Advances in ADHD genetics and genome sequencing technology

Daniel Howrigan, PhD
Senior Group Leader - Neale lab

STANLEY CENTER
FOR PSYCHIATRIC RESEARCH
AT BROAD INSTITUTE

MASSACHUSETTS
GENERAL HOSPITAL

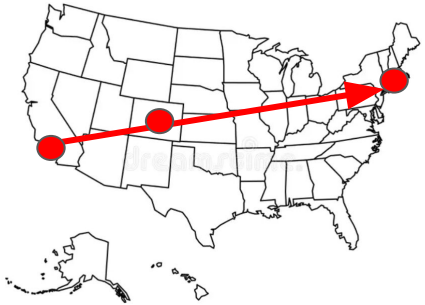
Today’s presentation

Part 1: Advances in ADHD genetics from PGC to iPSYCH

Part 2: Development of the Blended Genome Exome sequencing technology

About me

- Grew up in Southern California
- BA in Anthropology at UC Santa Barbara, California
- PhD in Psychology in Boulder, Colorado
- Past 13 years in Boston, Massachusetts
  - 4 years as a Postdoc in the Neale lab
  - 9 years as a Group Leader in the Neale lab



Winter surfing in Cape Cod, MA

METHODOLOGY ARTICLE Open Access

Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms

Daniel P Howrigan[1]

Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects

CNV and Schizophrenia Working Gro

Exome sequencing in schizophrenia-affected parent–offspring trios reveals risk conferred by protein-coding de novo mutations

Analytic and Translational Genetics Unit

PGC
Psychiatric Genomics Consortium

iPSYCH

hail

gnomAD

SCHEMA browser

Epi25 WES Browser

BipEx: Bipolar Exomes Browser

Autism Sequencing Consortium exome analysis

genebass
gene-based association summary statistics

Pan-UK Biobank
Pan-ancestry genetic analysis of the UK Biobank

All of Us
RESEARCH PROGRAM

FINNGEN

UKB SNP-Heritability Browser
Results from the Neale Lab

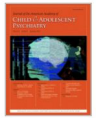
Global Biobank Meta-analysis Initiative (GBMI) consortium

Part 1: Back in 2010…

ELSEVIER

Journal of the American Academy of Child & Adolescent Psychiatry

Volume 49, Issue 9, September 2010, Pages 906-920



New research

Case-Control Genome-Wide Association Study of Attention-Deficit/Hyperactivity Disorder

Benjamin M. Neale Ph.D. [a c], Sarah Medland Ph.D. [b c], Stephan Ripke M.D. [a c], Richard J.L. Anney Ph.D. [d], Philip Asherson M.R.C.Psych., Ph.D. [e], Jan Buitelaar M.D. [f], Barbara Franke Ph.D. [f], Michael Gill M.B., Bch, BAO, M.D., MRCPsych, F.T.C.D. [d], Lindsey Kent M.D., Ph.D. [g], Peter Holmans Ph.D. [e], Frank Middleton Ph.D. [h], Anita Thapar M.D. [i], Klaus-Peter Lesch M.D. [l], Stephen V. Faraone Ph.D. [h], Mark Daly Ph.D. [a c], Thuy Trang Nguyen Dipl. Math. oec [j], Helmut Schäfer Ph.D. [j], Hans-Christoph Steinhausen M.D., Ph.D., D.M.Sc. [k], Andreas Reif M.D. [l], Tobias J. Renner M.D. [l]… Joseph Biederman M.D. [r s]

Objective

Although twin and family studies have shown attention-deficit/hyperactivity disorder (ADHD) to be highly heritable, genetic variants influencing the trait at a genome-wide significant level have yet to be identified. Thus additional genomewide association studies (GWAS) are needed.

Method

We used case-control analyses of 896 cases with *DSM-IV* ADHD genotyped using the Affymetrix 5.0 array and 2,455 repository controls screened for psychotic and bipolar symptoms genotyped using Affymetrix 6.0 arrays. A consensus SNP set was imputed using BEAGLE 3.0, resulting in an analysis dataset of 1,033,244 SNPs. Data were analyzed using a generalized linear model.

Results

No genome-wide significant associations were found. The most significant results implicated the following genes: *PRKG1, FLNC, TCERG1L, PPM1H, NXPH1, PPM1H, CDH13, HK1*, and *HKDC1*.

Back in 2010…

Journal of the American Academy of Child & Adolescent Psychiatry
Volume 49, Issue 9, September 2010, Pages 884-897

New research

Meta-Analysis of Genome-Wide Association Studies of Attention-Deficit/Hyperactivity Disorder

Benjamin M. Neale Ph.D. [a b], Sarah E. Medland Ph.D. [c d], Stephan Ripke M.D. [a b], Philip Asherson M.R.C.Psych., Ph.D. [e], Barbara Franke Ph.D. [f], Klaus-Peter Lesch M.D. [m], Stephen V. Faraone Ph.D. [g], Thuy Trang Nguyen Dipl. Math. oec. [h], Helmut Schäfer Ph.D. [h], Peter Holmans Ph.D. [i], Mark Daly Ph.D. [a d], Hans-Christoph Steinhausen M.D., Ph.D., D.M.Sc. [j k l], Christine Freitag M.D., M.A. [n], Andreas Reif M.D. [m], Tobias J. Renner M.D. [m], Marcel Romanos M.D. [m], Jasmin Romanos M.D. [m], Susanne Walitza M.D. [j m], Andreas Warnke M.D., Ph.D. [m], Jobst Meyer Ph.D. [o]...Stan Nelson M.D. [aj]

. . . . .

Hakon Hakonarson M.D., Ph.D. [x y], Josephine Elia M.D. [x], Alexandre Todorov Ph.D. [z], Ana Miranda M.D. [aa], Fernando Mulas M.D., Ph.D. [ab], Richard P. Ebstein Ph.D. [ac], Aribert Rothenberger M.D., Ph.D. [ad], Tobias Banaschewski M.D., Ph.D. [n], Robert D. Oades Ph.D. [ae], Edmund Sonuga-Barke Ph.D. [e af ag], James McGough M.D. [w], Laura Nisenbaum Ph.D. [ah], Frank Middleton Ph.D. [f], Xiaolan Hu Ph.D. [ai], Stan Nelson M.D. [aj],
Psychiatric GWAS Consortium: ADHD Subgroup

Method

We used data from four projects: a) the Children's Hospital of Philadelphia (CHOP); b) phase I of the International Multicenter ADHD Genetics project (IMAGE); c) phase II of IMAGE (IMAGE II); and d) the Pfizer-funded study from the University of California, Los Angeles, Washington University, and Massachusetts General Hospital (PUWMa). The final sample size consisted of 2,064 trios, 896 cases, and 2,455 controls. For each study, we imputed HapMap single nucleotide polymorphisms, computed association test statistics and transformed them to z-scores, and then combined weighted z-scores in a meta-analysis.

Results

No genome-wide significant associations were found, although an analysis of candidate genes suggests that they may be involved in the disorder.

Conclusions

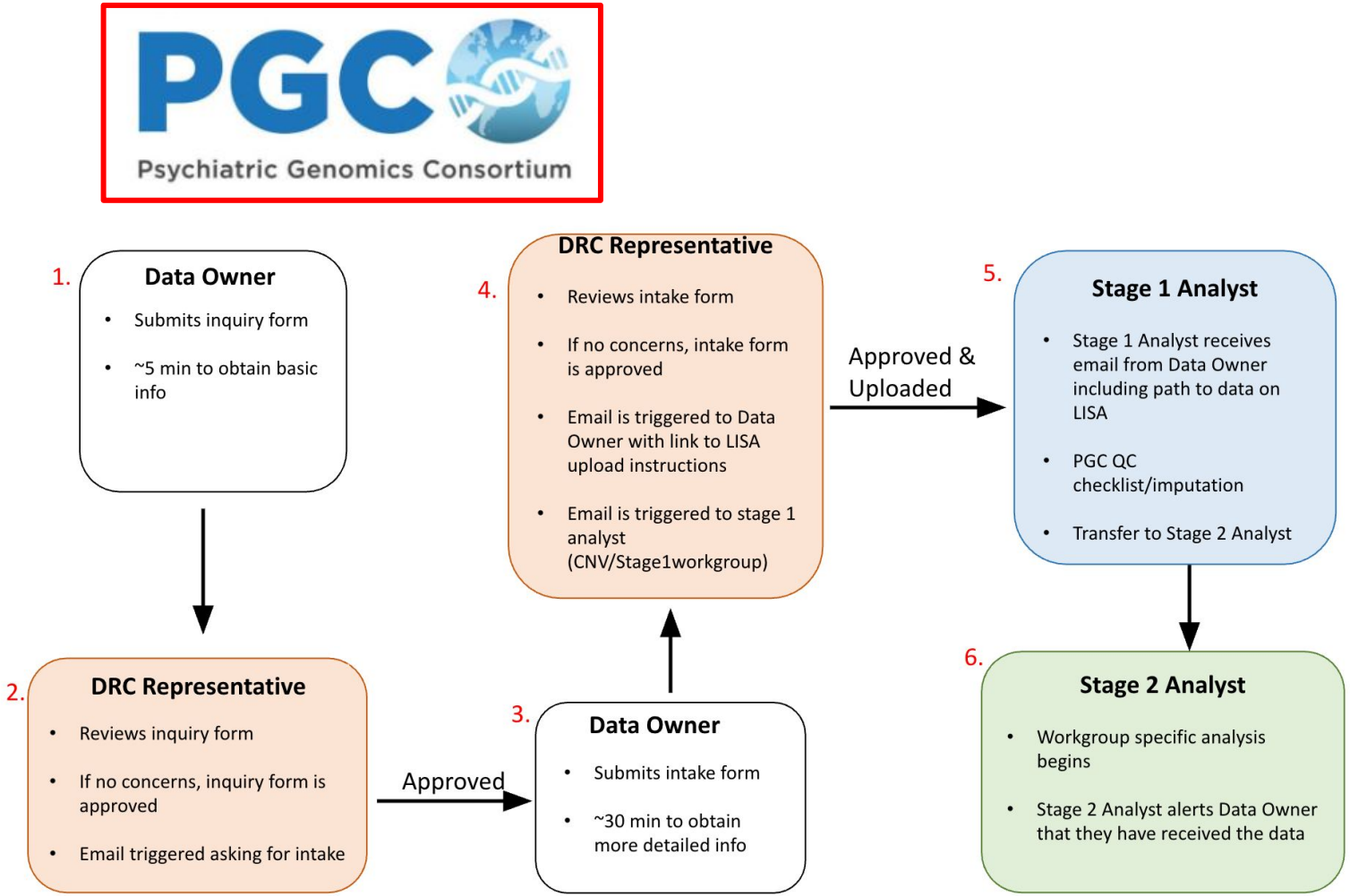
Given that ADHD is a highly heritable disorder, our negative results suggest that the effects of common ADHD risk variants must, individually, be very small or that other types of variants, e.g., rare ones, account for much of the disorder's heritability.

PGC

Psychiatric Genomics Consortium

PGC WORKING GROUPS

- Alzheimer's Disease
- Attention Deficit Hyperactivity Disorder
- Anxiety Disorders
- Autism Spectrum Disorder
- Bipolar Disorder
- Copy Number Variations
- Cross Disorder Analyses
- Cross-Population Analyses
- Eating Disorders
- Functional Genomics
- Major Depressive Disorder
- Obsessive-Compulsive Disorder and Tourette Syndrome
- Post-Traumatic Stress Disorder
- Schizophrenia
- Substance Use Disorders
- Suicide Working Group


PGC
Psychiatric Genomics Consortium
1.
Data Owner
Submits inquiry form
~5 min to obtain basic info
2.
DRC Representative
Reviews inquiry form
If no concerns, inquiry form is approved
Email triggered asking for intake
Approved
3.
Data Owner
Submits intake form
~30 min to obtain more detailed info
4.
DRC Representative
Reviews intake form
If no concerns, intake form is approved
Email is triggered to Data Owner with link to LISA upload instructions
Email is triggered to stage 1 analyst (CNV/Stage1workgroup)
Approved & Uploaded
5.
Stage 1 Analyst
Stage 1 Analyst receives email from Data Owner including path to data on LISA
PGC QC checklist/imputation
Transfer to Stage 2 Analyst
6.
Stage 2 Analyst
Workgroup specific analysis begins
Stage 2 Analyst alerts Data Owner that they have received the data

iPSYCH

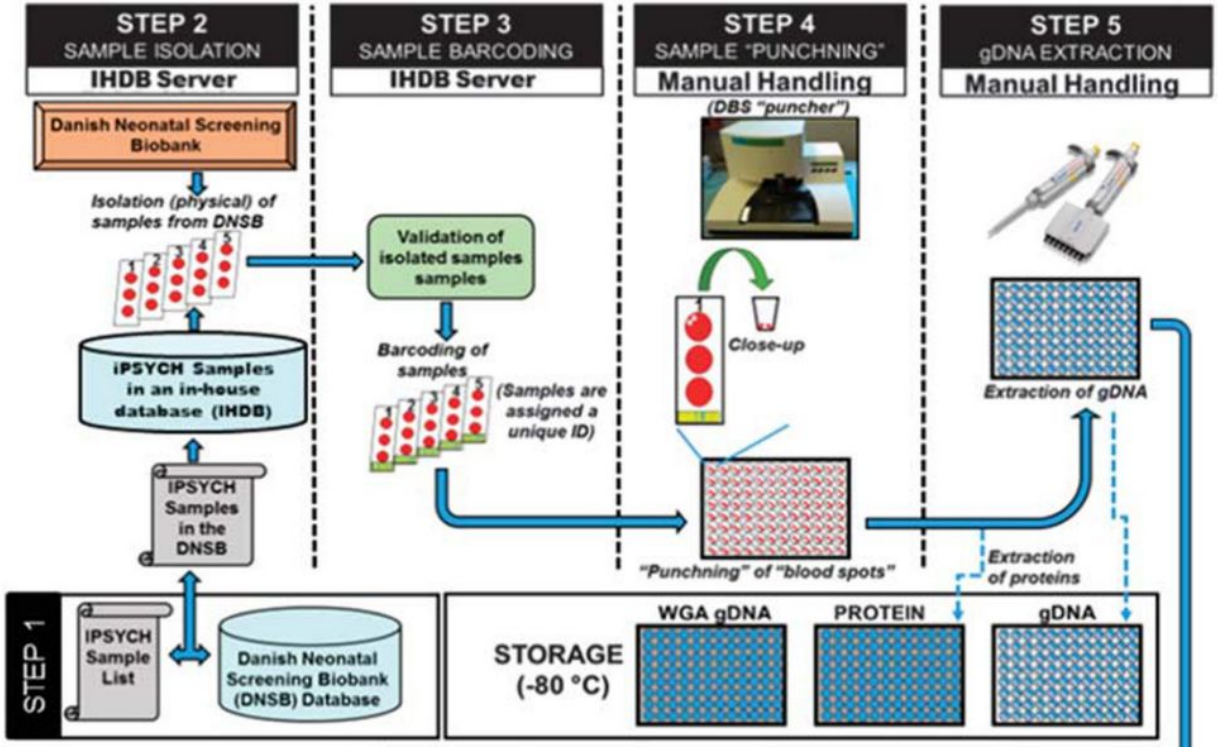
The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders

CB Pedersen[1,2,3,15], J Bybjerg-Grauholm[1,4,15], MG Pedersen[1,2,3], J Grove[1,5,6], E Agerbo[1,2,3], M Bækvad-Hansen[1,4], JB Poulsen[1,4], CS Hansen[1,4], JJ McGrath[1,2,7,8], TD Als[1,5], JI Goldstein[9,10,11], BM Neale[9,10,11], MJ Daly[9,10,11], DM Hougaard[1,4,16], O Mors[1,12,16], M Nordentoft[1,13,16], AD Børglum[1,5,16], T Werge[1,14,16] and PB Mortensen[1,2,3,5,16]

The Integrative Psychiatric Research (iPSYCH) consortium has established a large Danish population-based Case–Cohort sample (iPSYCH2012) aimed at unravelling the genetic and environmental architecture of severe mental disorders. The iPSYCH2012 sample is nested within the entire Danish population born between 1981 and 2005, including 1 472 762 persons. This paper introduces the iPSYCH2012 sample and outlines key future research directions. Cases were identified as persons with schizophrenia ($N = 3540$), autism ($N = 16\,146$), attention-deficit/hyperactivity disorder ($N = 18\,726$) and affective disorder ($N = 26\,380$), of which 1928 had bipolar affective disorder. Controls were randomly sampled individuals ($N = 30\,000$). Within the sample of 86 189 individuals, a total of 57 377 individuals had at least one major mental disorder. DNA was extracted from the neonatal dried blood spot samples obtained from the Danish Neonatal Screening Biobank and genotyped using the Illumina PsychChip. Genotyping was successful for 90% of the sample. The assessments of exome sequencing, methylation profiling, metabolome profiling, vitamin-D, inflammatory and neurotrophic factors are in progress. For each individual, the iPSYCH2012 sample also includes longitudinal information on health, prescribed medicine, social and socioeconomic information, and analogous information among relatives. To the best of our knowledge, the iPSYCH2012 sample is the largest and most comprehensive data source for the combined study of genetic and environmental aetiologies of severe mental disorders.

