylation of CpG (mCpG) dinucleotides. Imprinted genes in humans are important for fetal and placental development and a dozen clinical syndromes are linked to defective imprinting. Current estimates of the number of human imprinted genes vary widely but those that have been robustly validated remain fewer than 100. Here, we leverage our 5-base long-read HiFi genome sequencing (5mC-HiFi-GS) platform for single-molecular profiling of mCpG together with pedigree-based phasing from long contiguous reads. Specifically, we applied 5mC-HiFi-GS at high depth in 198 samples from 66 trios including early developmental (~6-8 weeks gestation) chorionic villi as well as on paternal germ cells and proband/parental peripheral blood tissue. Using genome-wide haplotype-resolved differential mCpG testing, we identified widespread parent-of-origin effect (POE) of mCpGs in placental cells (95% maternal) at novel loci (N~2000) with 86% replication rate in independent samples. Paternal POE was rare but integrating of 5mC-HiFi-GS from germ cells revealed sperm hypermethylation at 98% of loci showing signature of paternal imprinting. We further showed that only 10% of loci maintain POE across cell types and developmental stages restricting genomic imprinting mostly to trophoblasts. Single-nuclei multiome profiling of chorionic villi revealed 2-fold enrichment of allele-specific expression of genes mapping to our identified loci in fetal vs. maternal cells where unique co-measurements of maternal and fetal cells with genetic deconvolution allowed unbiased investigation of tissue impact of imprinting. Finally, we investigated disease relevance using over 10,000 rare disease cases by restricting to POE loci mapping near genes (N=15) highly intolerant to loss-of-function (LOF) (pLI = 1). Eight of the 15 genes are included in OMIM and four of these are candidates as novel imprinting disorder, since

only *de novo* or parentally biased transmission have been reported. Among seven non-OMIM genes, two show multiple transmissions of rare LOFs from either maternal or paternal lineage alone, where probands show neurodevelopmental disease and congenital anomalies. In conclusion, we have developed an enhanced map of POE of mCpG in human tissues significantly extending the current “imprintome”. Our results provide an improved and unbiased view of genomic imprinting in human and uncover previously underappreciated genes and variants likely crucial for human development with potential to expand human imprinting disease catalog.

Identification of *FXN* protomutation alleles explains the unequal population distribution of Friedreich ataxia

Authors: M. Tackett[1], C. Lam[1], E. Xiao[1], D. Lynch[2], S. Bidichandani[1]; [1]Univ. of Oklahoma Hlth.Sci. Ctr., Oklahoma City, OK, [2]Children's Hosp. of Philadelphia, Philadelphia, PA

Abstract:

Friedreich ataxia (FRDA) is typically caused by homozygous inheritance of an expanded GAA triplet-repeat in the *FXN* gene. FRDA is seen in Europe and South Asia, where the frequency of heterozygous carriers is 1%. FRDA is not seen in Sub-Saharan Africans and East Asians. Disease-causing expanded (E) alleles have >100 triplets, and have evolved from non-disease causing Long Normal alleles (LN, 12-30 triplets). LN and E alleles are found in FRDA-susceptible populations. However, Sub-Saharan Africans are the only non-susceptible population with LN alleles that have remarkably not transitioned to E alleles. In order to determine the molecular basis for the unequal global distribution of FRDA, we sequenced the entire *FXN* locus in multiple FRDA patients using short-read and long-read technologies. Eurasian and Sub-Saharan African GAA repeat lengths were determined by long-range PCR using genomic DNA from the 1000 Genomes Project. Long-read genomic sequences of a thousand African-Americans and Africans were obtained from the All of Us Research Program. Ancient human DNA sequences were obtained from the Allen Ancient DNA Resource. Haplotype analysis of Eurasian LN and E alleles showed that E alleles arose at least twice from a subset of LN alleles. These alleles, termed *protomutation* alleles, have 20-30 triplets. All Sub-Saharan African LN alleles are devoid of this key haplotype, and have remained under 20 triplets in length. We conclude that E alleles evolved from LN alleles via a key protomutation allele, which has existed in Europe and Western Asia for thousands of years. This occurred exclusively in Eurasia, which explains the current population distribution of FRDA.

Session 90: Tumor Genome Landscape Studies

Location: Four Seasons Ballroom 2&3

Session Time: Saturday, November 9, 2024, 8:00 am - 9:00 am

A common missense polymorphism in the *PARP1* gene is associated with distinct tumor transcriptomic, immune and clinical profiles in high grade serous ovarian cancers

Authors: G. Richenberg[1,2], T. Gaunt[1,2], S. Kar[1,2,3]; [1]MRC Integrative Epidemiology Unit, Univ. of Bristol, Bristol, United Kingdom, [2]Population Hlth.Sci., Bristol Med. Sch., Univ. of Bristol, Bristol, United Kingdom, [3]Early Cancer Inst., Univ. of Cambridge, Cambridge, United Kingdom

Abstract:

Poly (ADP-ribose) polymerase (PARP) inhibitors have had a transformative impact on the treatment of high-grade serous ovarian cancer (HGSOC). However, little is known about how germline variation in the *PARP1* gene affects PARP protein (the enzyme targeted by PARP inhibitors) levels and if such germline variation in turn influences the tumor multi-omic landscape of HGSOC. We used genome-wide association and fine-mapping analyses to identify a missense variant (rs1136410-C, frequency=16%) in *PARP1* that was associated (P=1.8x10-54) with lower circulating PARP1 levels in 33,657 women of European ancestry in the UK Biobank. This allele has been shown to reduce PARP1 enzyme catalytic activity in a dose-response manner. Therefore, we used the number of C alleles at rs1136410 as a marker of germline genetically predicted PARP inhibition. We performed germline (blood) genotype quality control and imputation into the 1000 Genomes reference panel on data from 389 HGSOC cases of genetically inferred European ancestry from The Cancer Genome Atlas (TCGA). We counted the number of C alleles at rs1136410 for each case in the TCGA cohort as a proxy for PARP inhibition. Multivariable linear and Cox models adjusted for age and stage were used to test for associations between rs1136410-C and HGSOC tumor transcriptomic, immune, and survival data in TCGA. We ranked 15,696 genes genome-wide based on the association between their tumor expression and *PARP1* rs1136410-C allele count and performed gene set enrichment analysis on the ranked list to identify associations between increasing numbers of PARP inhibitory C alleles at rs1136410 and downregulation of genes in the G2M cell cycle checkpoint (false discovery rate (FDR)=0.001) and upregulation of genes in the angiogenesis (FDR=0.03) pathways. Each additional copy of the PARP inhibitory C allele of rs1136410 was associated with increased CD8 T cell infiltration (Z-score=2.82; FDR=0.07) on evaluation of

13 tumor immune cell infiltrates quantified by the CIBERSORT algorithm. Among HGSOC cases harboring the rs1136410-C allele, those with pathogenic germline *BRCA1/2* (*gBRCA*) mutations showed better overall survival than those with wildtype *gBRCA* (HR=0.4; 95%CI=0.11-0.98) and this effect on survival was markedly stronger than among HGSOC cases without the PARP inhibitory rs1136410-C allele (HR=0.4; 95%CI=0.11-0.98). Our findings suggest that a common missense polymorphism in *PARP1* predictive of PARP inhibition has impacts beyond synthetic lethality on angiogenesis and anti-tumor immune infiltration in HGSOCs and offers the possibility of further understanding mechanisms of PARP inhibitor action and resistance.

Characterization of the immunosuppressive microenvironment driven by HBV-infected tumor cells in hepatocellular carcinoma through single-cell sequencing

⭐

Authors: S-J. Paek[1,2], J-H. Choi[1,2], H. Hwang[1,2], H. Woo[1,2,3], K-T. Kim[1,2]; [1]Dept. of Physiology, Ajou Univ. Sch. of Med., Suwon, Korea, Republic of, [2]Dept. of BioMed. Sci., Ajou Univ. Graduate Sch., Suwon, Korea, Republic of, [3]Ajou Univ. Data Ctr. for Biomedicine & Pharmacotherapeutics (AUDC-BMPT), Ajou Univ. Sch. of Med., Suwon, Korea, Republic of

Abstract:

Chronic hepatitis B virus (HBV) infection is a major etiological factor in hepatocellular carcinoma (HCC), marked by persistent inflammation and immune evasion within the tumor microenvironment (TME). While immunotherapy combined with antiviral treatment has emerged as a crucial strategy for managing patients with HBV-induced HCC, the mechanisms by which HBV-infected tumor cells inhibit the activation of immune system are not fully understood. To address this, we employed single-cell RNA sequencing (scRNA-seq, 10X Chromium) on HCC (n=40) and adjacent normal liver tissues (n=17). HBV sequences were identified by de novo assembly and subsequent mapping to a modified human genome reference. This method enabled the detection of HBV-infected cells from standard scRNA-seq data, along with the specific identification of HBV genotypes. Notably, HBV+ tumor cells were predominantly found in HBsAg+ patients containing genotype C, the most common HBV genotype in the Korean population.We observed significant transcriptional changes in HBV+ tumor cells driven by the HBx protein, affecting cytokine production and transcription factor expression contributing to oncogenic pathways. Additionally, HBV-infected cells were associated with an immunosuppressive microenvironment in HCC tissues from HBsAg+ patients. This environment was marked by reduced CD8+ T and natural killer cells along with diminished type 1 helper T cell activity.