Molecular Psychiatry (2018) **23,** 6–14; doi:10.1038/mp.2017.196; published online 19 September 2017

iPSYCH


STEP 2
SAMPLE ISOLATION
IHDB Server
Danish Neonatal Screening Biobank
Isolation (physical) of samples from DNSB
iPSYCH Samples in an in-house database (IHDB)
iPSYCH Samples in the DNSB
STEP 3
SAMPLE BARCODING
IHDB Server
Validation of isolated samples samples
Barcoding of samples
(Samples are assigned a unique ID)
STEP 4
SAMPLE "PUNCHNING"
Manual Handling
(DBS "puncher")
Close-up
"Punchning" of "blood spots"
STEP 5
gDNA EXTRACTION
Manual Handling
Extraction of gDNA
Extraction of proteins
STEP 1
iPSYCH Sample List
Danish Neonatal Screening Biobank (DNSB) Database
STORAGE (-80 °C)
WGA gDNA
PROTEIN
gDNA

Article | Published: 26 November 2018

Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder

Ditte Demontis, Raymond K. Walters, Joanna Martin, Manuel Mattheisen, Thomas D. Als, Esben Agerbo, Gísli Baldursson, Rich

- 20.2k cases
- 35k controls
- 12 GWAS sig. loci

Article | Published: 26 January 2023

Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains

Ditte Demontis, G. Bragi Walters, Georgios Athanasiadis, Raymond Walters, Karen Therrien, Trine Tollerup Nielsen, Leila Farajzadeh,

- 38.7k cases
- 184k controls
- 27 GWAS sig. loci

From 2010 to 2025…

Genome-wide associations with Attention-Deficit/Hyperactivity Disorder in >700,000 individuals

Raymond Walters
On behalf of the ADHD Working Group of the Psychiatric Genomics Consortium
And Collaborators

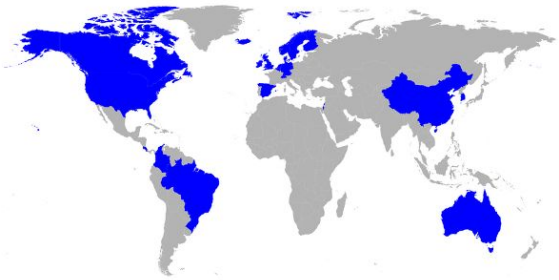
World Congress of Psychiatric Genetics
October 23, 2025

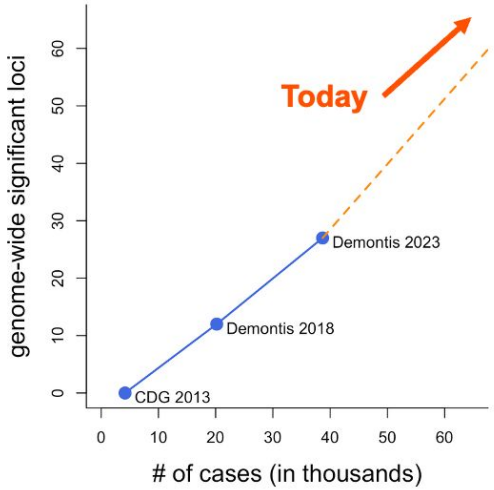
GWAS discovery in 77 cohorts quadruples sample size and expands mix of study designs

- Updated discovery meta-analysis includes 77 cohorts
  - 33 case/control cohorts with genotypes in PGC
  - 27 family-based cohorts with genotypes in PGC
  - 17 cohorts with external summary statistics

- (Slowly) expanding ancestral diversity
  - 3,120 AFR-like cases (5 cohorts)
  - 1,917 AMR-like cases (8 cohorts)
  - 1,497 EAS-like cases (3 cohorts)

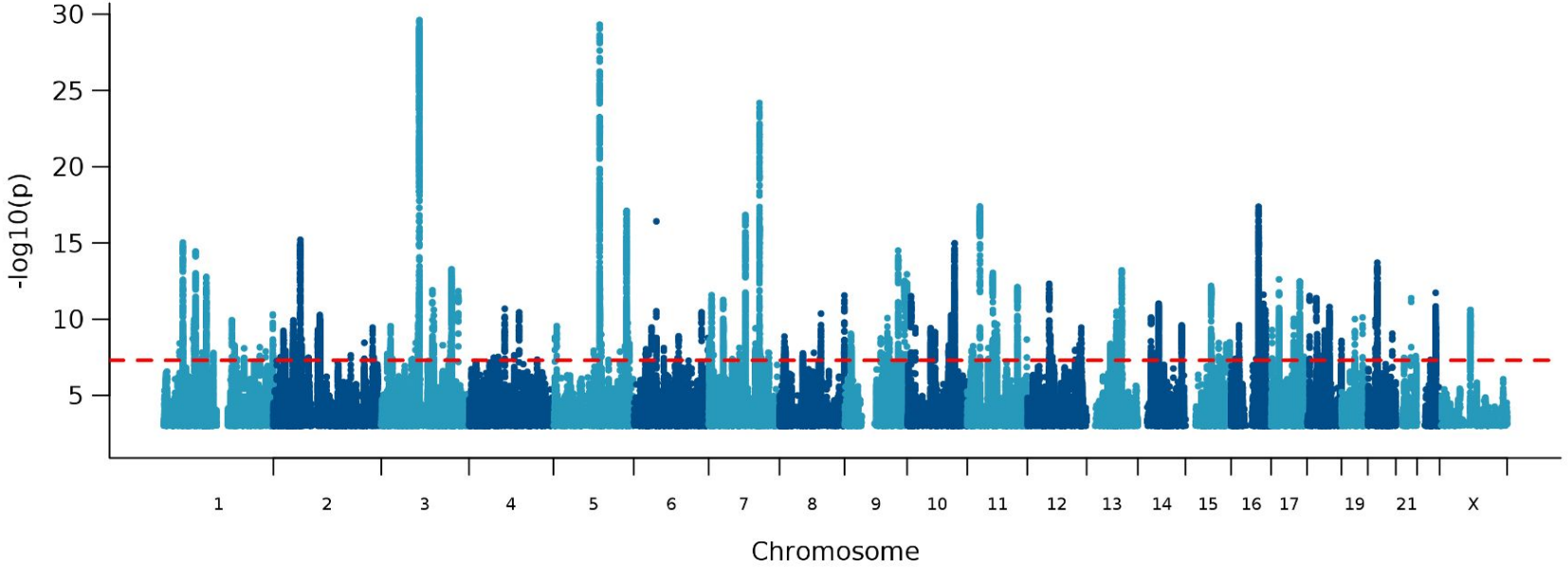
- **Total sample size increase from Demontis et al. 2023:**
  - **38,691 -> 170,683 cases**
  - **186,843 -> 1,528,137 controls**





Updated genome-wide meta-analysis of ADHD identifies 150 newly significant loci

- 170,683 cases, 1,528,137 controls
- **178 genome-wide significant loci**
  - Includes 13 only significant in multi-ancestry meta-analysis

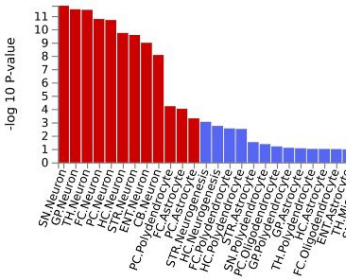


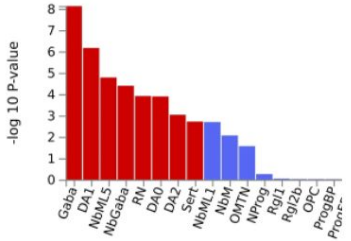
ADHD GWAS results are generally consistent across subgroups and cohort designs; some continental differences

- **Overall:** No loci with genome-wide significant heterogeneity across cohorts
- **By sex:** females vs. males: rg>1, se=.12
- **By ancestry:** Concordant effect sizes at top loci
  - Insufficient power for rg
- **By design:**
  - Registry vs. interview-based phenotyping: rg=.95, se=.08
  - Family vs. case/control or registry based: rg=.82, se=.18
- **By data freeze:** new cohorts vs. Demontis 2023: rg=0.97, se=.05
- **By continent:** European vs. US cohorts: rg=0.75, se=0.02

Increasing resolution of genome-wide enrichments highlights prenatal period, neuron development

- Strongly enriched across all brain tissues
- Remains most correlated with gene expression in early/mid prenatal period
- Enrichments highlight neurons, but not strongly specific to location or type

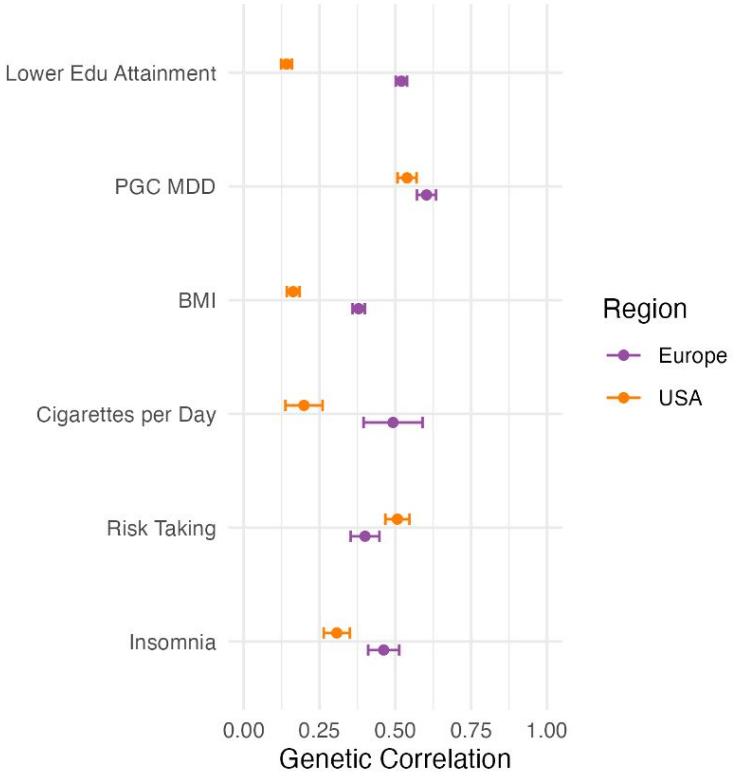




MAGMA gene property analysis with single cell RNAseq in adult mouse brain and human embryonic midbrain

European vs. US cohorts have differing correlation with educational attainment

- Pattern of genetic correlations differs between European and US ADHD cohorts
  - US weaker relationship with Education
- Some signs of differences at individual loci
  - Sensitive to individual cohorts



pTDT distinguishes genetic nurture from transmitted risk

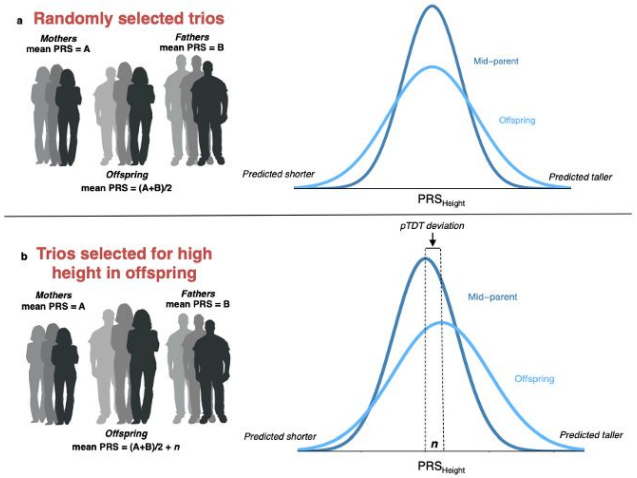
- Polygenic transmission disequilibrium test
  - Proband ascertainment -> higher polygenic risk than parental average
  - No impact if risk unrelated to transmitted genotype (i.e. genetic nurture)

- Evaluate in 2487 trios from 10 cohorts whether PGS are over-transmitted within families

Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders

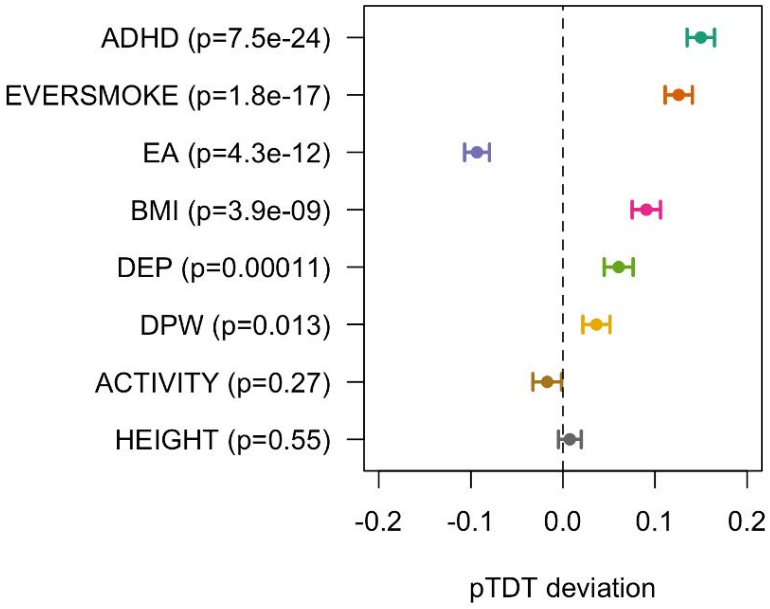
Daniel J Weiner, Emilie M Wigdor, Stephan Ripke, Raymond K Walters, Jack A Kosmicki, Jakob Grove, Kaitlin E Samocha, Jacqueline I Goldstein, Aysu Okbay, Jonas Bybjerg-Grauholm, Thomas Werge, David M Hougaard, Jacob Taylor, iPSYCH-Broad Autism Group, Psychiatric Genomics Consortium Autism Group, David Skuse, Bernie Devlin, Richard Anney, Stephan J Sanders, Somer Bishop, Preben Bo Mortensen, Anders D Børglum, George Davey Smith, Mark J Daly & Elise B Robinson

Nature Genetics **49**, 978–985 (2017) | Cite this article

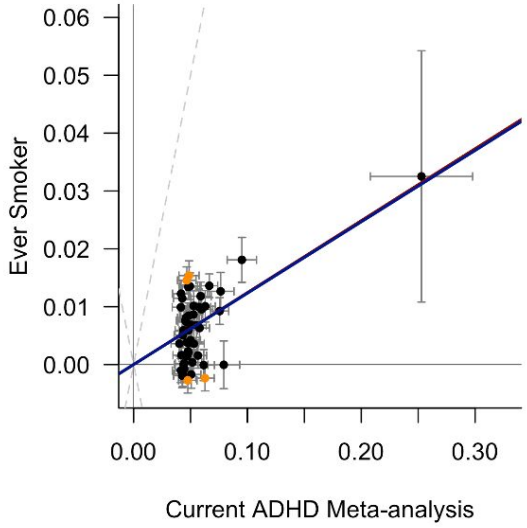


Clear within-family signal for transmission of polygenic risk for ADHD, some correlated traits

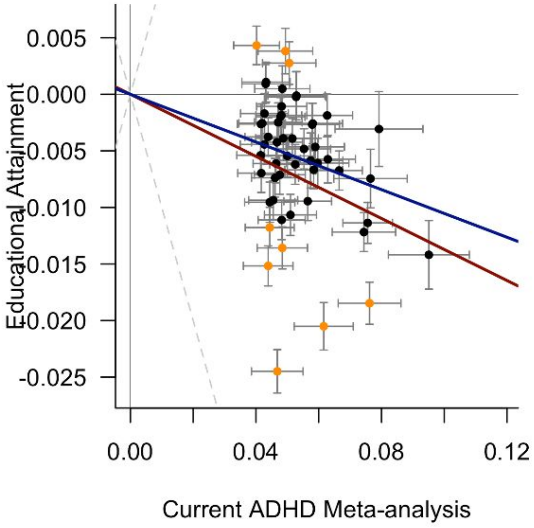
- ADHD trio probands inherit polygenic risk 0.15 SDs higher than chance
- Significant pTDT for smoking, educational attainment, BMI, depressive symptoms
- No significant over-transmission for alcohol consumption, physical activity



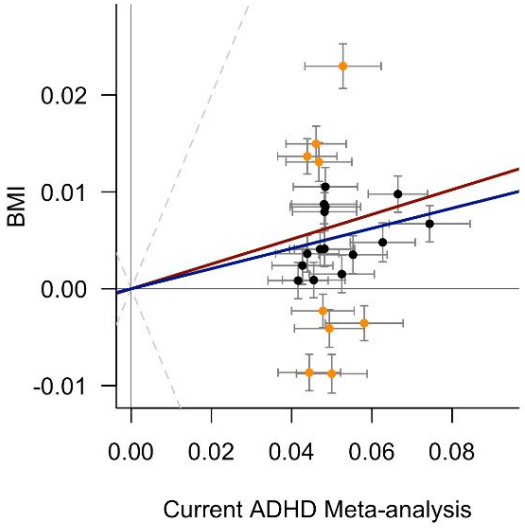
ADHD top loci have heterogeneous relationships with smoking, educational attainment, BMI



Global heterogeneity
p=1.7e-20



Global heterogeneity
p=1.7e-44



Global heterogeneity
p=4.2e-45

SCOUTJOY: Elliott et al., Nat Genet. 2024

Thanks to all the analysts making these updates possible!

Stage 1 Analysis Team

Daniel Howrigan

Danfeng Chen

Nik Baya

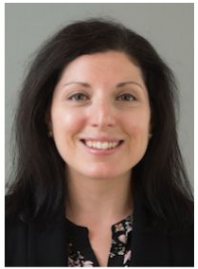
Andrew Marin

Wenhan Lu

Data Receiving

Tetyana Zayats

Project Management



Felecia Cerrato

iPSYCH/Stage 2

Ditte Demontis

Yorgos Athanasiadis

RICOPILI

Rapid Imputation and COmputational PIpeLIne for Genome-Wide Association Studies

QC	• Platform Matching • Renaming files and IDs • Filtering of SNPs and Samples
PCA	• Eigenstrat with population loadings • Determine which PCs are significant
Imputation	• 1KG reference panel based on input population
Analysis	• Logistic regression with significant PCs in PLINK • Meta-analysis using METAL

Document tabs

BHRC1 - Brazil GSA QC Analysis Report

Last Changes: June 30th, 2020

Andrew Marin, Stage 1 Analyst, amarin@broadinstitute.org

Stanley Center for Psychiatric Disease, Broad Institute, Cambridge, MA, USA

Primary Investigator: Giovanni Salum et al.

Original Filename:
/home/pgcdac/DWFV2CJb8Piv_0116_pgc_data/pgcdrc/add/incoming_datasets/bhrc1/

Primary Output Location:
/home/pgcdac/DWFV2CJb8Piv_0116_pgc_data/pgcdrc/add/working/wave2/bhrc1/primary_output

Sample Breakdown:
Dataset Abbreviation: bhrc1 Sample Ascertainment: Trios

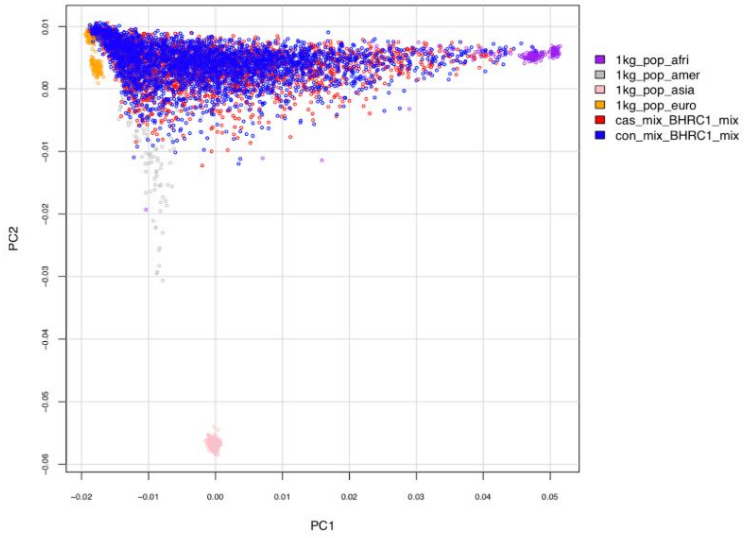
PRE-QC
Expected Continental Ancestry: Brazilian/ Mixed
Genotype Platform Used: GSAMD-24v1-0_20011747_A1.1.3
QC Analysis pipeline: Ricopili
Sample Size: 5364

Multiple Disease Types from Phenotype file:

- ANX = Ever diagnosed with any anxiety disorders (Proband Cases = 431; Parent cases = 524)

- BD = Ever diagnosed with mania (Probands Cases = 13; Parents cases = 147)

PCA With Reference (PCA Projection)



What about rare variants and ADHD?

iPSYCH

Article

Rare genetic variants confer a high risk of ADHD and implicate neuronal biology

https://doi.org/10.1038/s41586-025-09702-8

Received: 2 July 2024

Accepted: 2 October 2025

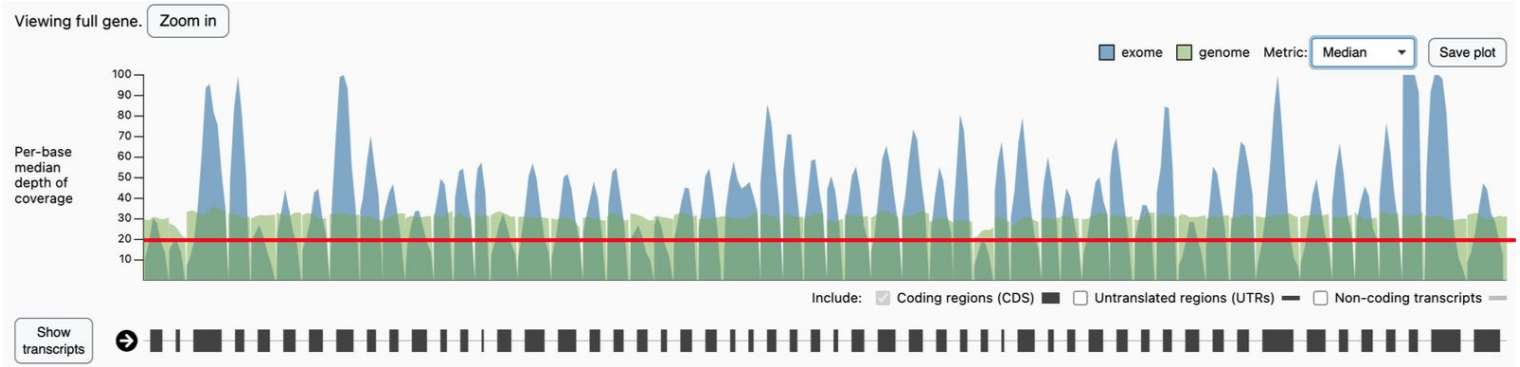
Published online: 12 November 2025

Open access

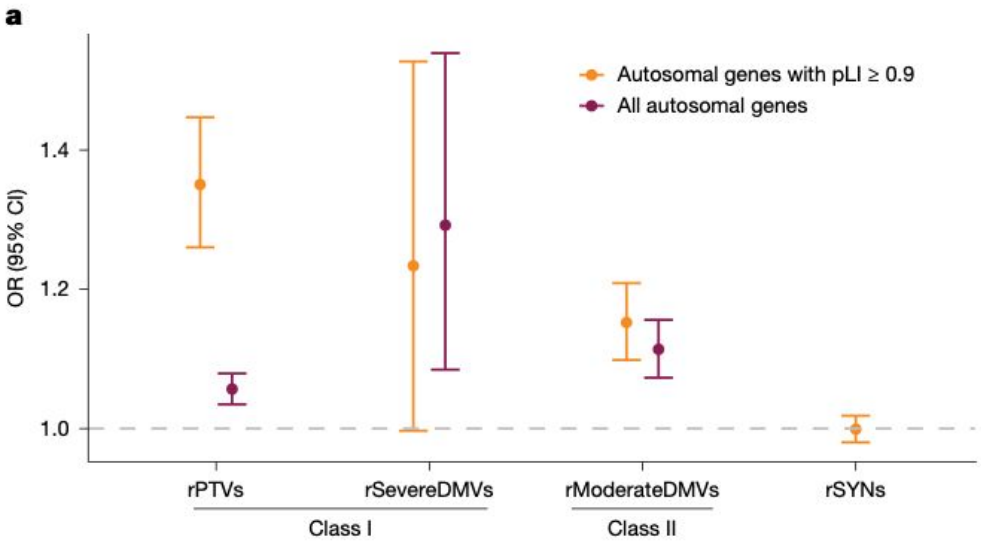
Check for updates

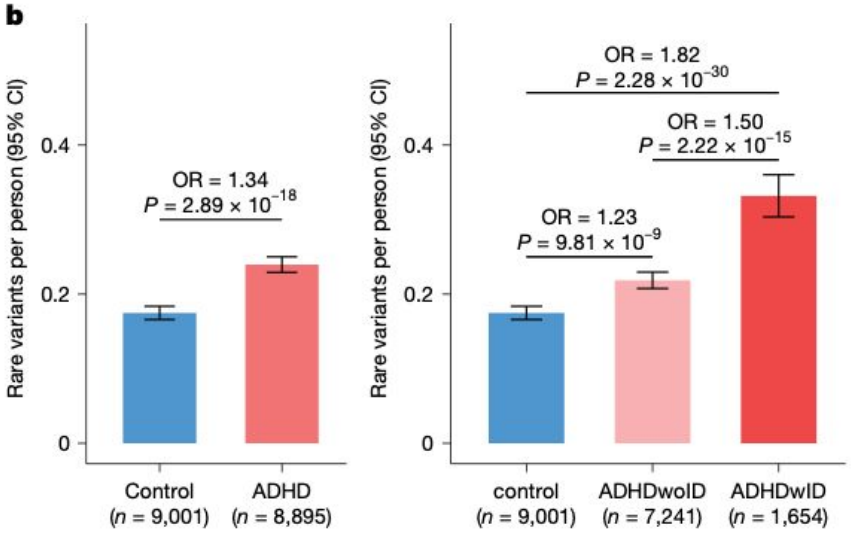
Ditte Demontis[1,2,3,4,28], Jinjie Duan[1,2,3,28], Yu-Han H. Hsu[4,5], Greta Pintacuda[4,5], Jakob Grove[1,2,3], Trine Tollerup Nielsen[1,2,3], Janne Thirstrup[1,2,3], Makayla Martorana[4,5], Travis Botts[4,5], F. Kyle Satterstrom[5,6], Jonas Bybjerg-Grauholm[2,7], Jason H. Y. Tsai[1,2,3], Simon Glerup[1], Martine Hoogman[8,9,10], Jan Buitelaar[8,9], Marieke Klein[8,9], Georg C. Ziegler[11], Christian Jacob[12], Oliver Grimm[13], Maximilian Bayas[13], Nene F. Kobayashi[13], Sarah Kittel-Schneider[14,15], Klaus-Peter Lesch[16,17,18], Barbara Franke[8,9,19], Andreas Reif[13,20], Esben Agerbo[2,21,22], Thomas Werge[2,23], Merete Nordentoft[2,24], Ole Mors[2,25], Preben Bo Mortensen[2,21,22], Kasper Lage[4,5,23], Mark J. Daly[5,6,26,27], Benjamin M. Neale[5,6] & Anders D. Børglum[1,2,3]

- 8.9k cases
- 54k controls
- 3 exome-wide sig. genes

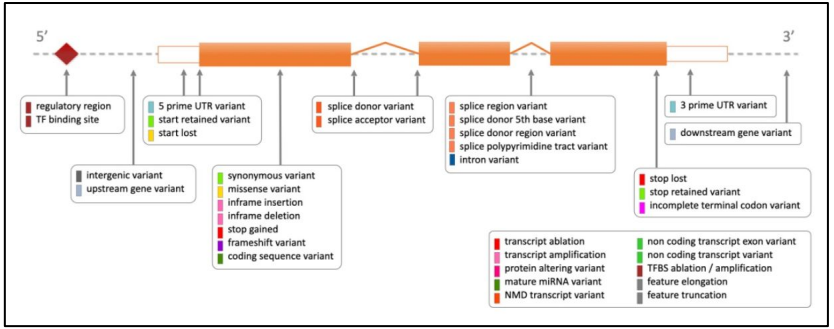


ADHD cases enriched for rare damaging coding variants

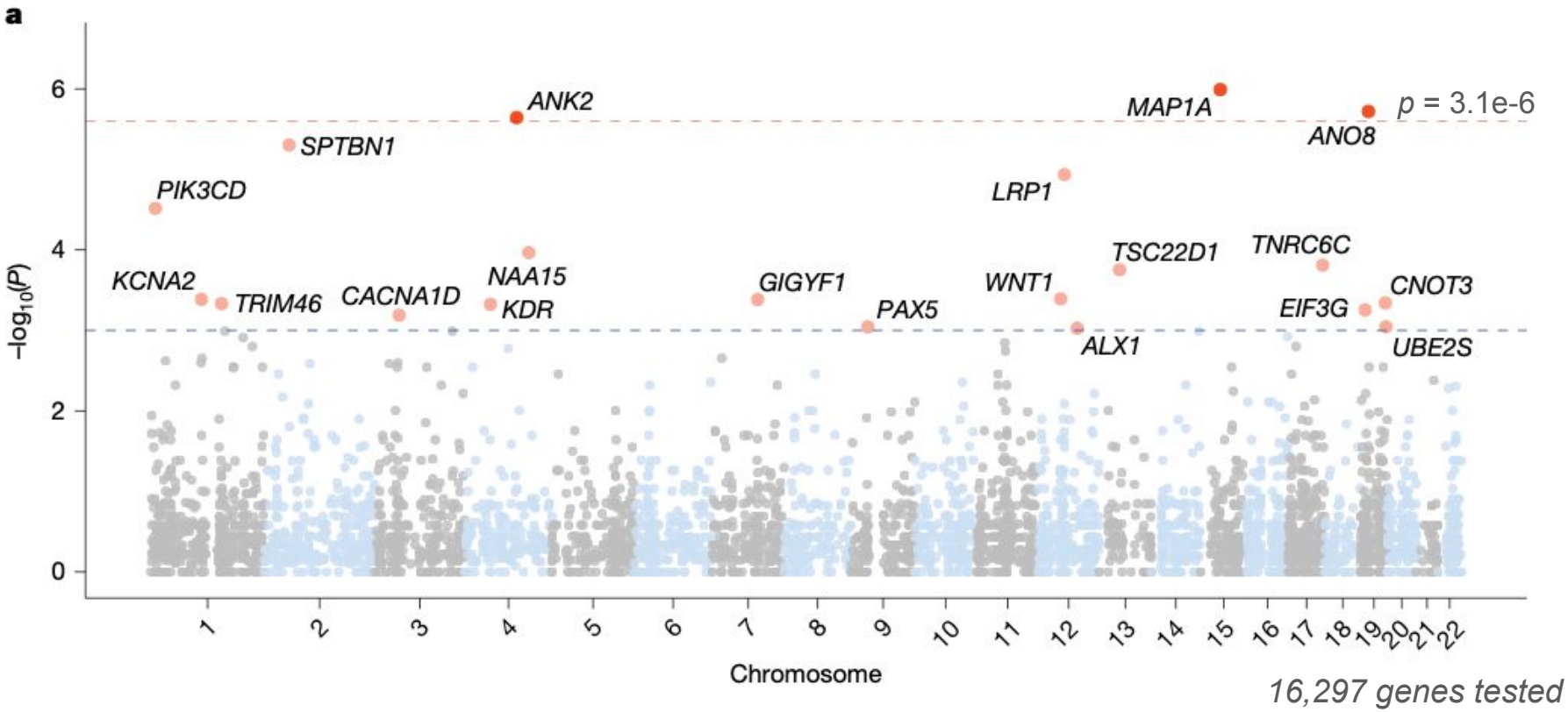




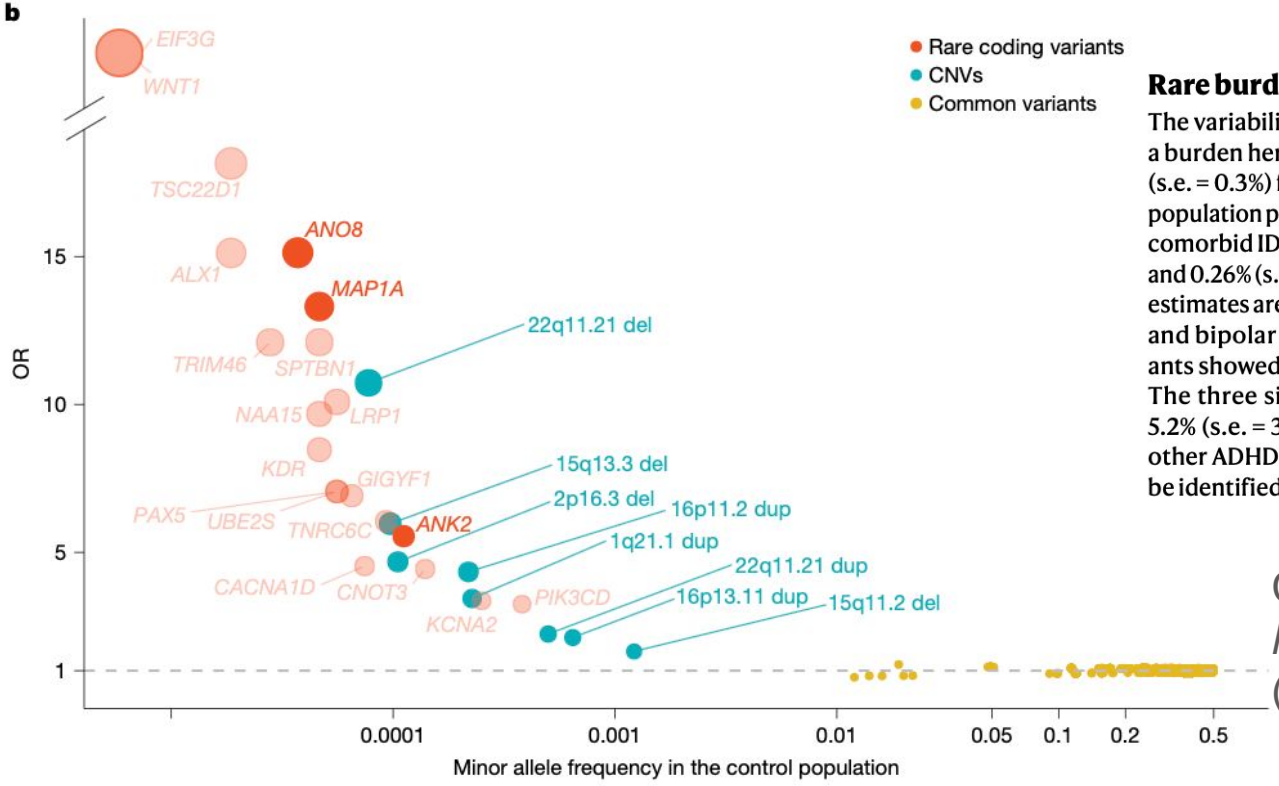
r = “rare”
PTV = Protein-truncating variant
DMV = Damaging Missense Variant
SYN = Synonymous Variant
ID = Intellectual Disability
pLI = probability of Loss-Of-Function Intolerance