Furthermore, an increase in CD8+ exhausted T cells, promoted by elevated LAMP3+ dendritic cells (DCs), indicated an immune evasion signature. These exhausted CD8+ T cells in HBsAg+ patients showed a distinct differentiation trajectory with tissue residency marker expression compared to activated Tem cells, exhibiting reduced cytotoxicity. Further, cell-cell interaction analysis highlighted T cell suppressive crosstalk between exhausted CD8+ T cells, LAMP3+ DCs, and HBV+ tumor cells, characterized by elevated immune checkpoint signaling. Specifically, the TIGIT-NECTIN2 axis was upregulated due to HBx-mediated alterations in transcription factor activity, suggesting that both chronic infection and direct HBV induced expression changes contribute to immune evasion.This study elucidates the unique characteristics of HBV-infected tumor cells and their role in establishing an immunosuppressive TME in HCC using standard scRNA-seq data. These insights provide a foundation for developing targeted therapies addressing the HBV-HCC axis, particularly by modulating immune checkpoint pathways to enhance anti-tumor immunity.

Single-cell RNAseq revealed multiple resistance mechanisms in patient-derived xenograft model of rectal cancer during treatment

Authors: Y. Song[1,2], R. Irwin[3], K. Hardiman[3,4], Z. Chong[1,2]; [1]Dept. of BioMed. Informatics and Data Sci., Univ. of Alabama at Birmingham, Birmingham, AL, [2]Dept. of Genetics, Univ. of Alabama at Birmingham, Birmingham, AL, [3]Dept. of Surgery, Univ. of Alabama at Birmingham, Birmingham, AL, [4]Dept. of Surgery, Birmingham VA Med. Ctr., Birmingham, AL

Abstract:

Background: Colorectal cancer is the third deadliest cancer in the US, with one-third of these cases being rectal cancer (RC). While approximately 20% of RC patients have a complete response to chemoradiation and may be able to avoid surgery, therapeutic resistance undermines treatment efficacy. The genetic heterogeneity observed in RC, particularly in treatment-resistant sub-clones, suggests that this diversity could be a harbinger of adverse outcomes. A deeper understanding of sub-clone treatment responses is vital for the development of targeted therapies.**Methods:** In our study, we utilized patient-derived xenograft (PDX) models and scRNA sequencing to analyze the cellular landscape of RC tumors. PDX models were established from primary tumors of two pre-treatment RC patients and subjected to a chemoradiation treatment regimen, including 2 weeks of 2 Gy radiation and 100mg/kg capecitabine daily followed by a week-long break. Tumor samples were collected at various time points for sequencing.**Results:** Our analysis revealed five distinct tumor clusters with varying responses to treatment in PDX model

from one patient. Two clusters demonstrated initial treatment sensitivity, followed by a resurgence, indicating potential therapy resistance. Marker gene expression profiles linked these clusters to critical oncogenic pathways, including HIF-1, p53, FoxO, and MAPK, which are integral to tumor progression and resistance. Cell-cell communication studies highlighted extensive interactions among tumor sub-populations, particularly in ECM-receptor interaction (LAMININ-ITG) and cell adhesion (CLDN3-CLDN3), which may influence treatment outcomes. Lineage tracing of tumor cells disclosed two differentiation pathways, with one showing increased resistance—a critical factor in treatment failure. Additionally, our investigation into copy number variations revealed genes located within the copy number alteration region (chr12p13.33-p13.31) exhibit regulatory interactions with the hypoxia-inducible factor-1 (HIF-1) signaling pathway, linking them to the dysregulation of tumor suppressor genes and oncogenes, further contributing to resistance. Meanwhile, the scRNA results from the other patient indicated a therapy-resistance tumor cluster associated with HIF-1, Hippo, Wnt signaling pathways.**Conclusion:** By examining two distinct PDX models, we identified both unique and shared resistance patterns, emphasizing the complexity and heterogeneity of RC tumors. Our comprehensive study provides new insights into the dynamic cellular mechanisms at play in RC, setting the stage for more effective, precision-targeted treatments.

Session 91: Unraveling the Complexity of Polygenic Inheritance

Location: Room 401

Session Time: Saturday, November 9, 2024, 8:00 am - 9:00 am

Beyond known genes and relationships for craniofacial abnormalities

Authors: E. Brokamp[1], N. Cox[2], A. Scalici[3], M. Shuey[2]; [1]Vanderbilt Univ. Med. Ctr., Nashville, TN, [2]Vanderbilt Univ Med Ctr., Nashville, TN, [3]Vanderbilt Univ., Nashville, TN

Abstract:

Genetic analyses and functional studies have long served as a foundation for understanding the etiology of craniofacial anomalies (CFAs). These studies have identified numerous causal genes, many of which are now offered as part of clinical genetic testing panels for individuals who are affected with CFAs. Despite the wealth of high-quality data produced by these studies, there remains a lot to be learned about the polygenic nature of these anomalies and the complex nature of the developmental processes underlying their development. The existence of large-scale genomic biobanks provides an opportunity to expand our understanding of the complex drivers of CFAs. Using Vanderbilt University Medical Center's biobank, BioVU, we performed a transcriptome wide association study to identify genes whose genetically predicted gene expression (GPGE) was significantly associated with CFAs. Because previous studies have identified 391 genes with known links to CFAs or craniofacial morphology and because congenital anomalies (CAs) frequently occur in multiples, we performed regression analyses to test for significant associations between the GPGE of these genes and other CAs. All CA diagnoses were made using phecodes, an approach that aggregates International Classifiers of Disease (ICD v9 and 10) billing codes to represent a particular phenotype. Ten genes (*ZNF77*, *MSH2*, *CD164, ANK1*, *LMO2*, *CHL1*, *L3MBTL2*, *ATP6V1E2*, *TIMM21*, *MRPL23*) had a significant association between their GPGE and CFAs (Bonferroni $p < 5.4x10^{-5}$). Interestingly, none of these ten genes were in our curated list of known craniofacial genes. Many of these genes, however, were identified in previous as studies as being involved in embryo development or associated with other CAs. From our curated list of 391 genes only four (1.02%) demonstrated even a moderately significant association with CFAs ($p < 0.001$). We did determine that 39 (10.01%) of these genes demonstrated moderately significant associations with other congenital anomalies. For example, increasing GPGE of *TXNRD2*, a gene present on multiple cleft lip/palate diagnostic genetic panels, was significantly associated with digestive CAs and CAs of the limbs ($p= 1.9x10^{-4}$ and $2.3x10^{-4}$, respectively)

but not CFAs (p = 0.05). Our preliminary work in a single center clinical biobank demonstrates the strength of utilizing big data resources for identifying genetic drivers of CAs. It also suggests that the root cause for many of these anomalies, including craniofacial, is complex and often polygenic in nature. Large biobank databases with associated phenotypes can serve as an impactful tool to further our understanding of CFAs.

Non additive interactions between rare variants and lifestyle factors contribute to obesity

Authors: D. Banerjee[1], S. Girirajan[2]; [1]Penn State, State College, PA, [2]Pennsylvania State Univ., State College, PA

Abstract:

Obesity is a complex disorder where a multitude of genes and environmental factors interact to express the disease phenotype. While previous efforts have primary concentrated on monogenic and polygenic architectures of obesity, we sought to identify oligogenic combinations of genes with rare protein truncating variants that are significantly associated with obesity in UK Biobank (N = 486,657) and All of Us (N = 245,388) cohorts. The 819 oligogenic combinations identified through our analysis have a substantial effect on body mass index (BMI; 4.12 kg/m^2 average increase; p < 0.001) and individuals carrying the combinations have higher odds of being overweight (OR = 1.98), obese (OR = 7.95) or severely obese (OR = 11.54) than those with monogenic and polygenic risk factor. Notably, we identified oligogenic partners for known monogenic obesity genes such as *MC4R*, *POMC* and *ADCY3* which interact non-additively with the monogenic risk factor to significantly increase BMI (p < 0.001). Furthermore, oligogenic combinations interact synergistically with common variant derived polygenic score for BMI (PGS_{BMI}; p < 0.001) and increase its penetrance, with combination carriers belonging to the highest PGS_{BMI} decile showing 5.5 times higher risk of obesity (BMI > 30 kg/m^2) and 10.2 times higher risk of severe obesity (BMI > 40 kg/m^2) than non-carriers. The increased penetrance amongst combination carriers towards obesity phenotypes was observed across all PGS_{BMI} deciles. Functional studies suggest that the oligogenic genes are involved in metabolic pathways and expressed in musculoskeletal tissues in contrast to previously known obesity genes that are mainly involved in neuronal pathways and expressed in brain tissues. Additionally, we found oligogenic genes interacting non-additively with obesogenic lifestyle factors to further exacerbate the disease risk (4.93 kg/m^2 average increase in BMI; p < 0.001). A collective risk model incorporating oligogenic combinations explained 12.31% of the BMI

variance compared to 8.53% explained by a model without these combinations. Overall, our findings challenge the existing landscape of genetic screening for complex disorders which primarily focuses on monogenic variants or polygenic risk scores and emphasize the importance of considering oligogenic risk when assessing predisposition to complex disorders such as obesity.

Polygenic risk of rheumatoid arthritis regulates the abundance of circulating regulatory T cells

Authors: K. Asahara[1], M. Kono[1,2], M. Nakano[1], T. Arakawa[1], T. Kawashima[1], K. Ishigaki[1,2,3]; [1]Lab. for Human Immunogenetics, RIKEN Ctr. for Integrative Med. Sci., Kanagawa, Japan, [2]Dept. of Microbiol. and Immunology, Keio Univ. Sch. of Med., Tokyo, Japan, [3]Keio Univ. Human Biology-Microbiome-Quantum Res. Ctr. (WPI-Bio2Q), Tokyo, Japan

Abstract:

Rheumatoid arthritis (RA) is a prototype of an autoimmune disease with an unknown etiology. Previous genetics studies successfully provided an effective polygenic risk score (PRS) model, identifying a small fraction of populations with substantially high genetic risk for RA. Theoretically, this enables us to intervene with donors with high RA-PRS to prevent disease onset. However, there are no established primary preventive measures for RA, which could become a major hurdle to implementing PRS in the health care system. In this study, we aimed to overcome the current situation and understand the immunological mechanism of how the RA polygenic risk dysregulates the immune system, contributing to the loss of self-tolerance and predisposing healthy donors to develop RA. For this aim, we sought to test the association between RA-PRS and immune cell abundance in peripheral blood. To maximize statistical power, we carefully designed our study structure. We utilized pre-generated genotype data and stock frozen peripheral blood mononuclear cell (PBMC) samples obtained from 52,892 donors in the Tohoku Medical Megabank, a community-based cohort in Japan. By calculating RA-PRS in this large cohort, we selected 120 healthy donors with extreme values of PRS, including 60 with the highest PRS (high-PRS) and 60 with the lowest PRS (low-PRS), and we treated PRS as a binary categorical variable. We then obtained PBMCs of these 120 donors and conducted cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) to unbiasedly quantify immune cell abundance in the peripheral blood. After stringent QC and batch correction, 470,289 cells remained in the analysis. We used weighted-nearest-neighbor method to integrate the cell surface protein marker and transcriptome data and identified 33 fine-resolution clusters.

We then evaluated cell type abundance differences between PRS-high and -low groups using covarying neighborhood analysis (CNA) and mixed-effects modeling of associations of single cells (MASC). CNA identified PRS-associated cell populations (global *P* value = 0.001), indicating the substantial impact of PRS on the cellular abundance. Among all clusters, MASC revealed the significant association with one cluster, CD45RA+ naïve regulatory T cell (T-reg): its frequency in high-PRS group was 1.08% and that in low-PRS group was 0.86% (OR = 1.21; *P* value = 0.001). This is the first study to identify a solid link between PRS and high-level immune phenotypes. Our results are consistent with previous studies reporting the RA heritability enrichment in the regulatory region of Tregs and provide additional genetic evidence supporting its causal role in RA risk.

An atlas of associations between plasma proteins biomarkers and polygenic risk scores for cancer and other complex human diseases

Authors: D. Chasioti[1], P. C. Haycock[2], G. Hemani[2], L. Andrews[2], C-Y. Chen[1], T. Gaunt[2], C. D. Whelan[1], B. B. Sun[1]; [1]Biogen Inc., Cambridge, MA, [2]Univ. of Bristol, Bristol, United Kingdom

Abstract:

Proteins are the primary drug targets, and routinely measured as disease biomarkers. Information regarding proteins can be valuable for advancement of precision medicine regarding polygenic disorders. We present an extensive catalogue of biologically meaningful associations between 33 disease-specific PRSs and 1,463 blood plasma proteins. We combine observational associations, disease-specific polygenic risk score (PRS), bi-directional Mendelian randomization (MR) and genetic colocalization approaches to build a genotype-phenotype map of causal protein-disease relationships. Disease-specific PRSs were developed utilizing data from the UKB-PPP study, to identify candidate proteins. Bi-directional Mendelian randomization and colocalization approaches were applied to capture potential causal biomarkers. Finally, pathway analysis was performed to increase results interpretability. Of the 33 diseases and 1,463 proteins, we identified 1,041 PRS-protein associations for 21 complex diseases ($p < 2.30e-6$). Of these, 92% indicate causal effects of disease liability on protein levels. To account for potential horizontal pleiotropy bias, we applied reverse Mendelian randomization (assesses the effect of disease liability on protein levels) combined with heterogeneity analysis. Although PRS was better powered than the MR approach, the two approaches were well correlated. The causality of the PRS findings was further assessed by cis and trans forward MR (assesses the effect of protein on disease). Forward cis-MR captured 20 proteins that were associated with disease (FDR<0.05), 10 of which colocalized. PRS results were more

strongly correlated with the trans-MR than cis-MR findings, consistent with a “reverse causation” interpretation, i.e. PRS results mostly reflect causal effects of disease liability on protein concentration rather than effects of proteins on disease. Among the proteins with evidence for causal effects on disease, we identified known drug targets (*PCSK9* for coronary artery disease, *IL2RA* for multiple sclerosis) and potential new drug targets (*BTN3A* for cancer). The potential role of the findings as pre-diagnostic biomarkers was further supported by observational studies. For example, lung cancer PRS identified proteins whose changes can increase the risk of lung cancer diagnosis as much as 93%. These results were well correlated with results from the LC3 study (prospective observational study of imminent lung cancer diagnosis). The findings indicate that disease PRS could be informative on the pathogenesis and the progression of complex diseases and ultimately, promote the efforts towards drug development.

Session 92: Genetic Information in Breast Cancer Risk Assessment and Screening

Location: Room 505

Session Time: Saturday, November 9, 2024, 9:30 am - 10:30 am

Comprehensive Genetic Risk Assessment for Breast Cancer in a Diverse Cohort: Preliminary Findings from the eMERGE Study

Authors: C. Liu[1], G. Wiesner[2], W. Chung[3], C. Katherine[1], J. Morse[2], K. McGuffin[2], J. Peterson[2], J. Linder[2], N. Lennon[4], C. Kachulis[4], A. Bick[2], K. Rita[1], C. Weng[1]; [1]Columbia Univ., New York, NY, [2]Vanderbilt Univ. Med. Ctr., Nashville, TN, [3]Boston Children's Hosp., Boston, MA, [4]Broad Inst. of MIT and Harvard, Cambridge, MA

Abstract:

The eMERGE network has enrolled 25,000 diverse individuals from ten different clinical sites. Each individual underwent genotyping and targeted sequencing, with their clinical data including family health histories (FHH) collected via surveys, electronic health records (EHR), and the MeTree FHH tool. These data were combined to generate individual Genomic Informed Risk Scores (GIRA) for 10 conditions and returned to the individuals and their healthcare providers. For breast cancer risk assessment, we created an integrated risk score using a customized BOADICEA model, incorporating a 309 SNP-based polygenic risk score (PRS), high-penetrance variants (including BRCA1, BRCA2, and PALB2), family history, and lifestyle, reproductive, and other clinical factors. Our preliminary analysis included 10,340 women who met the eligibility criteria (age >=18, self-reported BMI available, and without previous breast cancer diagnosis) for GIRA generation and return. Median age was 49 years (range, 18 - 75) and self-reported race/ethnicity included (multiple responses allowed): 5,499 (53%) White, 1704 (16%), Black, 2513 (24%) Hispanic, and 902 (9%) Asians. Of these, 8,715 (84%) completed targeted gene sequencing, identifying pathogenic variants in 14, 21, and 11 individuals (0.4%) for BRCA1, BRCA2, and PALB2, respectively. Furthermore, 318 individuals (5.4% of 5,878 PRS completed) had a PRS score above the top 5% of individuals with similar genetic ancestry, based on a model of PRS as a function of genetic ancestry trained on individuals of the All of Us Research Program. Incorporating all factors, 167 individuals (1.9% of 8,882 GIRA completed) had a lifetime breast cancer risk over 25%. This ratio is consistent across self-reported racial/ethnic groups: 86 (1.9%) White, 28 (1.8%) Black, 50 (2.1%) Hispanic, and 14 (1.9%) Asian. Notably, unadjusted analyses of BOADICEA factors showed 91 (54%) individuals with a life-time risk over 25% had a PRS in the top 5%, significantly more than the 3.8% in the low-risk group (p

< 0.05), indicating the strong influence of PRS in the integrated risk. Our preliminary findings suggest similar GIRA scores across racial/ethnic groups based upon PRS and non-genetic risk factors.

Investigating genotype-estrogen interactions in breast cancer through a combined molecular and epidemiological approach

Authors: I. Elfaki[1], L. Kellman[1], R. Meyers[1], D. Porter[1], P. Middha[2], T. Nakase[3], L. Kachuri[4], P. Khavari[1]; [1]Stanford Univ., Palo Alto, CA, [2]Univ. of California in San Francisco, San Francisco, CA, [3]Stanford Univ., Stanford, CA, [4]Stanford Univ., San Francisco, CA

Abstract:

13% of women in the US will develop breast cancer (BC) in their lifetime. BC risk is determined by a multitude of genetic and environmental risk factors, but the interaction of these factors on the molecular level is not well understood. These genotype-environment interaction (GxE) effects are difficult to capture and often left out of risk models. A better understanding of GxE risk could improve risk prediction and early detection of breast cancer. Estrogen (E2) exposure is an important risk factor for the development of all subtypes of BC. E2 signals through the G protein-coupled estrogen receptor (GPER-1) and the E2 receptor (ER), causing transcriptional changes. In this study, we took combined functional genomics and epidemiological approaches to characterize a subset of GxE effects in BC by identifying noncoding GWAS BC risk SNPs whose transcriptional activity is modulated by E2 exposure. For a set of 1,604 SNPs implicated in breast cancer risk, we used massively parallel reporter assays (MPRA) in ER+ BC cell lines in the presence and absence of E2 to examine their transcription-directing potential. We identified 73 SNPs for which differences in transcriptional activity between their alleles were modulated by the presence of estrogen (with FDR < 0.1 for evidence of interaction), termed **estrogen-modulated SNPs (emSNPs)**. Integrating with publicly available RNA-Seq, ATAC-Seq, ChIP-Seq data and motif enrichment tools, we show that emSNPs are enriched for motifs of TFs known to be involved in E2 signaling, including ESRRA, ZBTB7A, and BMAL1, as well as TFs yet to be associated experimentally. Next, we combined emSNPs into a polygenic risk score model (PRS) and tested it in 13,026 BC cases and 108,265 controls in the UK Biobank. MPRA-informed PRS was associated with postmenopausal BC and showed statistically significant interactions with 2 risk factors: age of first live birth (p = 0.0026) and reproductive interval (p = 0.0024). By directly interrogating the effects of E2 on the transcriptional activity of breast cancer risk SNPs, we were able to detect statistically significant GxE effects that have been missed in previous studies. Our work shows that

despite the complexity of these interactions in humans, a portion of GxE risk can be modeled in 2D culture, improving our ability to find these signals on a population level and advancing our molecular understanding of these interactions. Future directions include performing reporter assays for emSNPs in the context of ER and GPER1 knockouts to determine whether the observed GxE effects are mediated through ER or GPER1 signaling.