Three ADHD risk genes reaching exome-wide significance



Rare variants of large effect explain a much smaller fraction of ADHD liability than common variants



Rare burden heritability

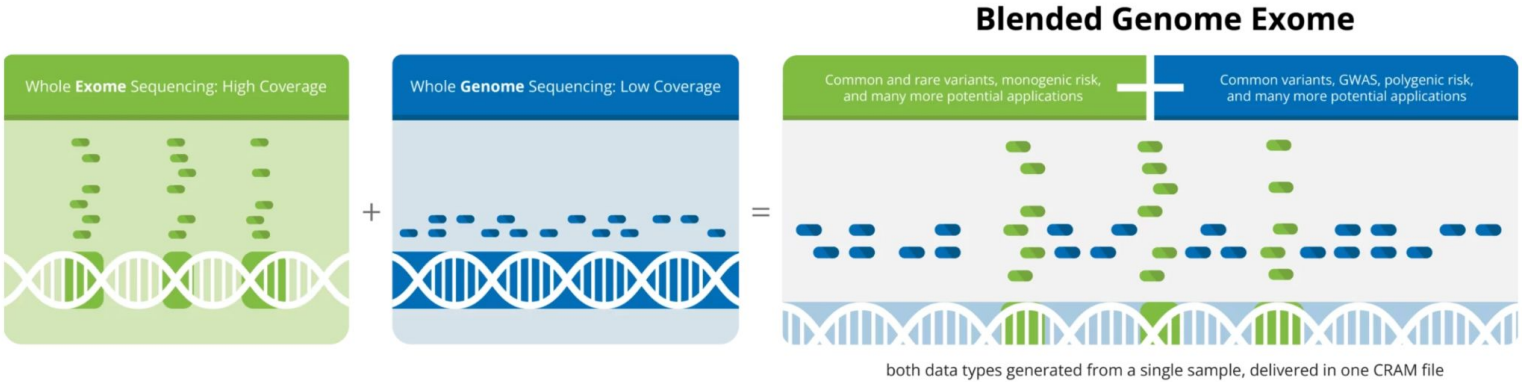
The variability in the phenotype explained by rare variants revealed a burden heritability of 2.5% (s.e. = 0.7%) for class I variants and 0.1% (s.e. = 0.3%) for class II variants for ADHD on the liability scale, using a population prevalence of 5% (Supplementary Table 10). When excluding comorbid ID, the burden heritability decreased to 1.43% (s.e. = 0.74%) and 0.26% (s.e. = 0.27) for class I and class II variants, respectively. These estimates are in line with findings for schizophrenia (1.7% (s.e. = 0.3%)) and bipolar disorder (1.8% (s.e. = 0.3%))[26]. Rare synonymous variants showed no evidence of non-zero burden heritability for ADHD. The three significant genes (*MAP1A*, *ANO8* and *ANK2*) explained 5.2% (s.e. = 3.4%) of the class I burden heritability, suggesting that other ADHD risk genes implicated by rare coding variants remain to be identified.

Compare to **14%** (s.e.=1%) for h^2 *SNP* from GWAS (Demontis et al., 2023)

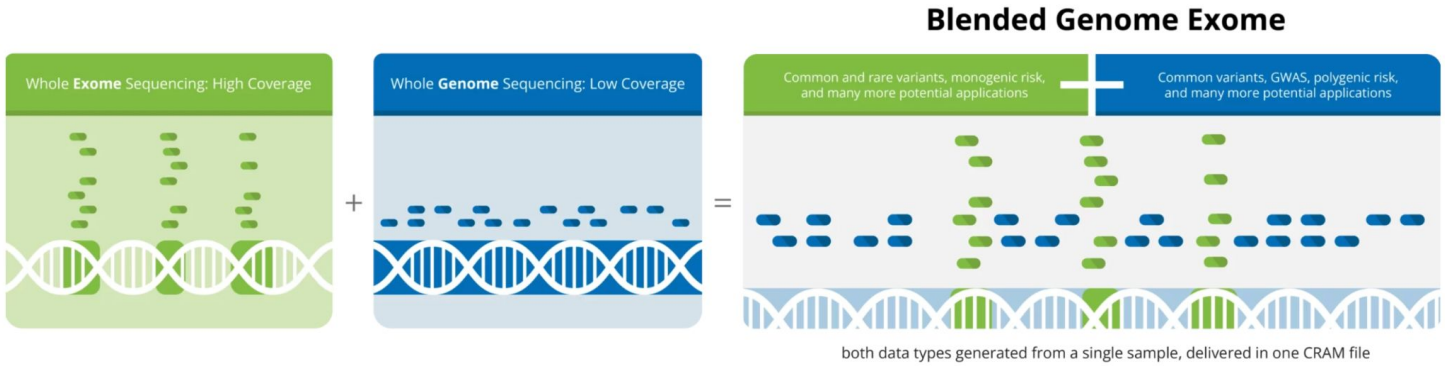
We have learned a lot in 15 years of ADHD genetics

- ADHD risk gene discovery is alive and well for both common and rare variants
- Genetics is helping to disentangle the relationship between ADHD epidemiology and other behavioral and cognitive phenotypes
- Many areas for further discovery and refinement
  - Still early days for exome sequencing
  - Need more extensive phenotyping / recontacting of future cohorts
  - Integration with other omics technologies
  - Expanding the success of GWAS / RVAS outside of predominantly European cohorts

Part2:
DNA sequencing with the Blended Genome Exome (BGE)



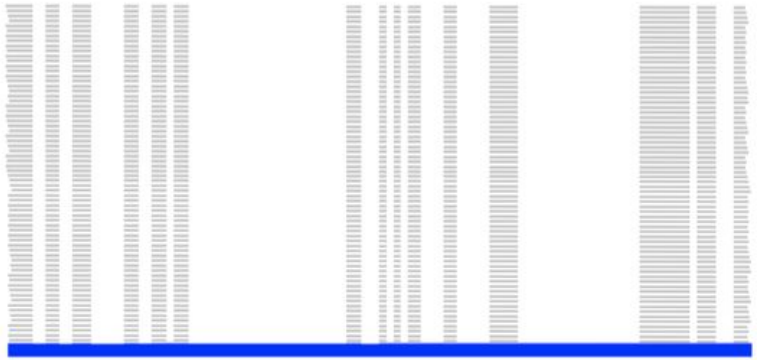
1. Motivation for developing the Blended Genome Exome protocol
2. Technical Development of BGE
3. Scaling and evaluation of BGE in a large multi-ancestry cohort
4. Current BGE projects and resources



Motivation

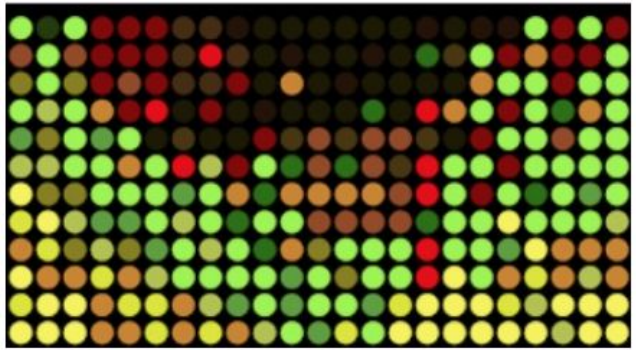
- Prior to BGE, the main product was an exome + genotyping array as separate products
- The main genotyping array used was the Global Screening Array (GSA)

Exome



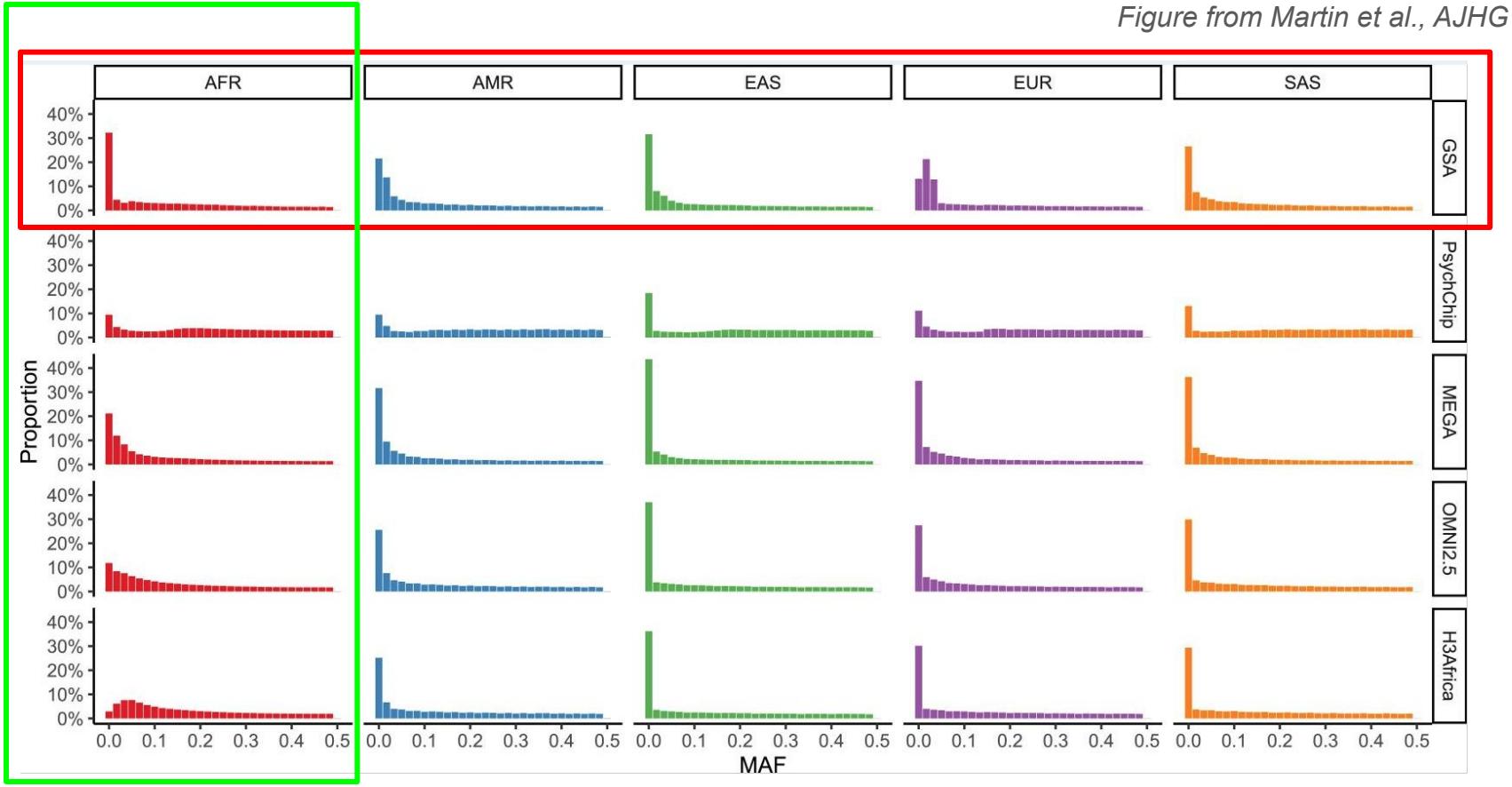
+

Genotyping Array



The diversity problem in the Global Screening Array

Figure from Martin et al., AJHG 2021



Motivation

- Alternatives to the exome + array strategy

    1. Sequence array SNPs with additional capture probes (Genotyping-by-sequencing)

    2. Generate low pass whole genome sequencing (WGS) in the same sequence run

    3. Deep WGS ($$$)

Option 1 does exist…

TWIST BIOSCIENCE

PRODUCTS APPLICATIONS RESOURCES COMPANY

Diversity SNP Panel

OVERVIEW ORDERING RESOURCES

As the first release in Twist's emerging Targeted Genotyping-By-Sequencing (GBS) portfolio, the Twist Diversity SNP panel leverages Twist's best-in-class DNA synthesis platform to generate a global panel of more than 600,000 probes governing approximately 1.4 million SNPs.

Used separately as a stand-alone genotyping panel or as a spike-in into Twist's Human Comprehensive Exome panel, this assay gives researchers a new ethnicity-neutral gold standard to use in generating genotyping data to match with their sequencing and other genomic data.

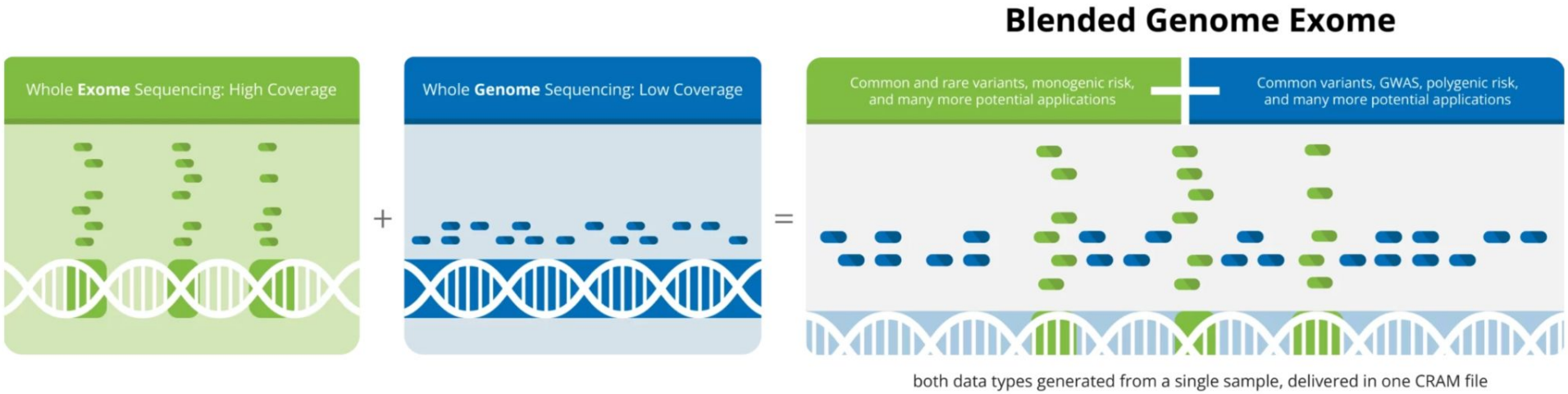
Twist Bioscience Collaborates with Regeneron for Production of Genotyping by Sequencing Panel to Enable Diverse Genome-wide Screening

June 14, 2021

-- Population Genetics Sequencing Panel Incorporates Global Genetic Variations for Superior Study of Disease and Target Discovery --

SOUTH SAN FRANCISCO, Calif.--(BUSINESS WIRE)--Jun. 14, 2021-- Twist Bioscience Corporation (Nasdaq: TWST), a company enabling customers to succeed through its offering of high-quality synthetic DNA using its silicon platform, today announced it collaborated with Regeneron Genetics Center LLC (RGC), a wholly-owned subsidiary of Regeneron (Nasdaq: REGN), for the production of a custom next-generation sequencing (NGS) population genetics genotyping assay. Arising from a need to incorporate the genetic differences of global populations, this assay is designed to gain new insights into disease mechanisms, identify novel drug targets, and accelerate drug discovery and development. Twist will market the assay as the Twist Diversity SNP Panel, and will make the content available to researchers globally for their population genomics studies.