Finding the pathogenic variant in the haystack: using breast-cancer-related family history from electronic health records to identify patients who should be prioritized for genetic testing

Authors: D. Kiser[1], G. Elhanan[1], A. Bolze[2], I. Neveux[1], K. A. Schlauch[1], W. J. Metcalf[1], E. Cirulli[2], C. McCarthy[1], L. A. Greenberg[1], S. Grime[3], J. M. S. Blitstein[1], W. Plauth[3], J. J. Grzymski[1]; [1]Univ. of Nevada, Reno Sch. of Med., Reno, NV, [2]Helix, San Mateo, CA, [3]Renown Hlth., Reno, NV

Abstract:

Background Pathogenic or likely pathogenic (P/LP) variants in the *ATM*, *BRCA1*, *BRCA2*, *CHEK2*, and *PALB2* genes substantially increase the risk of breast and other forms of cancer. However, most patients with P/LP variants have not been identified, resulting in missed opportunities for preventative care. Ideally, population-wide genetic testing would be used to detect every affected patient, but limited resources may necessitate prioritization of patients whose family history indicates increased risk.
Question Can we identify patients with increased likelihood of having P/LP variants associated with breast cancer by applying criteria from the seven-question family history questionnaire (FHS7) to electronic health records (EHR)?
Methods In a population of 835,886 patients aged 18-79 years with at least one encounter where family or personal medical history was assessed, 38,007 patients were sequenced and P/LP variants in the *ATM*, *BRCA1*, *BRCA2*, *CHEK2*, and *PALB2* genes identified. Family history relevant for determining if patients were positive for FHS7 criteria (FHS7+) were extracted from a structured EHR table using both discrete codes and free-text comments.
Results 4,107 (15.6%) of 26,301 sequenced females and 835 (7.1%) of 11,699 sequenced males were FHS7+. Among females, being FHS7+ was associated with increased likelihood of P/LP variants in *BRCA1* and *BRCA2* (OR 3.3, 95% confidence interval [CI] 2.5-4.5), *CHEK2* (OR 1.6, 95% CI 1.1-2.4), and *PALB2* (OR 2.8, 95% CI 1.2-6.2), while the number needed to test to detect one patient with a P/LP variant in any gene (NNT) declined from 58 overall to 32. Among males, being FHS7+ was associated with increased likelihood of P/LP variants in *BRCA1* and *BRCA2* (OR 3.3, 95% CI 1.9-5.6), while the NNT declined

from 54 overall to 31. Utilizing free-text comments in addition to discrete family history codes was essential for identifying at-risk patients, as it tripled the number of patients meeting criteria. Among the 796,618 unsequenced patients with no evidence of genetic testing in their EHR, 24,538 (3.1%) were FHS7+.
Conclusion In a large sequenced population, patients having an EHR that met FHS7 criteria were significantly more likely to have P/LP variants associated with breast cancer. We also identified 24,538 patients in our health system who have likely not been tested but who have an increased likelihood of having P/LP variants according to FHS7. Prioritizing these patients for genetic testing would help maximize the benefits obtained from limited public health resources.

Polygenic risk score (PRS) significantly improves breast cancer (BC) risk assessment for diverse ancestries

Authors: M. Kucera[1], T. Simmons[1], **E. Hughes**[1], D. Pruss[1], B. Roa[1], T. Judkins[1], A. Younus[1], S. Cummings[1], S. Jammulapati[1], K. M. Timms[1], A. W. Kurian[2], H. J. Pederson[3], P. W. Whitworth[4], S. Wagner[1], D. Muzzey[1], J. S. Lanchbury[1], A. Gutin[1]; [1]Myriad Genetics, Inc., Salt Lake City, UT, [2]Stanford Univ. Sch. of Med., Stanford, CA, [3]Cleveland Clinic, Cleveland, OH, [4]Nashville Breast Ctr., Nashville, TN

Abstract:

Background: Accurate BC risk assessment is essential to identify women for whom screening and preventive interventions may be lifesaving. Incorporation of PRS into clinical models can improve risk prediction, but most PRS have shown suboptimal performance among non-Europeans. We previously described a multiple-ancestry PRS (MA-149) based on 56 ancestry-informative and 93 BC-associated single-nucleotide polymorphisms (SNPs). MA-149 achieved accuracy for diverse populations by characterizing genetic ancestry at each SNP in terms of fractions attributable to reference ancestries and applying ancestry-specific risks and frequencies. MA-149 significantly outperformed the Tyrer-Cuzick (TC) model, and integration of MA-149 with TC improved predictive accuracy by roughly two-fold over TC alone. Here, we aimed to improve MA-149 by expanding the set of BC SNPs and refining ancestry-specific risks.
Methods: We developed a novel stepwise regression methodology accounting for linkage disequilibrium to select an optimal set of BC SNPs. Women referred for hereditary cancer testing and negative for pathogenic variants in BC genes were divided into consecutive cohorts to (1) refine ancestry-specific SNP risks, (2) develop the PRS (*N*=184,322), and (3) conduct independent validation (*N*=146,110). Predictive accuracy and calibration of the

new PRS were evaluated in the full cohort and subpopulations of different ancestries. Odds ratios (OR) from multivariable regression are reported per standard deviation.
Results: A set of 385 SNPs (56 ancestry, 329 BC) was selected for the new PRS (MA-385). Within each ancestral group, MA-385 improved upon clinical factors and showed a significant improvement in risk prediction compared to MA-149. Among women of non-European descent, MA-385 was a more effective risk stratifier (OR=1.47, 95% CI: 1.43-1.52) than MA-149 (OR=1.40, 95% CI: 1.35-1.45). The strongest associations were observed in Ashkenazi Jewish (OR=1.80, 95% CI: 1.52-2.14) and Hispanic (OR=1.63, 95% CI: 1.53-1.73) women. MA-385 identified more women at >2-fold increased risk than MA-149 (6.3 % vs 2.0%). Goodness-of-fit tests showed that MA-385 was well-calibrated while a European-derived PRS was miscalibrated for non-Europeans.
Conclusions: MA-385 was well-calibrated, improved upon clinical factors, and outperformed existing PRS in all tested ancestries. Incorporation of MA-385 into risk assessment could improve the early detection and prevention of BC.

Session 93: Modeling Ataxia and Neuropathy

Location: Room 501

Session Time: Saturday, November 9, 2024, 9:30 am - 10:30 am

Sumo1 mutation modifies behavioral performances in Fragile X associated tremor/ataxia mouse model

Authors: Y. Gu, D. Nelson; Baylor Coll. of Med., Houston, TX

Abstract:

Fragile X associated tremor/ataxia Syndrome (FXTAS) is a late onset neurodegenerative disorder that displays intention tremor, ataxia, cognitive and autonomic impairment. It is associated with premutation-length (55-200) CGG repeats in *FMR1*. Some 40% males and 16% females carrying *FMR1* premutation develop the signs and symptoms of FXTAS by age 70, suggesting that genetic and environmental modifiers can impact development of the condition. A search for genetic modifiers of FXTAS and Fragile X associated primary ovarian insufficiency (FXPOI) identified proteasome 20S subunit beta 5 (*PSMB5*) and small ubiquitin like modifier 1 (*SUMO1*). Both were able to affect phenotypes in a Drosophila model of the disorder. To further define the mechanism of FXTAS and outline possible therapeutic approaches, we created a mouse model of FXTAS, that impacts Purkinje neurons with specific expression of CGG repeats using Purkinje neuron specific L7 driven Cre/loxP. We also created knockout mice for *Psmb5* and *Sumo1* to study whether reduced expression of these genes can modify cerebellar phenotypes and bred heterozygotes into the FXTAS animals. Several behavioral tests including rotarod, open field activity, and parallel rod footslip were performed in the FXTAS Purkinje cell mouse model carrying heterozygous knockout alleles for *Psmb5* and *Sumo1* and appropriate controls (FmrpolyG^{+}, L7Cre^{+}, Psmb5$^{+/-}$) and (FmrpolyG^{+}, L7Cre^{+}, Sumo1$^{+/-}$) at several ages (4, 7, 14 months). We found that the FXTAS Purkinje cell mouse model showed behavioral phenotypes beginning at 14 months, consistent with the late age onset nature of the human condition. Male mice expressing the FmrpolyG+ allele in Purkinje neurons exhibited significant impairment in rotarod performance when compared with wild type control animals at 14 months. Psmb5$^{+/-}$, FmrpolyG^{+}, L7Cre^{+} male mice were similarly impaired in rotarod performance as FXTAS male animal when compared with wild type males at 14 months. In contrast, mice heterozygous for knockout of *Sumo1* showed altered phenotypes mice in several behavioral tests, with improved performance in rotarod in both males and females, more mobile time, episodes, less immobile time, episodes in males in parallel rod footslip. less stereotypy time, episode count, and activity count in females in open field activity. These

data suggest the possibility that reduction of *Sumo1* can improve abnormal phenotypes in mouse models and conform with prior data from fly models. Small molecules that target *Sumo1* may provide a path toward therapeutic intervention. Additional studies of reduction of *Sumo1* on other mouse models of FXTAS and analyses of brain pathology are underway.

SNX13 and SNX14 influence neuronal lipid homeostasis and associate with spinocerebellar ataxia and intellectual disability syndromes

Authors: N. Akizu[1], F. Chiara[2], S. Lee[1], W. Henne[3], F. Alkuraya[4], J. Gleeson[5], B. Keren[6], A. Afenjar[6], G. Ravenscroft[2], M. Clementina[7], V. B. Sanchez[1]; [1]Children's Hosp. of Philadelphia, Philadelphia, PA, [2]Harry Perkins Inst. of Med. Res. and Ctr. for Med. Res., Univ. of Western Australia, Nedlands, Australia, [3]Dept. of Cell Biology, UT Southwestern Med. Ctr., Dallas, TX, [4]King Faisal Specialist Hosp. & Res. Ctr., Riyadh, Saudi Arabia, [5]Univ. of California, San Diego, La Jolla, CA, [6]UF de Génomique du Développement, Département de Génétique médicale, Groupe Hosp.ier Pitié-Salpêtrière, AP-HP Sorbonne Université, Paris, France, [7]Dept. of Systems Pharmacology and Translational Therapeutics, Perelman Sch. of Med., Univ. of Pennsylvania, Philadelphia, PA

Abstract:

Genetic variants associated with disturbed lipid homeostasis are implicated in a plethora of neurological disorders. We previously found that recessive loss of function variants in *Sorting Nexin 14* (*SNX14*) cause a childhood onset spinocerebellar ataxia with intellectual disability syndrome (SCAR20) associated with neuronal lipid storage and homeostasis defects. To determine how SNX14 regulates lipid homeostasis in neurons we analyzed its neuronal interactome and identified SNX13 as the major interacting partner. SNX13 is the closest homologue of SNX14 that together with SNX19 and SNX25 form a newly characterized SNX-RGS subfamily of SNX proteins. SNX-RGS proteins are characterized by having a PXA and a PXC domain, both of which were recently predicted to interact intramolecularly to form non-vesicular interorganelle lipid transfer tunnels. Our current data shows that these two domains are also necessary for the interaction between SNX13 and SNX14, suggesting that they may cooperate to modulate interorganelle lipid transfer in neurons. To test this hypothesis, we knocked down (KD) *SNX13* and/or *SNX14* in human pluripotent stem cell-induced neurons (i^{3}Neurons) using CRISPRi, and analyzed the effects on lipid storage and metabolism through confocal microscopy and ultra-high performance lipidomics respectivelly. Results revealed that loss of either *SNX13* and/or *SNX14* alter lipid droplet biogenesis and lipid metabolism in

i[3]Neurons. Remarkably, our recent genetic studies have uncovered loss of function variants in *SNX13* as the genetic basis of a SCAR20-like syndrome in four patients of three unrelated families. Together, our work shows that both SNX13 and SNX14 influence neuronal lipid homeostasis leading to a pediatric onset cerebellar neurodegenerative disorder when disrupted.

Investigating R-Loop Formation as a Potential Pathomechanism in Spinocerebellar Ataxia 27B Using iPSC-Derived GABAergic Neurons

Authors: G. Spurdens[1], A. Rebelo[2], D. Pellerin[3], M. Danzi[4], C. Yanick[1], M. Saporta[5], B. Brais[6], S. Zuchner[1]; [1]Univ. of Miami Miller Sch. of Med., Miami, FL, [2]Univ Miami, Miami, FL, [3]McGill Univ., Montreal, QC, Canada, [4]Univ. of Miami, Dakota Dunes, SD, [5]Univ. of Miami, Miami, FL, [6]Mc Gill Univ., Outremont, QC, Canada

Abstract:

Spinocerebellar Ataxia 27B (SCA27B) is a prevalent adult-onset ataxia characterized by an intronic repeat expansion in FGF14, yet its underlying pathomechanisms remain elusive. Notably, SCA27B arises from the expansion of a GAA repetitive motif, akin to Friedreich's Ataxia (FRDA), the first disorder to be associated with this pathological repetitive motif. Leveraging the well-established understanding of pathomechanisms in FRDA, we hypothesize that SCA27B shares comparable underlying pathophysiological pathways. Here, utilizing induced pluripotent stem cells (iPSCs) derived from SCA27B patient fibroblasts, we differentiated them into GABAergic neurons to model the cerebellar phenotype more accurately. Employing CUT&Tag-IT methodologies, we aim to elucidate potential R-loop formations proximal to the SCA27B locus, exploring their likely involvement in disease pathology. Our ongoing investigation hypothesizes that R-loop formation within this GAA repetitive region may disrupt transcription, leading to reduced expression of Fibroblast Growth Factor 14 (FGF14) and contributing to the observed phenotype. If substantiated, these findings could shed light on a previously unidentified pathomechanism in SCA27B, underscoring the significance of R-loop dysregulation in repeat expansion disorders and suggesting promising avenues for therapeutic exploration.

Predictive modeling to define the locus heterogeneity of tRNA synthetase-related peripheral neuropathy

Authors: A. Cale, A. Antonellis; Univ. of Michigan, Ann Arbor, MI

Abstract:

Aminoacyl-tRNA synthetases (ARSs) are ubiquitously expressed, essential enzymes that ligate amino acids to tRNAs in the mitochondria or cytoplasm. Variants in seven genes encoding a cytoplasmic ARS cause dominant axonal peripheral neuropathy, which presents the question: how do variants in cytoplasmic ARSs, which are essential in all tissues, lead to phenotypes restricted to the peripheral nervous system? While defects in protein translation and activation of the integrated stress response have been implicated downstream of certain neuropathy-associated ARS variants, a unifying pathological mechanism that explains the locus heterogeneity has not been identified. Interestingly, all seven neuropathy-associated ARSs function as dimers and most pathogenic alleles are loss-of-function missense variants; these observations suggest a dominant-negative effect.

If a dominant-negative effect is the primary disease mechanism of ARS-mediated dominant peripheral neuropathy, then certain variants in any dimeric, cytoplasmic ARS could exert a dominant-negative effect and lead to dominant neuropathy. To test this, we engineered missense mutations in threonyl-tRNA synthetase (*TARS1*), a dimeric, cytoplasmic ARS not yet implicated in neuropathy, and are testing the effects of the mutations in multiple relevant models: yeast, worm, and mouse. We have tested seven variants to date: five variants are loss-of-function alleles in a humanized yeast complementation assay; and two of these loss-of-function alleles also repress wild-type *TARS1* in a humanized yeast dominant toxicity assay, consistent with a dominant-negative effect. We next introduced the *TARS1* allele with the most severe dominant-negative effect (H590A) into the endogenous worm *tars-1* locus using CRISPR-Cas9. We have confirmed that it is a loss-of-function allele in worm and are testing heterozygous worms for dominant neurologic and motor behavior phenotypes. Preliminary data suggest that heterozygous worms have decreased motility compared to wild-type worms. We have also successfully generated a mouse line that is heterozygous for H590A *Tars1* and will test these animals for neuropathic phenotypes in established behavioral and histopathological assays. Here, I will present all of our unpublished data from this study in which we aim to elucidate the mechanism by which numerous alleles across seven ARS genes converge on a dominant peripheral neuropathy phenotype.

Session 94: More than One Way to Break a Gene - Variant Effects on RNA

Location: Room 401

Session Time: Saturday, November 9, 2024, 9:30 am - 10:30 am

Unveiling the hidden role of RNA stability as a link between genetic variation and disease

Authors: E. Huang[1], T. Fu[1], G. Yan[1], L. Zhang[1], K. Amoah[1], R. Yamamoto[1], S. Terrazas[1], J. Hervoso[1], M. Paulsen[2], B. Magnuson[2], M. Ljungman[2], J. Li[1], X. Xiao[1]; [1]UCLA, Los Angeles, CA, [2]Univ. of Michigan, Ann Arbor, MI

Abstract:

Gene expression is jointly modulated by transcriptional regulation and mRNA stability, yet the latter is often overlooked in studies on genetic variants. Leveraging metabolic labeling data (Bru/BruChase-seq) and a new computational pipeline, RNAtracker, we categorize genes and their corresponding variants as either allele-specific RNA stability (asRS) or allele-specific RNA transcription (asRT) events. RNAtracker considers allele-specific expression measurements at multiple time points after transcription initiation and features the usage of a beta-binomial model to estimate the probability of various gene states (including asRS and asRT). These categorizations delineate whether the mechanism underlying transcript abundance differences between two alleles is due to differential degradation or transcription. After applying stringent confidence filters, we identify over 5,000 asRS variants among 680 genes across a panel of 14 ENCODE cell lines. These variants directly overlap conserved microRNA target regions and allele-specific RNA binding protein sites, including those of the RNA-stabilizing protein YBX3, the deubiquitinase UCHL5, and the RNA helicase required for nonsense-mediated decay (NMD) UPF1, illuminating mechanisms through which stability is mediated. Furthermore, we identified additional causal asRS variants using our massively parallel screen for variants that affect post-transcriptional mRNA abundance (MapUTR). Compared to asRT, asRS genes exhibit higher prevalence in the majority of cell lines analyzed. The predominance of asRS events was recapitulated in our analysis of tissue-specific eQTLs, which revealed higher enrichment of asRS compared to asRT variants. Notably, asRS genes were enriched in immune-related pathways, with many asRS variants significantly associated with multiple autoimmune diseases. Our analyses nominated RNA stability as a key mechanism underlying established genotype-disease relationships (e.g., variants in *ATG16L1* for Crohn's Disease and *IQGAP1* for multiple sclerosis). This work highlights

RNA stability as a critical, yet understudied mechanism linking genetic variation and disease.