BROAD
CLINICAL LABS



https://broadclinicallabs.org/clinical-blended-genome-exome-sequencing/

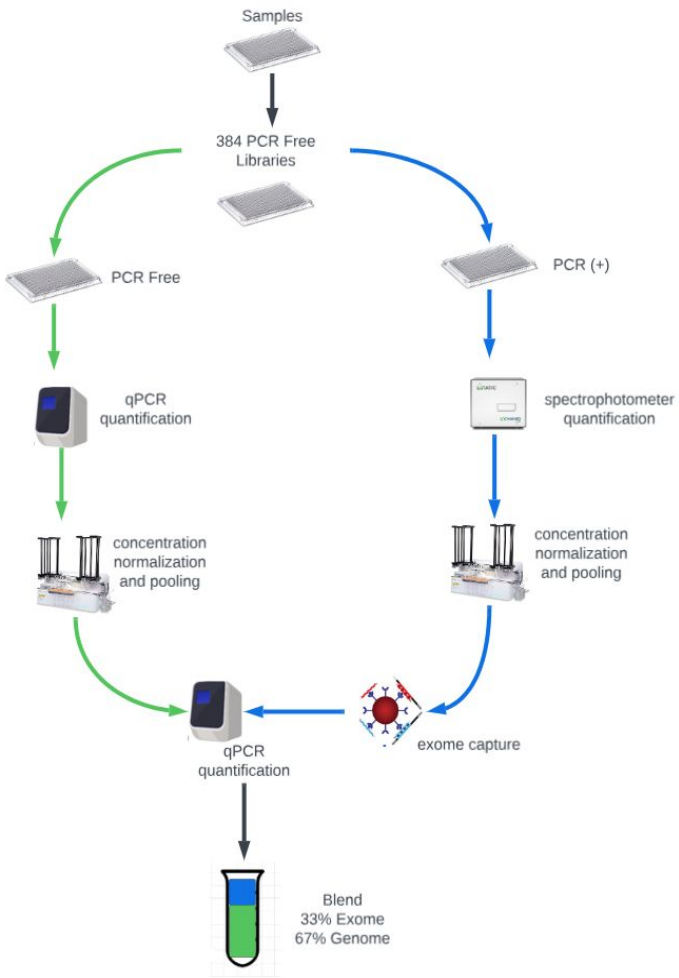
Technical Development of BGE

Blended Genome Exome (BGE) as a Cost Efficient Alternative to Deep Whole Genomes or Arrays

Matthew DeFelice, Jonna L. Grimsby, Daniel Howrigan, Kai Yuan, Sinéad B. Chapman, Christine Stevens, Samuel DeLuca, Megan Townsend, Joseph Buxbaum, Margaret Pericak-Vance, Shengying Qin, Dan J. Stein, Solomon Teferra, Ramnik J. Xavier, Hailiang Huang, Alicia R. Martin, Benjamin M. Neale

doi: https://doi.org/10.1101/2024.04.03.587209

Low pass WGS library



Exome capture library

The high-throughput technology behind BGE

Can run over 60 samples through a single lane of sequencing!

Gory details:

- Enzymatic fragmentation (NEBNext Ultra II FS kit)
  - NEB – New England Biosciences
- Quarter reaction volumes
- 384 sample batches (have 192 indexed adapters now)
- 384 well SPRI cleanups
  - SPRI – Solid Phase Reversible Immobilization
- Multiple additions of sample + bead to magnet
- Reduced cost exome capture
- Tempest for fast non-contact dispense destination normalization (384 in minutes!)

Lessons learned from Covid Dx and Covid Seq!

Low pass imputation using GLIMPSE software

GLIMPSE Home Overview Installation Documentation Benchmark GitHub

glimpse

Genotype Likelihoods IMputation and PhaSing mEthod

CATCH A GLIMPSE OF YOUR LOW DEPTH SEQUENCING DATA

GLIMPSE is a phasing and imputation method for large-scale low-coverage sequencing studies.

Main features of the method:

1. **Accurate imputed genotype calls.** Our method takes advantage of reference panels to produce high quality genotype calls.
2. **Accurate phasing.** GLIMPSE outputs accurate phased haplotypes for the low-coverage sequenced dataset.
3. **Low-coverage sequencing outperforms SNP arrays.** Imputation using low-coverage sequencing data is competitive to SNP array imputation. Results for European and African-American populations are interactively available on the website.
4. **A cost-effective paradigm.** GLIMPSE realises whole genome imputation from the HRC reference panel for less than 1$.

GLIMPSE tools is available under the MIT licence on the Github repository https://github.com/odelaneau/GLIMPSE.

HUGE thanks to Kai Yuan for being our GLIMPSE workflow expert

THE HUANG LAB

Kai Yuan
Research Fellow
Email: kyuan@broadinstitute.org

Kai Yuan is a postdoctoral fellow in the Massachusetts General Hospital and the Broad Institute, advised by Dr. Hailiang Huang. He obtained his PhD in computational biology from the Partner Institute for Computational Biology, Chinese Academy of Sciences. During his PhD training, he worked on population genetics, especially for the admixed populations and developed several methods to infer population

Low pass imputation using GLIMPSE software

"Variable position" = SNP in reference panel, not in sequence data

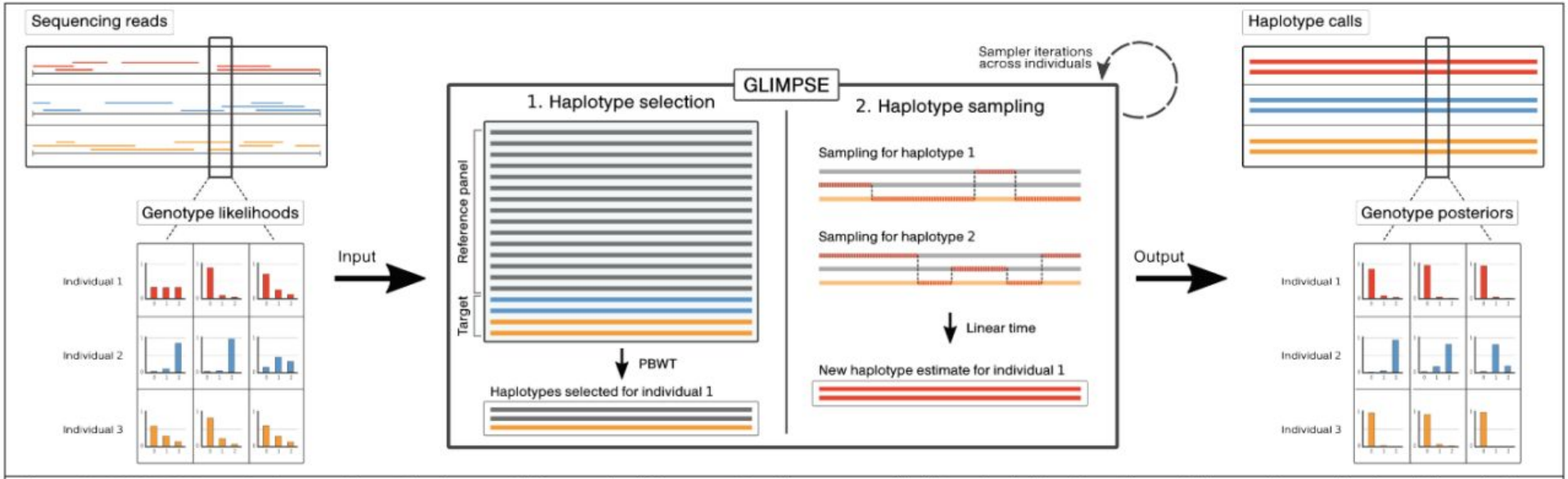
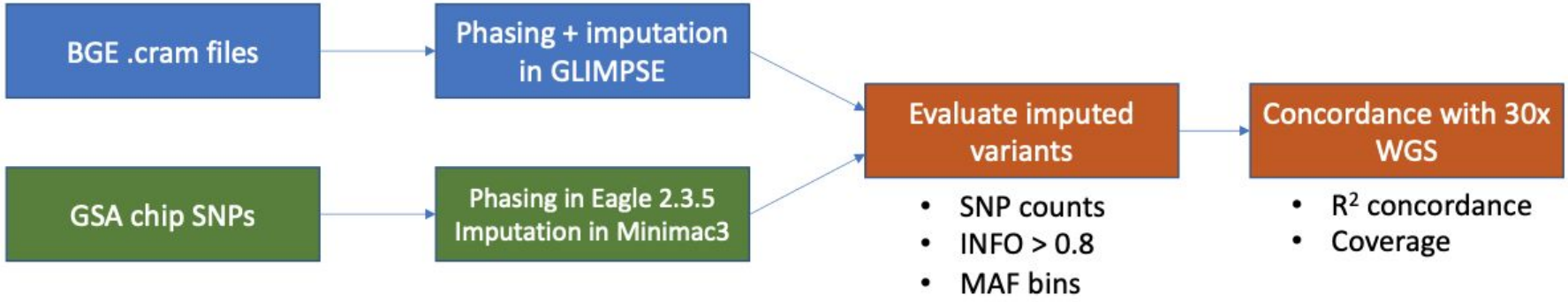


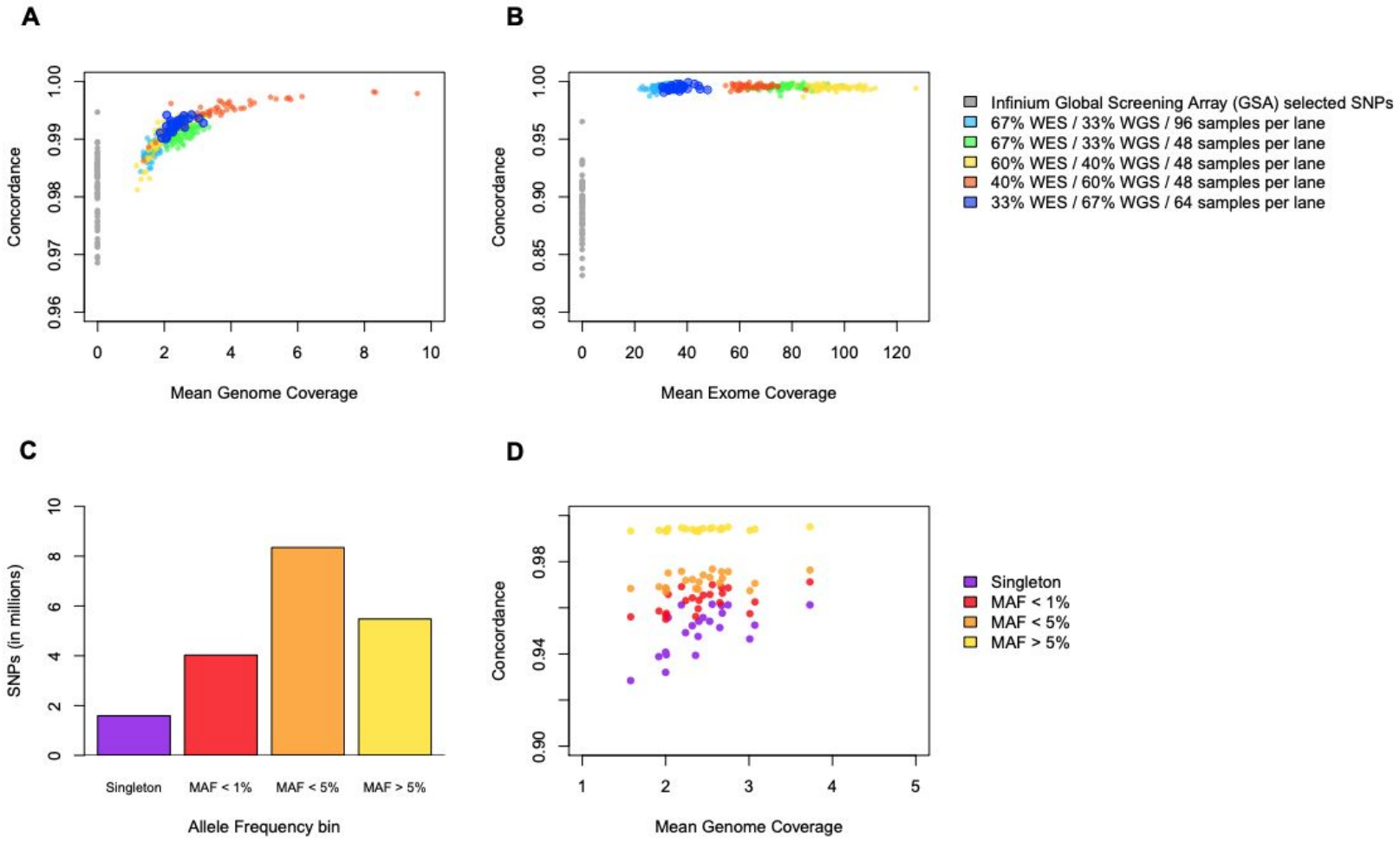
Figure1: GLIMPSE method overview. The input of the method is a matrix of genotype likelihoods defined at all variable positions obtained directly from the sequencing reads (left). GLIMPSE refines the genotype likelihoods using a Gibbs sampler scheme. At each iteration a new pair of haplotypes for each individual is estimated (middle). This involves two main steps: (1.) the haplotype selection using a reference panel and the current estimate of all other target haplotypes (middle, left) and (2.) a linear time sampling algorithm based on the Li and Stephens model (middle, right). As an output, GLIMPSE produces consensus-based haplotype calls and genotype posteriors at every variable position (right).

Evaluating BGE performance using deep whole genomes

- Using high quality calls from 30x genomes as our "truth" dataset
  - Compare low-pass GLIMPSE imputation against Global Screening Array (GSA) chip

- Pilot sample sets
  - Early rounds: 31 to 62 Hispanic samples
  - Later rounds: 23 African samples from PUMAS (Ethiopia and South Africa)




A
Concordance
1.00
0.99
0.98
0.97
0.96
0
2
4
6
8
10
Mean Genome Coverage
B
Concordance
1.00
0.95
0.90
0.85
0.80
0
20
40
60
80
100
120
Mean Exome Coverage
Infinium Global Screening Array (GSA) selected SNPs
67% WES / 33% WGS / 96 samples per lane
67% WES / 33% WGS / 48 samples per lane
60% WES / 40% WGS / 48 samples per lane
40% WES / 60% WGS / 48 samples per lane
33% WES / 67% WGS / 64 samples per lane
C
SNPs (in millions)
0
2
4
6
8
10
Singleton
MAF < 1%
MAF < 5%
MAF > 5%
Allele Frequency bin
D
Concordance
0.98
0.94
0.90
1
2
3
4
5
Mean Genome Coverage
Singleton
MAF < 1%
MAF < 5%
MAF > 5%

What is BGE again?