Splice Switch: An investigation on the effect of a sQTL on PAPR2 isoforms and subsequent influenza A virus susceptibility

Authors: G. Connelly[1], L. Wang[2], B. Schott[1], D. Ko[1]; [1]Duke Univ., Durham, NC, [2]Duke Univ, Durham, NC

Abstract:

Natural genetic variation contributes to a range of clinical outcomes to influenza A virus (IAV) infections, ranging from asymptomatic to death. Previous human genome-wide association studies (GWAS) on IAV infection phenotypes have struggled to capture genetic variants that impact susceptibility. This is likely due to the limited power of existing studies. To address this issue, our lab has developed a scalable, cell-based GWAS approach called scHi-HOST (single cell High-throughput Human in vitrO Susceptibility Testing). In a single scHi-HOST experiment, we use scRNA-seq to quantify viral burden in each cell to calculate infection phenotypes and collect host transcriptomics to identify expression quantitative trait loci (eQTLs) and splicing quantitative trait loci (sQTLs). From a scHi-HOST IAV screen of 96 LCLs we have identified a sQTL that leads to a change in protein product of the gene *PARP2*, which encodes an ADP-ribosyltransferase. This sQTL is a single nucleotide polymorphism (SNP) which causes a change in the preferred splice donor. Entropy modeling shows that the minor allele leads to an increase in potential U1 snRNP binding affinity at the SNP location. The major allele, however, causes a decreased potential U1 snRNP binding affinity score at the same location resulting in the use of an earlier splice donor. This difference in splice donor causes individuals homozygous for the major allele to produce a protein product that is 13 amino acids shorter than individuals homozygous for the minor allele. We have verified this difference in protein product via western immunoblot from LCLs of both homozygous genotypes. Additionally, heterozygous individuals produce protein products of both sizes. Genetic association of scHi-HOST IAV phenotypes demonstrated that homozygous minor allele individuals are more susceptible to IAV infection compared to homozygous major allele individuals. RNAi in lung epithelial cells (A549s), with confirmed knock-down of *PARP2*, led to decreased IAV infection indicating pro-viral role. We are currently determining the mechanism through which the different protein isoforms leads to differences in IAV infection by creating stable over-expression cell lines of each PARP2 protein isoform and measuring genotype specific ADP-ribosylation. Our findings elucidate how a human genetic variant causes changes in protein

product and therefore, differences in IAV susceptibility in addition to shedding light on the role of PARPs during infection.

Modulation of the impact of genetic mutations on human health by transcriptional adaptation

Authors: M. El-Brolosy[1,2,3], A. Fischer[2], A. Hoang[2], A. Oak[2,4], O. Corradin[2,4], M. Daly[5,3], J. S. Weissman[2,6,4], K. Karczewski[5,3]; [1]Harvard Society of Fellows, Cambridge, MA, [2]Whitehead Inst. for BioMed. Res., Cambridge, MA, [3]Broad Inst. of MIT and Harvard, Cambridge, MA, [4]MIT, Cambridge, MA, [5]Massachusetts Gen. Hosp., Boston, MA, [6]Howard Hughes Med. Inst., Massachusetts Inst. of Technology, Cambridge, MA

Abstract:

Genetic variants that affect gene expression, such as protein-truncating variants (PTVs), can have a strong impact on human health. However, even these high-impact variants often have less pronounced effects than expected. One explanation for such behavior is due to transcriptional adaptation (TA), a recently described genetic robustness mechanism where pLoF variants triggering mutant mRNA decay can induce sequence-dependent upregulation of the unaltered copy of the gene, or paralogs. Despite advances in understanding the molecular mechanisms of TA, its physiological role and influence on genetic disorder landscapes remain unexplored.

Here, we explore the physiological role of TA by comparing PTVs that elicit mutant mRNA decay through nonsense-mediated decay (NMD) and deleterious missense mutations that do not. Using pQTL data from the UK Biobank's plasma proteomics, we observed that many heterozygous PTVs lead to a less-than-expected 50% decrease in protein levels, suggesting dosage compensation, possibly mediated through TA. Notably, we found that PTVs are more likely to upregulate paralogous genes compared to missense mutations, potentially leading to functional compensation. Utilizing data from gnomAD, ClinVar, and the UK Biobank, we identified 686 genes where PTVs are associated with milder phenotypic outcomes compared to missense mutations. Specifically, from gnomAD, we identified genes where pLoFs were less constrained than deleterious missense mutations. From ClinVar, we identified genes with a significantly higher number of pathogenic missense mutations than PTVs, that aren’t due to the natural distribution of these variant classes. Finally, we used rare variant association information from the UK Biobank (GeneBass) to identify genes where missense mutations had a significant association with a phenotype, but pLoFs did not.

Notably, CRISPR/Cas9 Perturb-seq experiments on 291 of these genes in iPSCs showed

significant upregulation of paralogous genes when perturbing 71 of these genes with PTVs, some of which include genes where missense mutations are linked to Mendelian disorders. For example, missense mutations in *CALM1*, encoding the calcium binding protein Calmodulin 1, are associated with cardiovascular disorders. PTVs introduced to *CALM1* triggered the upregulation of two calcium binding paralogs, *CABP5* and *EFCAB1*, potentially explaining why PTVs in *CALM1* are not associated with cardiovascular disorders. This analysis will enhance our understanding of how different variants influence disease outcomes and may provide new explanations on cases where PTVs are better tolerated compared to missense mutations.

Systematic analysis of nonsense variants uncovers peptide release rate as a novel modifier of nonsense-mediated mRNA decay efficiency

Authors: S. Jagannathan[1,2], D. Kolakada[3], R. Fu[4], N. Biziaev[5], A. Shuvalov[5], M. Lore[3], A. E. Campbell[1], M. A. Cortázar[1], M. P. Sajek[1,6], J. R. Hesselberth[1], N. Mukherjee[1], E. Alkalaeva[5]; [1]Dept. of Biochemistry and Molecular Genetics, Univ. of Colorado Anschutz Med. Campus, Aurora, CO, [2]RNA BioSci. Initiative, Univ. of Colorado Anschutz Med. Campus, Aurora, CO, [3]Molecular Biology Graduate Program, Univ. of Colorado Anschutz Med. Campus, Aurora, CO, [4]New York Genome Ctr., New York, NY, [5]Engelhardt Inst. of Molecular Biology RAS, Moscow, Russian Federation, [6]Inst. of Human Genetics, Polish Academy of Sci., Poznan, Poland

Abstract:

Nonsense variants underlie many genetic diseases. The phenotypic impact of nonsense variants is determined by Nonsense-mediated mRNA decay (NMD), which degrades transcripts with premature termination codons (PTCs). NMD activity varies across transcripts and cellular contexts via poorly understood mechanisms. Here, by leveraging human genetic datasets, we uncover that the amino acid preceding the PTC dramatically affects NMD activity in human cells. We find that glycine codons in particular support high levels of NMD and are enriched before PTCs but depleted before normal termination codons (NTCs). Gly-PTC enrichment is most pronounced in human genes that tolerate loss-of-function variants. This suggests a strong biological impact for Gly-PTC in ensuring robust elimination of potentially toxic truncated proteins from non-essential genes. Biochemical assays revealed that the peptide release rate during translation termination is highly dependent on the identity of the amino acid preceding the stop codon. This release rate is the most critical feature determining NMD activity across our massively parallel reporter assays. Together, we conclude that NMD activity is significantly modulated by the

“window of opportunity” offered by translation termination kinetics. Integrating the window of opportunity model with the existing framework of NMD would enable more accurate nonsense variant interpretation in the clinic.

Session 95: Phenomenal PheWAS

Location: Four Seasons Ballroom 2&3

Session Time: Saturday, November 9, 2024, 9:30 am - 10:30 am

Genome-wide association studies in a large Korean cohort identify novel quantitative trait loci for 36 traits and illuminate their genetic architectures

Authors: Y. Jee[1], Y. Wang[2], K. Jung[3], J. Lee[3], H. Kimm[3], R. Duan[1], A. L. Price[1], A. R. Martin[2], P. Kraft[4]; [1]Harvard T.H. Chan Sch. of Publ. Hlth., Boston, MA, [2]Massachusetts Gen. Hosp., Boston, MA, [3]Graduate Sch. of Publ. Hlth., Yonsei Univ., Seoul, Korea, Republic of, [4]Natl. Cancer Inst., Rockville, MD

Abstract:

Genome-wide association studies (GWAS) have identified many loci associated with complex traits. However, most of these studies were conducted in populations of European ancestry, limiting opportunities for biological discovery and generalizability. Here we report GWAS findings from 153,950 individuals across 36 quantitative traits in the Korean Cancer Prevention Study-II (KCPS2) Biobank. We discovered 616 novel genetic loci in KCPS2. As thyroid-stimulating hormone is measured in KCPS2 participants but is not typically measured at baseline in biobanks, we identified a particularly high novelty rate (55/100 loci); one such novel locus includes a missense variant (rs75326924) in *CD36*, reiterating its importance in hepatic fatty acid storage and transport in relation to hypothyroidism. This variant is common in KCPS2 (minor allele frequency [MAF]=0.068) but entirely absent in European populations. We also conducted meta-analysis of 21 traits across KCPS2, Korean Genome and Epidemiology Study (KoGES), Biobank Japan (BBJ), Taiwan Biobank (TWB), and UK Biobank (UKB) (N_{total}=928,679), which identified 11,861 loci, 3,524 of which were not significant in any of the contributing GWAS. We assessed the genetic architecture of these traits and demonstrated different heritability estimates across East Asian cohorts as well as across East Asian and European ancestry populations, reflecting differences in study design, sample size, and linkage disequilibrium. The S parameters linking MAF and effect sizes were similar across the biobanks (median=-0.59), suggesting a pervasive action of negative selection on the trait-associated variants. The median polygenicity estimates for the 8 traits available in all four studies were largest in UKB (median=0.02), followed by BBJ (median=0.007), KCPS2 (median=0.006), and TWB (median=0.001), which follows the same order as the sample sizes of the biobanks. We also highlight associations with alleles that are common in East Asian but rare in European populations, including a known pleiotropic missense variant in *ALDH2* (rs671) associated with 26 traits. Fine-

mapping analyses identify rs671 as a likely causal variant for diverse sets of traits including liver enzymes, blood pressure, and lipid values. Our findings highlight how broadening the population diversity of GWAS participants can aid discovery and provide insights into the genetic architecture of complex traits in East Asian populations. By increasing the sample size and ancestral diversity of GWAS samples, our analysis may help identify novel targets for prevention and treatment and offer equitable access to precision medicine to diverse populations.