Fragmented Blood/saliva DNA

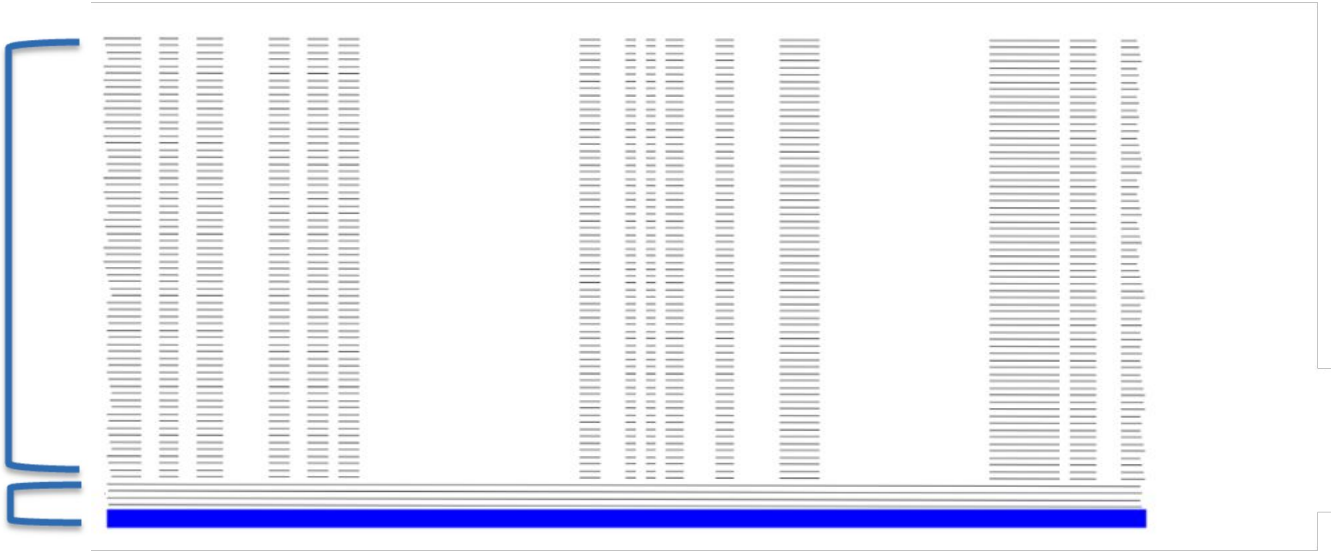
33% exome

67% genome

Sequenced reads after blending

30x whole exome

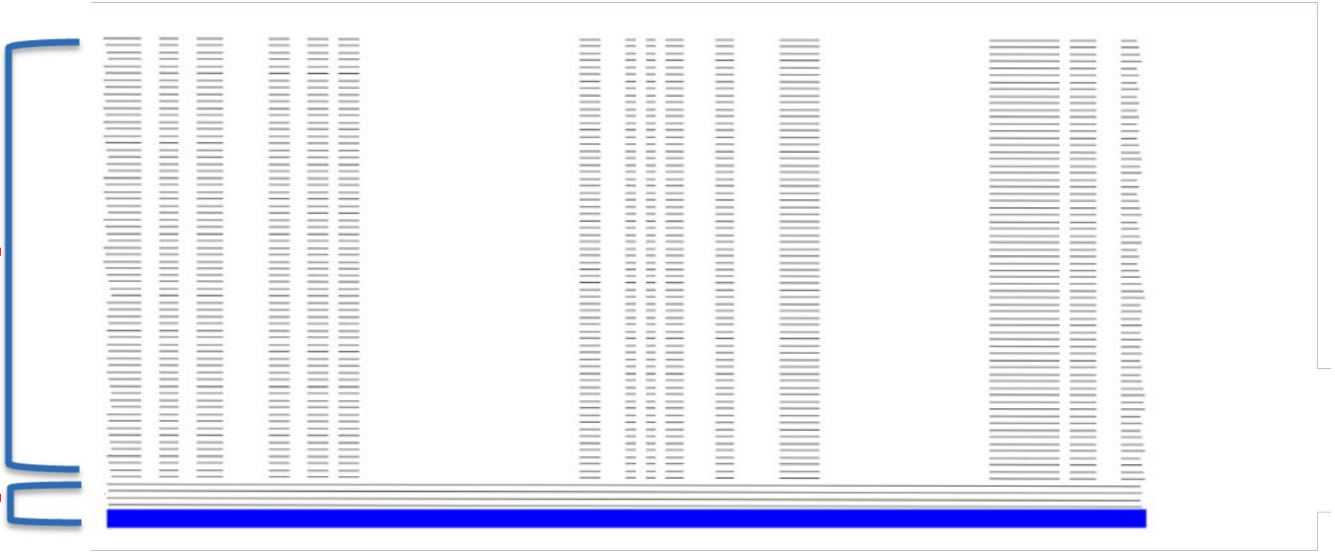
4x whole genome



What is BGE again?

33% exome

67% genome



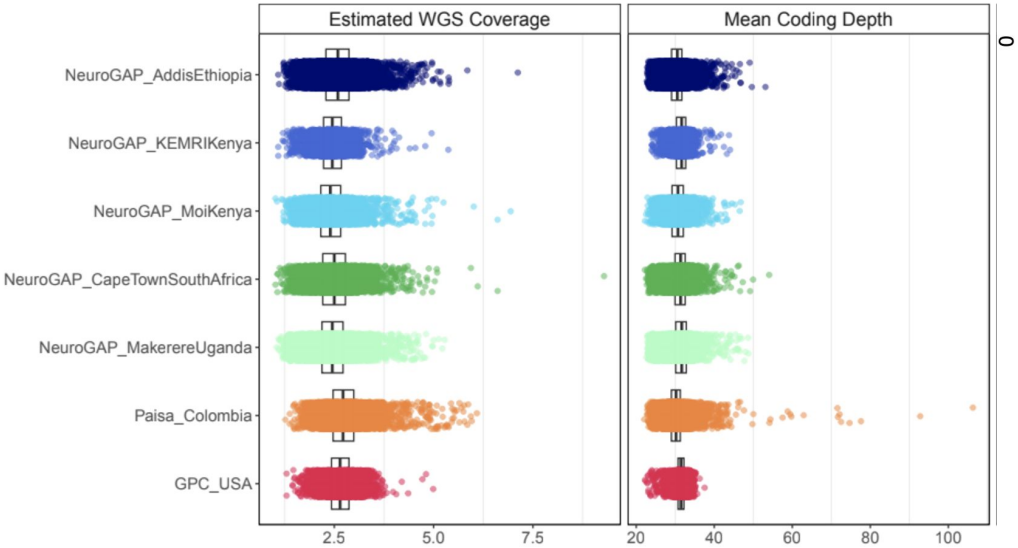
~~30x whole exome~~

25-40x whole exome

~~4x whole genome~~

1-4x whole genome

Mean coverage distributions



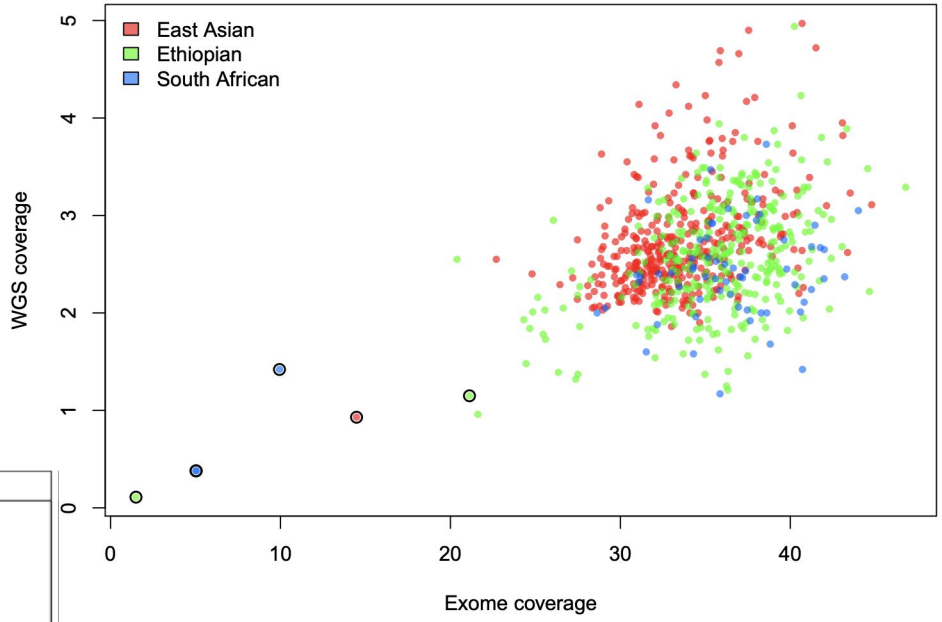
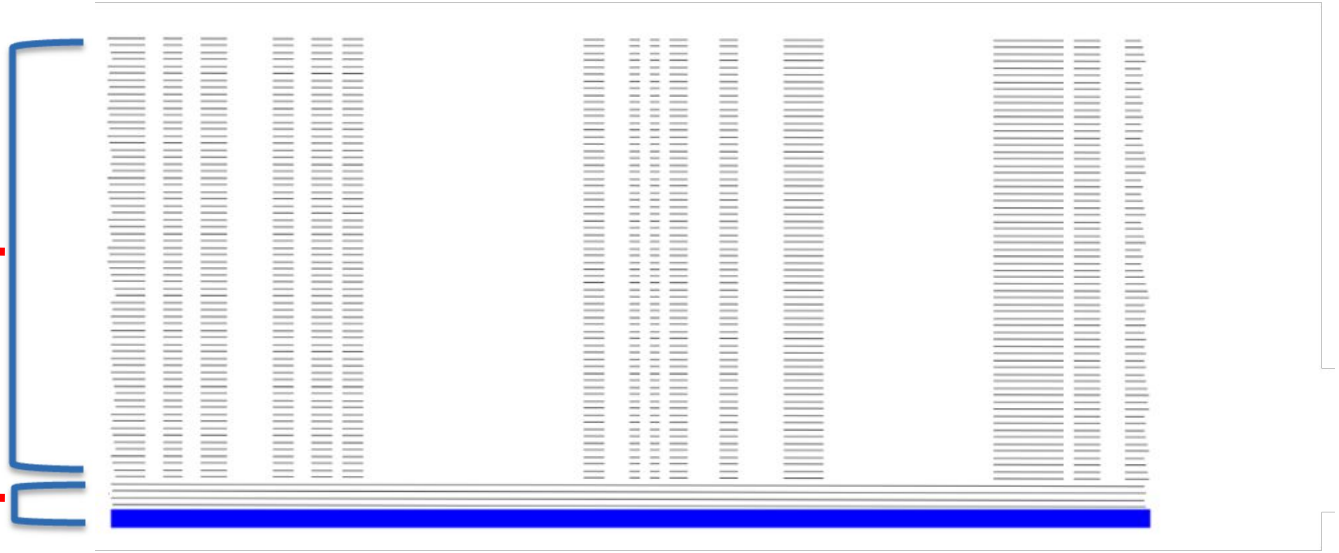


Figure from Julia Sealock

What is BGE again?

33% exome

67% genome



~~30x whole exome~~

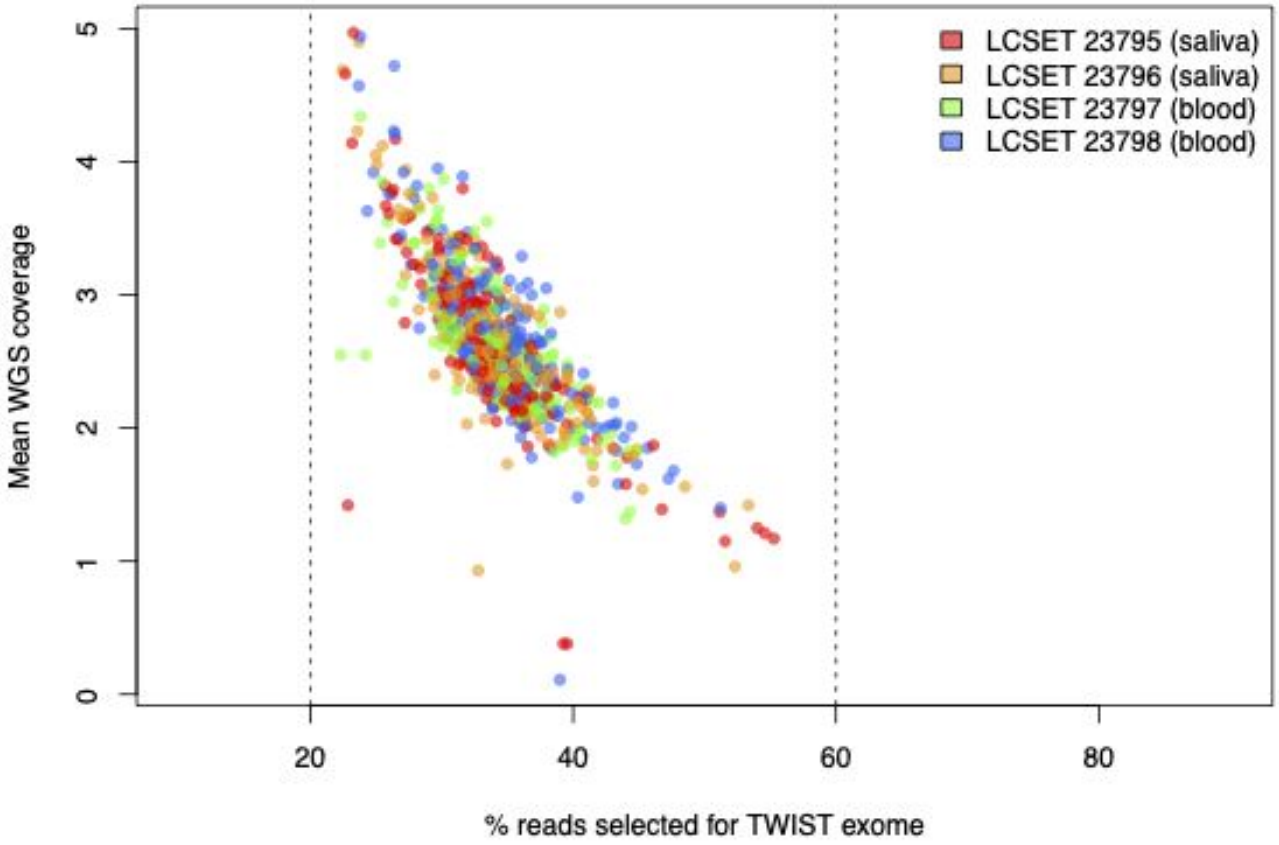
~~25-40x whole exome~~

90% of exome at 10x

~~4x whole genome~~

~~1-4x whole genome~~

At least 45% of reads


LCSET 23795 (saliva)
LCSET 23796 (saliva)
LCSET 23797 (blood)
LCSET 23798 (blood)
Mean WGS coverage
0
1
2
3
4
5
20
40
60
80
% reads selected for TWIST exome

Scaling and evaluation of BGE in a large multi-ancestry cohort

A blended genome and exome sequencing method captures genetic variation in an unbiased, high-quality, and cost-effective manner

Toni A Boltz, Benjamin B Chu, Calwing Liao, Julia M Sealock, Robert Ye, Lerato Majara, Jack M Fu, Susan Service, Lingyu Zhan, Sarah E Medland, Sinéad B Chapman, Simone Rubinacci, Matthew DeFelice, Jonna L Grimsby, Tamrat Abebe, Melkam Alemayehu, Fred K Ashaba, Elizabeth G Atkinson, Tim Bigdeli, Amanda B Bradway, Harrison Brand, Lori B Chibnik, Abebaw Fekadu, Michael Gatzen, Bizu Gelaye, Stella Gichuru, Marissa L Gildea, Toni C Hill, Hailiang Huang, Kalyn M Hubbard, Wilfred E. Injera, Roxanne James, Moses Joloba, Christopher Kachulis, Phillip R Kalmbach, Rogers Kamulegeya, Gabriel Kigen, Soyeon Kim, Nastassja Koen, Edith K. Kwobah, Joseph Kyebuzibwa, Seungmo Lee, Niall J Lennon, Penelope A Lind, Esteban A Lopera-Maya, Johnstone Makale, Serghei Mangul, Justin McMahon, Pierre Mowlem, Henry Musinguzi, Rehema M. Mwema, Noeline Nakasujja, Carter P Newman, Lethukuthula L Nkambule, Conor R O'Neil, Ana Maria Olivares, Catherine M. Olsen, Linnet Ongeri, Sophie J Parsa, Adele Pretorius, Raj Ramesar, Faye L Reagan, Chiara Sabatti, Jacquelyn A Schneider, Welelta Shiferaw, Anne Stevenson, Erik Stricker, Rocky E. Stroud II, Jessie Tang, David Whiteman, Mary T Yohannes, Mingrui Yu, Kai Yuan, NeuroGAP-Psychosis, Dickens Akena, Lukoye Atwoli, Symon M. Kariuki, Karestan C. Koenen, Charles R. J. C. Newton, Dan J. Stein, Solomon Teferra, Zukiswa Zingela, Carlos N Pato, Michele T Pato, Carlos Lopez-Jaramillo, Nelson Freimer, Roel A Ophoff, Loes M Olde Loohuis, Michael E Talkowski, Benjamin M Neale, Daniel P Howrigan, Alicia R Martin

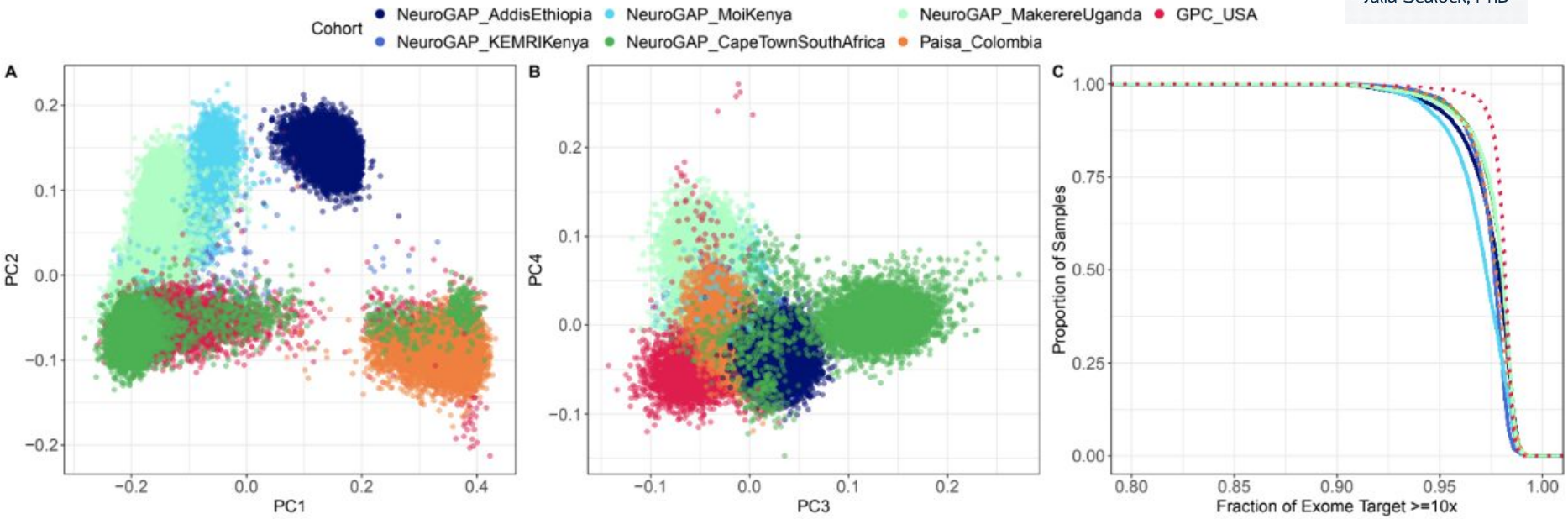
doi: https://doi.org/10.1101/2024.09.06.611689

- Application to 50k samples
- Verification of rare SNV, SV, and common SNP capture

PUMAS = Populations Underrepresented in Mental illness Association Studies

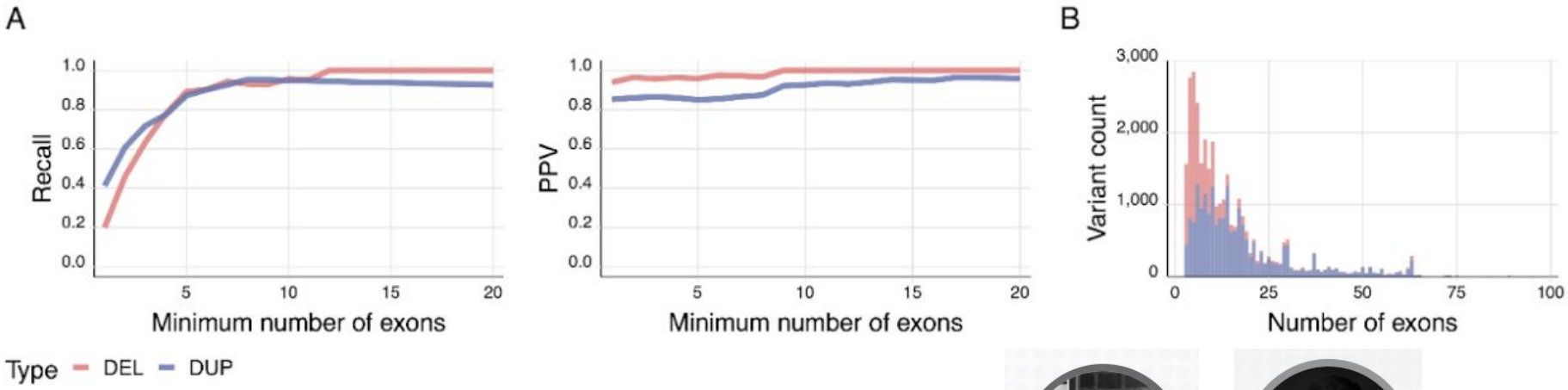
Cohort name and location	# Total Pre-QC	# Total Post-QC
NeuroGAP AddisEthiopia	11,715	11,027
NeuroGAP KEMRIKenya	3,078	2,889
NeuroGAP MoiKenya	5,040	4,716
NeuroGAP CapeTownSA	8,747	5,779
NeuroGAP MakerereUganda	11,306	10,727
Paisa Colombia	9,007	8,200
GPC USA	4,553	3,926
Total	**53,446**	**47,264**

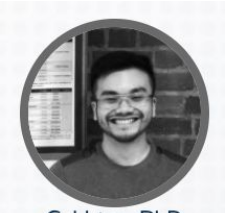

Julia Sealock, PhD


Cohort
NeuroGAP_AddisEthiopia
NeuroGAP_MoiKenya
NeuroGAP_MakerereUganda
GPC_USA
NeuroGAP_KEMRIKenya
NeuroGAP_CapeTownSouthAfrica
Paisa_Colombia
A
PC2
PC1
B
PC4
PC3
C
Proportion of Samples
Fraction of Exome Target >=10x

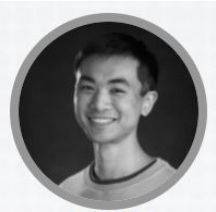
Structural variation in the exome capture using GATK-gCNV

Concordance against deep whole genome sequencing (SFARI = 400 samples)





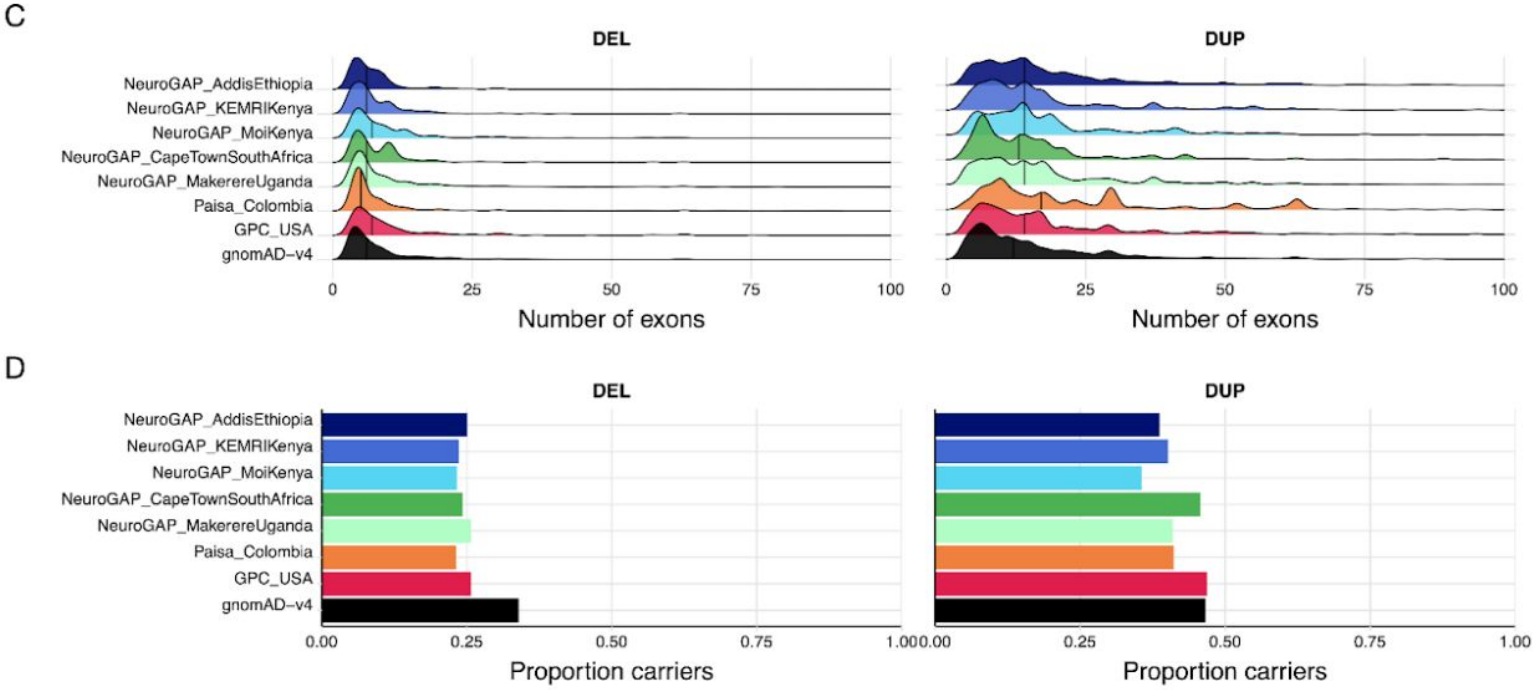
Cal Liao, PhD



Bob Ye

Structural variation in the exome capture using GATK-gCNV

Evaluation of PUMAS SV calls against gnomAD v4



Imputation evaluation in PUMAS cohort

- Evaluated against GSA array genotypes instead of 30x WGS

Dataset	N samples with BGE and GSA data
NeuroGAP - AddisEthiopia	158
NeuroGAP - KEMRIKenya	183
NeuroGAP - MoiKenya	157
NeuroGAP - CapeTown South Africa	162
NeuroGAP - Makerere Uganda	178
Paisa - Colombia	1,191
GPC - USA	3,932
QIMR - AGBP	3,664
QIMR - QSkin	3,545

Cohort	Sample Size	Cost	Cost Per Sample
NeuroGAP	35,279	$12,763.23	0.361
Paisa	8,317	$3,025.45	0.363
GPC	3,967	$1,406.46	0.354
QIMR	7,499	$2,732.96	0.364
Total	**55,062**	**$19,928.10**	**0.361**

GLIMPSE2

Website:
https://odelaneau.github.io/GLIMPSE/

Preprint:

https://www.biorxiv.org/content/10.1101/2022.11.28.518213v1.full.pdf

About

GLIMPSE2 is a set of tools for low-coverage whole genome sequencing imputation. GLIMPSE2 is based on the GLIMPSE model and designed for reference panels containing hundreads of thousands of reference samples, with a special focus on rare variants.

Citation

If you use GLIMPSE in your research work, please cite the following papers:

Rubinacci et al., Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. BiorXiv (2022)

Rubinacci et al., Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nature Genetics 53.1 (2021): 120-126.

Get started now | View source code on GitHub

GLIMPSE1

At the moment, GLIMPSE2 performs imputation only from a reference panel of samples. To use the joint-model, particularly useful for many samples at higher coverages (>0.5x) and small reference panels, please visit the GLIMPSE1 website.

Abstract

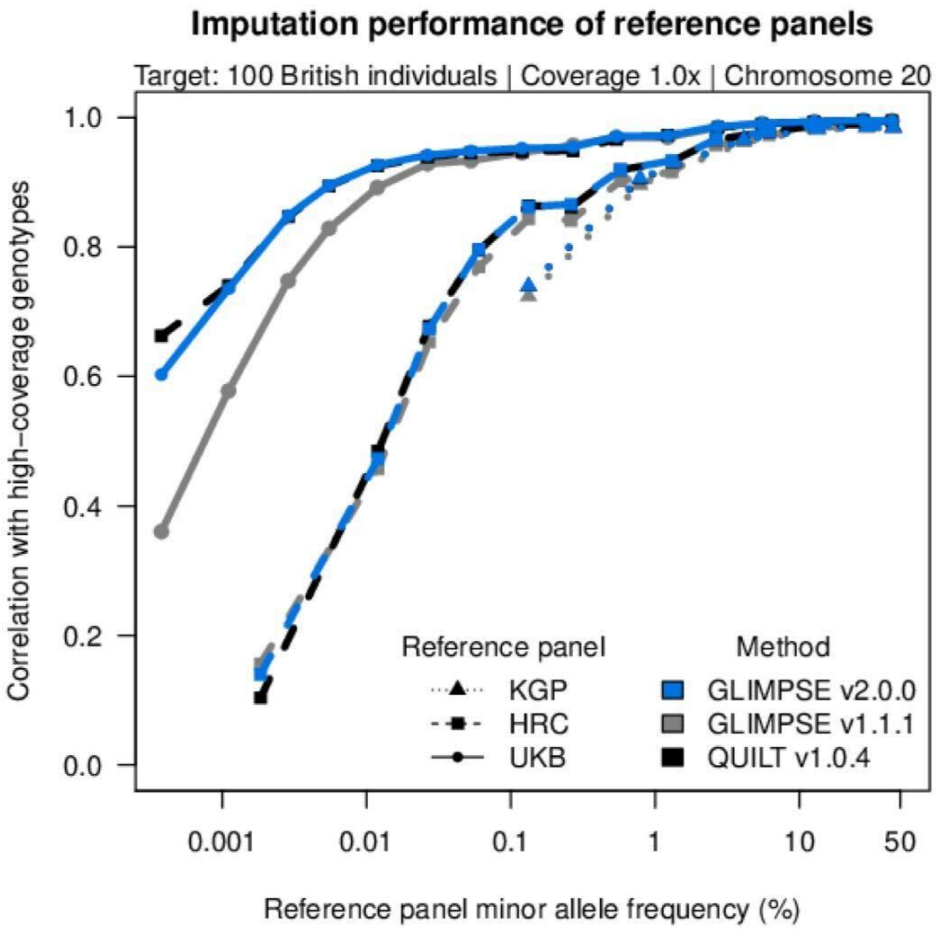
Recent work highlights the advantages of low-coverage whole genome sequencing (lcWGS), followed by genotype imputation, as a cost-effective genotyping technology for statistical and population genetics. The release of whole genome sequencing data for 150,119 UK Biobank (UKB) samples represents an unprecedented opportunity to impute lcWGS with high accuracy. However, despite recent progress[1,2], current methods struggle to cope with the growing numbers of samples and markers in modern reference panels, resulting in unsustainable computational costs. For instance, the imputation cost for a single genome is 1.11£ using GLIMPSE v1.1.1 (GLIMPSE1) on the UKB research analysis platform (RAP) and rises to 242.8£ using QUILT v1.0.4. To overcome this computational burden, we introduce GLIMPSE v2.0.0 (GLIMPSE2), a major improvement of GLIMPSE, that scales sublinearly in both the number of samples and markers. GLIMPSE2 imputes a low-coverage genome from the UKB reference panel for only 0.08£ in compute cost while retaining high accuracy for both ancient and modern genomes, particularly at rare variants (MAF < 0.1%) and for very low-coverage samples (0.1x-0.5x).

GLIMPSE2 features

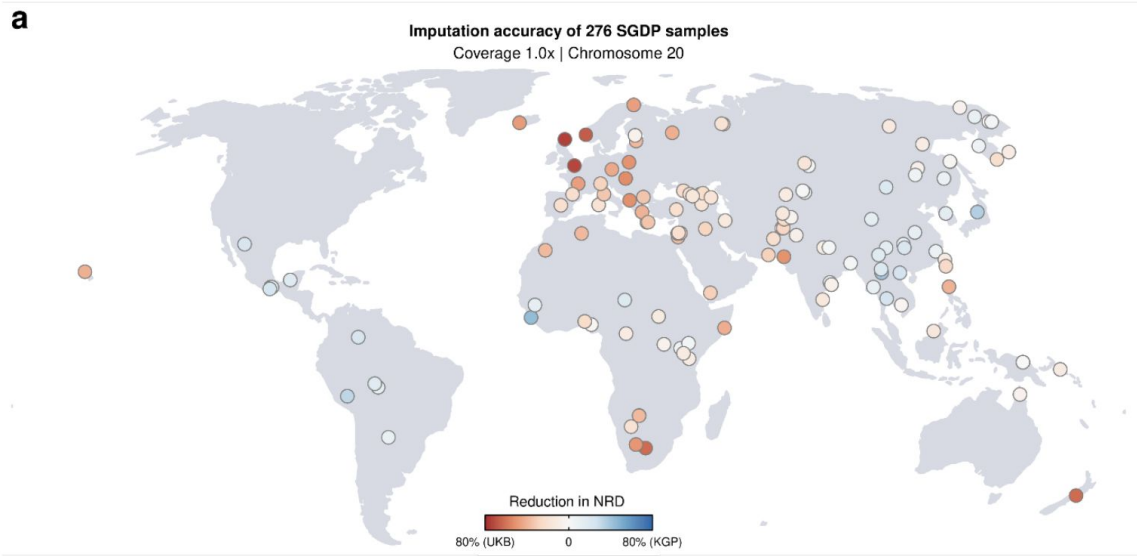
GLIMPSE2 is designed to perform imputation based only on the reference panel, and optimises this task with seven novel key features:

1. A sparse memory representation of the reference panel that stores efficiently the large number of rare variants it contains (**Section S1.2.2.1**),
2. An efficient implementation of the hidden Markov model (HMM) that speeds up probability computations by leveraging the sparsity of the reference panel (**Section S1.2.2.2**),
3. A new data structure based on the Positional Burrows Wheeler Transform (PBWT) that speeds up haplotype matching by leveraging the sparsity of the reference panel (**Section S1.2.2.3**),
4. A sparse file format for the reference panel, containing also the pre-computed PBWT data structures, that allows fast loading times (**Section S1.2.2.4**),
5. A genotype caller that internalises the pile-up of the sequencing data and the computations of genotype likelihoods (**Section S1.2.2.5**),
6. A model extension to impute small indels and low-quality bi-allelic variants separately from SNPs (**Section S1.2.2.6**),
7. An optimised iteration scheme that integrates an initialisation step based on rare variant sharing (**Section S1.2.2.7**).

a


Imputation performance of reference panels
Target: 100 British individuals | Coverage 1.0x | Chromosome 20
Correlation with high−coverage genotypes
1.0
0.8
0.6
0.4
0.2
0.0
0.001
0.01
0.1
1
10
50
Reference panel minor allele frequency (%)
Reference panel
KGP
HRC
UKB
Method
GLIMPSE v2.0.0
GLIMPSE v1.1.1
QUILT v1.0.4

Using SGDP (Simons Genome Diversity Project) to evaluate imputation accuracy in diverse genomes



Supplementary Figure 8: Performance of SGDP samples using different reference panels. (**a-b**) Comparison between KGP and the UKB reference panels to impute 276 SGDP samples across 129 world-wide populations at 1.0x coverage on chromosome 20. (**a**) Per-sample comparison. Each circle represents one sample of SGDP and is coloured according to the reduction in NRD achieved when using the UKB reference panel (red) or KGP (blue). Location represents the geographical origin of the sample. (**b**) Population-level comparison. Samples belonging to the same population (x-axis) have been considered together (number shown in the x-axis label), showing the reduction of NRD between the two panels (y-axis). Populations have been coloured and ordered according to the continent and country of origin. Striped bars represent populations where KGP performs better than UKB reference panels.