The phenomic landscape of gain- and loss-of-function genetic variants across diverse human populations

Authors: M. Kars, D. Stein, Y. Itan; Icahn Sch. of Med. at Mount Sinai, New York, NY

Abstract:

Phenome-wide association studies (PheWAS) have been instrumental in investigating genotype-phenotype correlations in large-scale biobanks. Recently, several PheWAS resources have become available, utilizing electronic health records based on International Classification of Diseases (ICD) diagnoses and genotypes predominantly from European populations. However, current resources have certain limitations, such as the lack of clinically-relevant phenotype definitions, underrepresentation of non-European populations and a limited scope of laboratory measurements. Additionally, none of these resources have distinguished between the phenotypic consequence of loss-of function (LoF) and gain-of-function (GoF) variants, which have been shown to result in different phenotypes due to their distinct effects on proteins. To address these limitations, we developed a comprehensive PheWAS resource containing ancestry-specific and panancestral analysis results. We utilized The Mount Sinai BioMe BioBank, a hospital-based biobank comprising African, European and Hispanic/Latinx American participants, and predicted LoF and GoF missense variants obtained from LoGoFunc, a machine learning classifier tailored to predict LoF, GoF and neutral variants. Utilizing two separate whole-exome sequencing (WES) cohorts of BioMe, including 27,739 and 14,186 participants, we systematically performed association testing of approximately 26,000 LoF, 2,000 GoF and over 100,000 predicted neutral variants across 1,500 binary phenotypes (phecodeX), and nearly 500 quantitative phenotypes curated from laboratory measurements in each ancestral group. In total, we performed over a billion association tests, yielding nearly 517 million high quality associations. Our analyses revealed population-specific associations for numerous variants and genes. For instance, we identified a significant association between a predicted GoF variant in *CACNA1C* and an

increased risk of ventricular tachycardia specifically in the African American cohort. Similarly, a predicted LoF variant in *PSEN1* was associated with an increased risk of dementias specifically in the Hispanic/Latinx American cohort, while a predicted LoF variant in *HFE* was associated with an increased risk of disorders of iron metabolism in the European American cohort. These findings underscore the importance of considering genetic diversity in disease genomics and the potential contribution of our comprehensive resource in advancing precision medicine research.

Phenome-wide study reveals multiple diseases and biomarkers causally associated with alcohol consumption

Authors: N. Kassaw; Univ. of South Australia, Adelaide, Australia

Abstract:

Aims: This study seeks to elucidate the causal relationships between alcohol consumption and a variety of diseases, biomarkers, and physiologic measures. Methods: A phenome-wide association study combined with Mendelian randomization analyses (PheWAS-MR) was conducted on 337,463 white-British participants from the UK Biobank. We analyzed PheWAS signals meeting the false-discovery rate threshold using five complementary MR methods, followed-up by sensitivity analyses to assess robustness. We also explored the effects of varying alcohol consumption levels using non-linear MR analysis and examined the causal role of genetic predisposition to alcohol intake on biomarkers and physiologic measures. Results: Genetic predisposition to alcohol consumption was associated with 28 diseases across ten categories, reflecting 22 distinct conditions. In addition to the strong association with ‘alcohol-related disorders’ (OR per 1 gram/day: 7.02, 95% CI: 5.26, 9.37), robust evidence was observed for increased risks of ‘cerebrovascular diseases’ (1.63, 95% CI: 1.20, 2.21), 'essential hypertension' (1.34, 95% CI: 1.07, 1.67), 'electrolyte imbalance' (1.82, 95% CI: 1.34, 2.48), disorder of magnesium (4.39, 95% CI: 2.06, 9.39), 'open wounds of head, neck, and trunk' (2.15, 95% CI: 1.39, 3.33), and 'symptoms involving nervous and musculoskeletal systems' (2.16, 95% CI: 1.60, 2.91). We also observed a suggestive higher risk for 12 additional diseases, majority in mental and digestive- disorders and a lower risk for three others (benign neoplasms of connective and other soft tissue, urinary calculus, and migraines). Non-linear relationships indicated varying risk intensities for seven diseases with increasing alcohol intake, although all exhibited incremental risks or benefits with modest curvature (all Pnon-linearity < 0.05). Additionally, we established a robust causal association between alcohol intake and several biomarkers and physiological measures, including bilirubin, urine sodium, urea, and blood pressure. Conclusion: Our

comprehensive analysis supports a causal role of alcohol on multiple diseases and biomarkers, highlighting the need for further research across diverse populations to better understand alcohol's public health impacts and guide global alcohol policies.

Phenome-Wide Association of *APOE* Alleles in the *All of Us* Research Program

Authors: E. Khajouei[1], V. Ghisays[2], I. Piras[3], K. L. Martinez[1], M. Naymik[3], P. Ngo[1], T. C. Tran[4], J. Denny[4], T. Wheeler[1], M. J. Huentelman[3], E. M. Reiman[2], **J. H. Karnes**[1]; [1]Univ. of Arizona, Tucson, AZ, [2]Banner Alzheimer's Inst., Phoenix, AZ, [3]TGen, Phoenix, AZ, [4]NIH, Bethesda, MD

Abstract:

Background: *APOE* variants have known roles in lipid metabolism, as well as neurodegenerative and cardiovascular disease. However, prior studies of phenotypic associations are largely limited to those with European ancestry and differential risk has not been widely conducted by sex or ancestry. We utilized a phenome-wide association study (PheWAS) approach to explore *APOE*-associated phenotypes in the ancestrally diverse *All of Us* Research Program (*AoU*).
Methods: We determined *APOE* alleles for 181,880 *AoU* participants with whole genome sequencing and electronic health records (EHR) data, representing seven ancestry groups. We tested association of *APOE* allele diplotypes, ordered based on previously published disease risk (ε2/ε2<ε2/ε3<ε3/ε3<ε2/ε4<ε3/ε4<ε4/ε4), with 3492 EHR-derived phenotypes. Analyses were performed with Bonferroni-adjusted alpha in the overall cohort, by ancestry, and by sex with adjustment for 16 principal components, age, sex at birth, and length of available EHR records.
Results: In the overall cohort, PheWAS identified 17 significant associations, including an increased odds of hyperlipidemia (OR 1.15[1.14-1.16] per *APOE* allele group; $P=1.81\times10^{-129}$), dementia (OR 1.32[1.26-1.39]; $P=7.71\times10^{-29}$), and Alzheimer's disease (OR 1.55[1.40-1.70]; $P=5.02\times10^{-19}$), and a reduced odds of fatty liver disease (OR 0.93[0.90-0.95]; $P=1.62\times10^{-9}$) and chronic liver disease. Phenotypic odds ratios were similar across sexes, except for an increased number of cardiovascular associations in males, and decreased odds of Noninflammatory disorders of vulva and perineum in females (OR 0.89[0.84-0.94]; $P=1.07\times10^{-5}$). Significant associations were observed in four of the seven ancestral groups, namely African (AFR), Admixed American (AMR), European (EUR), and Other (OTH), and were largely consistent with the overall analysis. Unique associations included Transient retinal arterial occlusion in EUR, and first degree atrioventricular block in AMR.
Conclusion: We replicated extensive phenotypic associations with *APOE* alleles in a large,

diverse cohort. We also provided evidence of unique phenotypic associations by sex and by ancestry and a comprehensive catalog of *APOE*-associated phenotypes. Limitations of our analysis include relatively low samples sizes in ancestry groups other than EUR, AMR, AFR, and OTH and potential misclassification in clinical diagnoses in *AoU*.

Session 96: Technology for Translation

Location: Room 405

Session Time: Saturday, November 9, 2024, 9:30 am - 10:30 am

A randomized study of a digital genetic health portal (MyCancerGene) for patients who have received germline cancer genetic test results as compared to usual care

Authors: S. Brown[1], S. Howe[1], B. Egleston[2], D. Fetzer[1], S. Domchek[1], L. Fleisher[2], L. I. Wagner[3], K-Y. Wen[4], A. Anantharajah[1], C. Cacioppo[1], J. Cappadocia[1], J. E. Ebrahimzadeh[1], C. Langer[1], E. M. Wood[1], M. Weinberg[1], K. Karpink[1], L. Gutstein[1], A. Tahsin[1], S. Posen[1], E. Selmani[1], A. Bradbury[1]; [1]The Univ. of Pennsylvania, Abramson Cancer Ctr. and Div. of Hematology-Oncology, Philadelphia, PA, [2]Fox Chase Cancer Ctr., Temple Univ., Philadelphia, PA, [3]The Univ. of North Carolina, Gillings Sch. of Global Publ. Hlth., Chapel Hill, NC, [4]Thomas Jefferson Univ., Philadelphia, PA

Abstract:

Background: Given the increasing complexity of genetic testing (e.g. increasing rates of VUS results and changes in risk estimates and medical guidelines) there is an increasing need for longitudinal follow-up. Digital tools could provide a method for updates, education and improved patient outcomes in the setting of complex genetic information. **Methods:** Patients who completed genetic testing were randomized to MyCancerGene access or usual care (UC) and completed assessments of genetic knowledge, distress and behaviors at baseline, 4 weeks and 6 months. We used regressions after multiple imputation to compare differences between arms, using an intention-to-treat analysis. In secondary analyses, we explored differences among those who did and did not access MyCancerGene, and interactions to explore intervention subgroup effects. **Results:** 267 patients were randomized to MyCancerGene access (n=135) or UC (n=132). Participants included 20% men, 16% positive and 23% VUS results; 15% were non-white race and 3.4% Hispanic. 74.1% of participants accessed MyCancerGene within 4 weeks; 83.7% accessed within 6 months. Outcomes were not statistically significant at 4 weeks in the intention-to-treat analysis. At 6 months, there was significant increase in knowledge in the intervention arm (+0.22 v -0.37 UC, p=0.057). There was significant decrease in negative responses to testing for those who accessed MyCancerGene within 4 weeks (9.58 without log-in v 2.66 with log-in; p=0.005). Among those with low literacy, MyCancerGene was associated with 6-month declines in disease specific distress (interaction p<0.01) and general anxiety (interaction p<0.01), and

increases in genetic knowledge (interaction p=0.02), while the intervention had opposite effects in those with high health literacy. In those without college degrees, MyCancerGene increased disease specific distress at 6 months, but decreased distress for those with college degrees (interaction p=0.01). Among those with positive results compared to others, MyCancerGene increased 6-month depression (p=0.001) and negative responses to genetic testing (p<0.01). **Conclusions:** The MyCancerGene intervention increases knowledge at 6 months and reduces negative responses to testing for those who access the intervention, providing a potential tool to improve longitudinal outcomes to genetic testing. MyCancerGene may be of greater benefit for lower literacy patients, and a better understanding of the impact of small increases in distress for some subgroups and longitudinal data on the impact on health behaviors will inform the potential to improve longitudinal clinical outcomes after genetic testing.

GenAI-powered approaches in advancing genetic testing education and communication: an exploratory study in Pharmacogenomics

Authors: M. Murugan[1], E. Venner[2], C. Ballantyne[2], K. M. Robinson[3], J. C. Coons[3], L. Wang[1], G. Metcalf[2], A. Anderson[1], P. Empey[3], R. Gibbs[1], B. Yuan[2]; [1]Baylor Coll. Med., Houston, TX, [2]Baylor Coll. of Med., Houston, TX, [3]Univ. of Pittsburgh, Pittsburgh, PA

Abstract:

Background: A shortage of genetic professionals and limited genetic knowledge among health care providers (HCPs) create significant barriers to genetic services, especially in underserved settings. These barriers lead to delayed access to genomic testing, inadequate interpretation, and ineffective communication of results. GenAI, particularly large language models, show promise in bridging these gaps. This study explores the feasibility of GenAI to improve education and communication of genetic testing results. **Methods**: We conducted a proof-of-concept study to evaluate the feasibility of a GenAI-based pharmacogenomic AI assistant (PGx AI) to bridge knowledge gaps in the return of results to HCPs and patients, focusing on the SLCO1B1, ABCG2, and CYP2C9 genes, and statin therapies. The PGx AI was developed with OpenAI's GPT-4 using advanced techniques including retrieval-augmented generation, prompt engineering, guardrails, and a curated knowledge base from the Clinical Pharmacogenetics Implementation Consortium to generate responses beyond GPT-4's training data limitations. The PGx AI was evaluated on a specialized question catalog, with responses benchmarked against ChatGPT 3.5. Domain experts evaluated both sets of responses on key criteria such as accuracy, relevancy, risk, language, bias, citation, and

hallucination. **Results**: For provider-focused queries (n=47), the PGx AI achieved 85% overall performance across all criteria, significantly higher than ChatGPT 3.5's 69% (P-value: 8.11×10^{-20}) with a notable 11% reduction in hallucinations than ChatGPT. For patient-focused queries (n=33), the PGx AI performed well, achieving 82% performance compared to ChatGPT 3.5's 78% (P-value: .000643). Despite strong performance, expert feedback identified areas for improvement, such as accuracy, relevancy, reading levels, and interpretation of domain-specific terminology. **Conclusion**: The PGx AI, utilizing innovative GenAI approaches, demonstrated strong performance in informing genetic testing results, though challenges remain. We plan to fine-tune an LLM with relevant guidelines and standards, expand this study to a larger PGx panel with variable gene-drug interactions, and evaluate it as part of Baylor College of Medicine's Learning Health System's IMAGINE program. This will help improve and gain deeper insights into the performance, uptake, utility, and limitations of the PGx AI. While not yet ready for widespread clinical use, this work highlights the potential of GenAI to enhance education and accessibility in genomic medicine, warranting further evaluation in broader genetic domains.

Machine learning predictions to shorten diagnostic odysseys in Level IV NICUs

Authors: B. Chaudhari, A. Antoniou; Nationwide Children's Hosp., Columbus, OH

Abstract:

Background Genetic diseases are highly prevalent in Level IV Neonatal Intensive Care Units (NICUs). Despite the availability of rapid genomic sequencing (rGS), most genetic diagnoses in this population are made post-discharge and associated with long diagnostic odysseys (time to diagnosis or genomic testing).
Objective Train and test a machine learning (ML) classifier to predict subsequent diagnostic odyssey. Measure the potential costs (extra rGS) and benefits (shortened diagnostic odyssey, fewer non-rGS genetic tests) of following testing recommendations made by ML in a Level IV NICU validation cohort.
Methods
Subjects were 27,024 patients born 2010-2019, admitted to a Level III/IV NICU (training) and 2,241 patients born 2020-2021 and admitted to a Level IV NICU (validation). Of note, rGS has been clinically available at the Level IV NICU, without restriction, since 2020. Subjects were labeled "positive" for training if they began a diagnostic odyssey by 18 months of age, even if the odyssey began post-discharge. Features for ML included gestational age, birth weight Z-score, Human Phenotype Ontology (HPO) terms extracted by ClinPhen from 2,322,303 clinical notes, as well as data from infectious disease testing

through ICU day 7. Using 5-fold cross validation, we evaluated multiple models. The model with fewest HPO features and fold-mean average precision (AP) score within 0.1 of the maximum was selected for validation. F1-maximizing decision boundary was fixed for simulations.
We simulated diagnostic odysseys for the validation cohort under a testing policy which initiated genetic testing when positive ML predictions were made over the first 7 days in the Level IV NICU, but did not change the actual genetic tests ordered (Sim 1) as well making the first genetic test ordered rGS (Sim 2). These were compared to real diagnostic odysseys as experienced.
Results The selected XGBoost model was trained on 89 features and had a C-statistic of 0.88 and AP of 0.75. Under Sim 2, 512 rGS were ordered, but 70 karyotypes, 169 microarrays, and 170 NGS panels were avoided. 167 subjects who, in reality, did not receive any genetic testing received rGS. The actual median diagnostic odyssey of 21 days dropped to 12 days in Sim 1 and 5 days in Sim 2 (pairwise p < 0.05). Conclusion(s) ML predictions in the first week of admission may radically shorten diagnostic odysseys for patients admitted to Level IV NICUs. The retrospective nature of this analysis precludes assessment of the benefits or burdens of genetic testing recommendations made for the 167 subjects who, in reality, received no genetic testing.

The All of Us Research Program data release 2024 (CDR v8): Powering genomic research through All of Us

Authors: A. Musick[1], E. Banks[2], M. Basford[3], E. Boerwinkle[4], M. Cicek[5], K. Doheny[6], T. Dutka[7], E. Eichler[8], S. Gabriel[9], R. Gibbs[10], P. Harris[3], D. Glazer[11], G. Jarvik[12], H. Rehm[13], D. Roden[14], S. Topper[15], L. Lichtenstein[2], The All of Us Research Program Genomics Investigators; [1]NIH - All of Us Program, Bethesda, MD, [2]Broad Inst., Boston, MA, [3]Vanderbilt Univ. Med. Ctr., Nashville, TN, [4]Univ. of Texas Hlth.Sci. Ctr. at Houston, Houston, TX, [5]Mayo Clinic, Rochester, MN, [6]Ctr. for Inherited Disease Res. (CIDR), JHU, Baltimore, MD, [7]Natl. Inst. Hlth., Bethesda, MD, [8]Univ of Washington, Seattle, WA, [9]Broad Inst. of MIT and Harvard, Cambridge, MA, [10]Baylor Coll. Med., Houston, TX, [11]Google, Mountain View, CA, [12]Univ Washington Med Ctr., Seattle, WA, [13]Massachusetts Gen. Hosp., Boston, MA, [14]VUMC, Nashville, TN, [15]Color Hlth., Berkeley, CA

Abstract:

The NIH's *All of Us* Research Program aims to enroll one million or more participants who reflect the diversity of the United States and create one of the world's largest and most diverse research resources. The program is well-positioned to drive precision medicine by

combining data from surveys, electronic health records, physical measurements, genomics, and digital health technologies. Here we describe the program's most recent data release which overall reflects ~50% increase in participants with data available (N = 413,457 → 633,547), including ~415k whole genome sequences (~98k with structural variant (SV) calls), >2.8k long-read sequences, as well as auxiliary data and variant annotations. Genomic data are generated at one of four *All of Us* Genome Centers using identical library construction protocols, sequencers, software, and software configuration. The *All of Us* Data and Research Center performed additional QC, joint calling, and data curation. The release also provides data for common analysis use cases, including inferred continental ancestries and results of principal components analysis, variant annotations, relatedness and maximal set of unrelated individuals, and multiple variant file formats (e.g., plinkbed, bgen, Hail), and commonly used subsets of data, e.g., common variants, exomes, ClinVar variants. *All of Us* identified >1.25 billion genetic variants, including >300 million previously unreported genetic variants, >4 million in coding regions. Structural variant calls in ~98k participants identified >1.5M SVs. Long-read data, including SNV, in/dels and SVs, are made available on two references, hg38 and T2T-CHM13, as well as de novo assembly. Principal component analysis inferred almost half of participants are from non-European ancestries. Summary-level data are publicly available on the AoU data browser, including aggregated genetic variant counts. In summary, the *All of Us* Research Program allows approved researchers worldwide to access the data through a passport model with agreements that cover all researchers at an institution. Registered researchers run analyses in the *All of Us* Researcher Workbench, a secure, cloud-based Trusted Research Environment. There is no charge for data access, and researchers are given initial credits for cloud computing costs. Currently, ~800 domestic and international institutions have approved data use agreements in place, and >11k researchers have registered for access. The *All of Us* Research Program is enabling novel health research and medical breakthroughs due to the unprecedented scale and diversity of the participants and data sources.