NRD = Non-Reference Discordance

Red dots = UKB ref panel worked best

Blue dots = KGP ref panel worked best

Deeper color = larger difference between reference sample accuracy

Note:
1KG = KGP = 1000 genomes

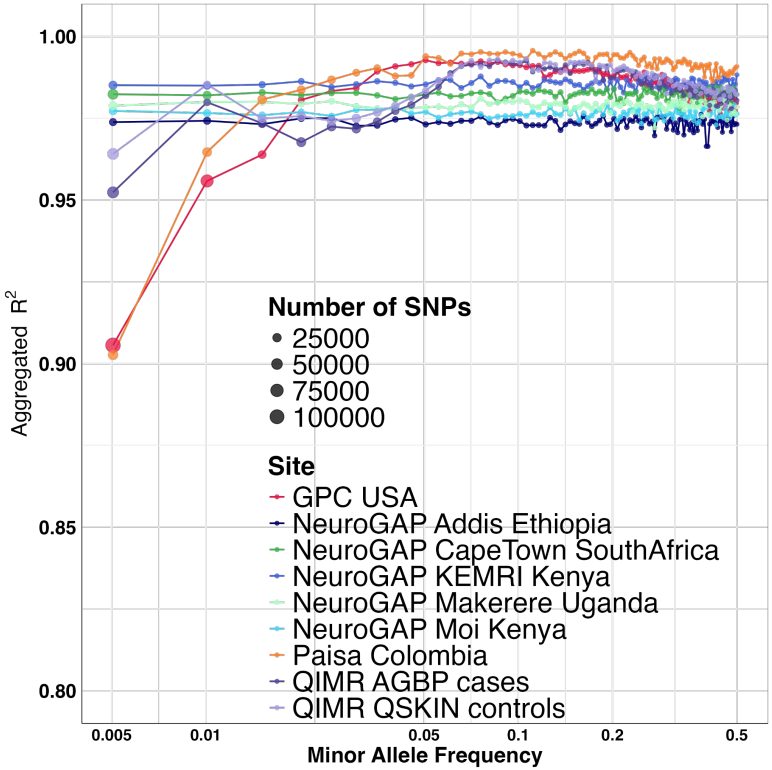
Imputation with GLIMPSE2

WGS reference panel
- Hg38 1KG+HDGP (4.1k samples)
- Indels, singleton + doubleton SNPs removed
- 67 million SNPs imputed

Imputation QC
- INFO score > 0.8
- Genotype posterior (GP) = 1

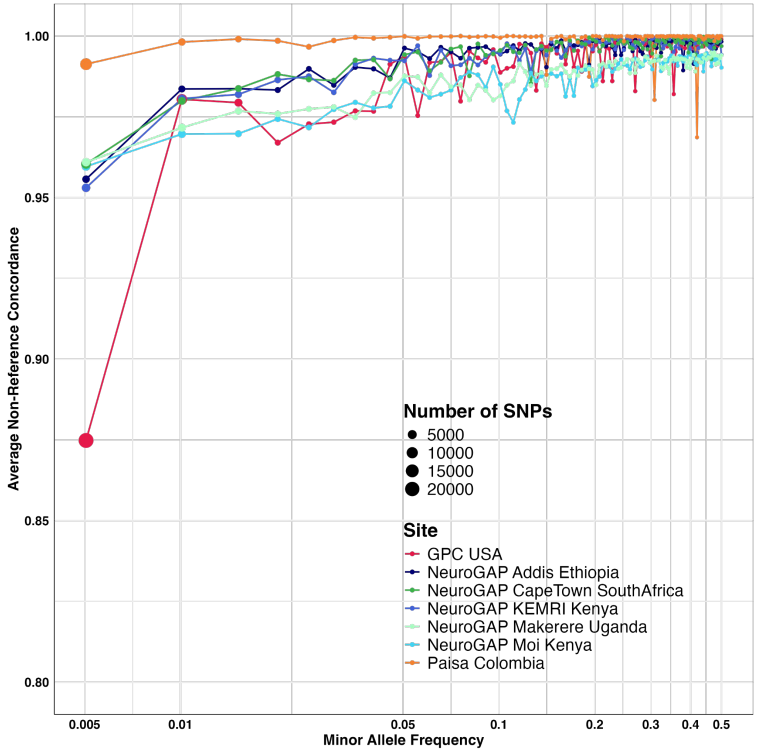
Results largely consistent with pilot samples

Concordance with GSA array genotypes

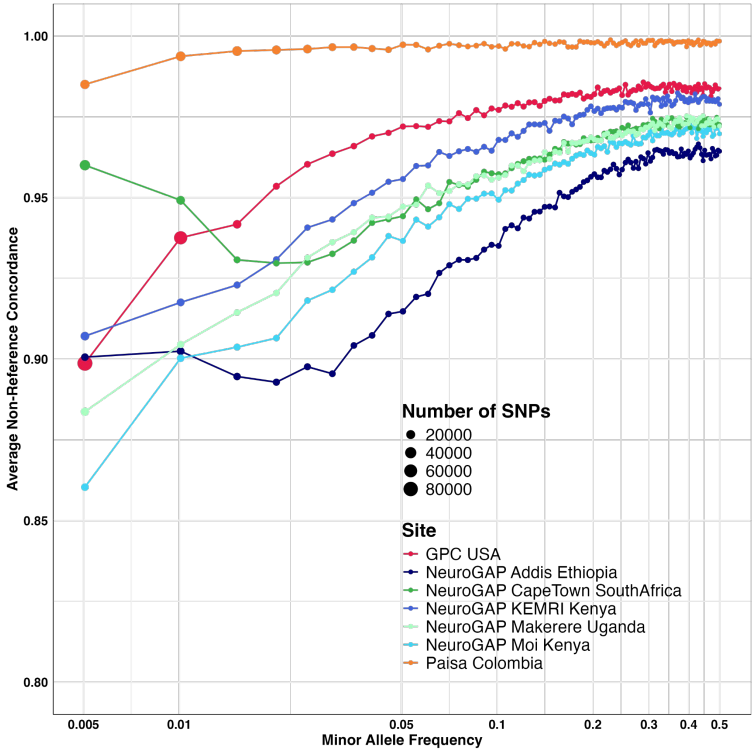


Non-reference concordance across coding / non-coding SNPs

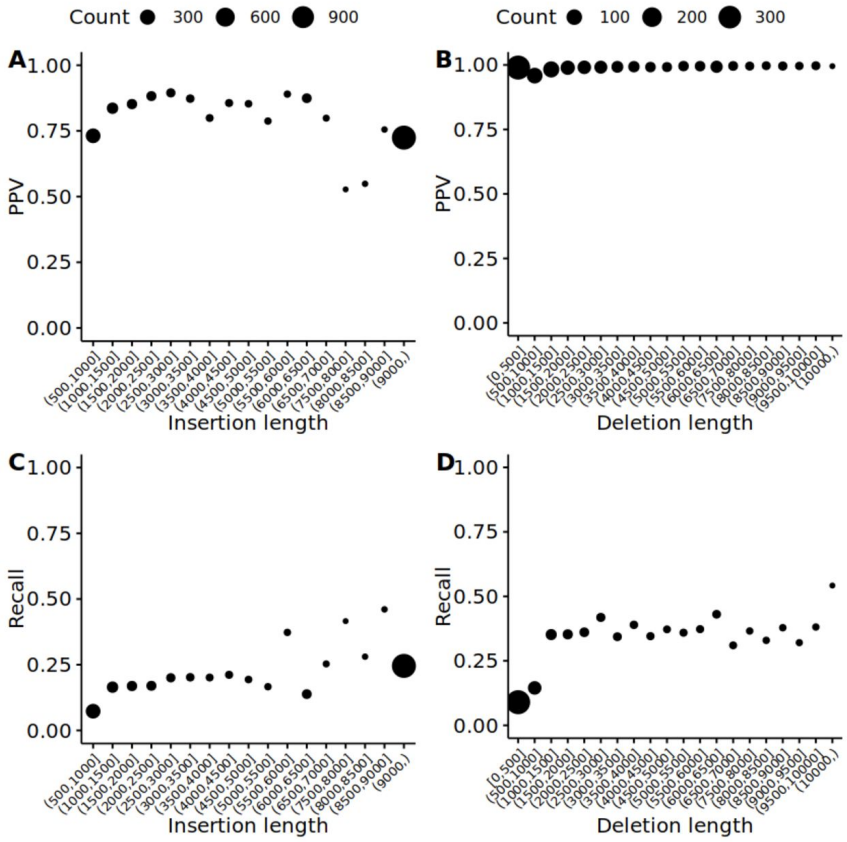
Coding SNPs



Non-coding SNPs



How well can we call structural variants (SVs) in the BGE low pass genome?



PPV = positive predictive value, or "what proportion of these called SVs are seen in the 30x genomes?"

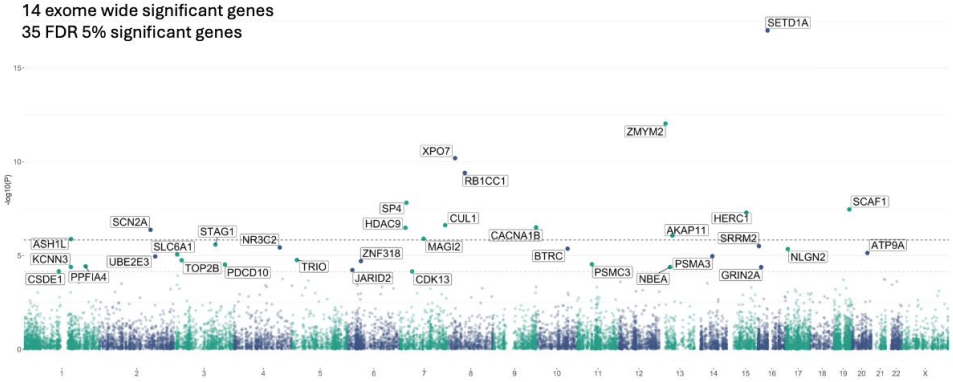
Recall = "How much of the 30x genome SV calls did we capture"

Current BGE projects and resources

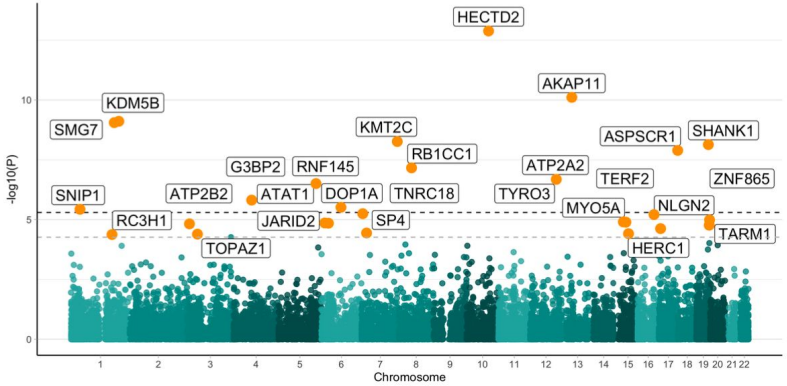
Julia Sealock, PhD

Cal Liao, PhD

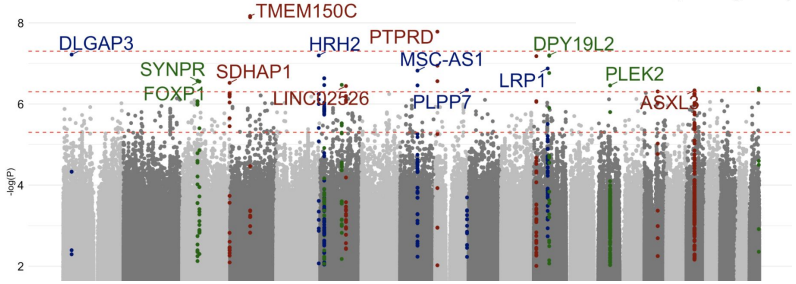
Lerato Majara



SCHEMA 2
59k SCZ

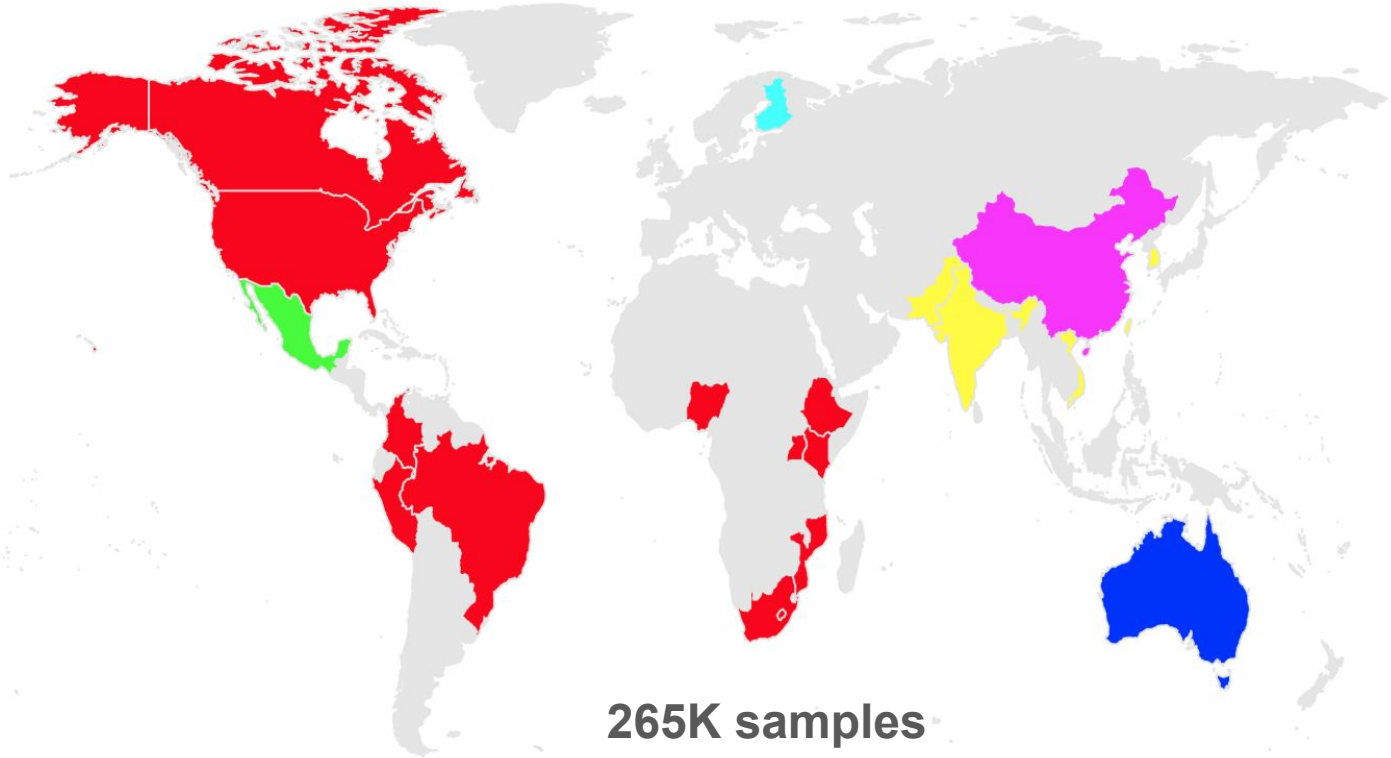


BipEx 2
45k BPD



NeuroGAP GWAS
11k SCZ
6k BPD

Ongoing BGE sample collections



PUMAS
Total = 140k

A-BIG-NET
Total = 43K

BioX
Total = 30K

NeuroMEX
Total = 6.5K

QIMR
Total = 8.3K

MGBB
Total = 4.8K

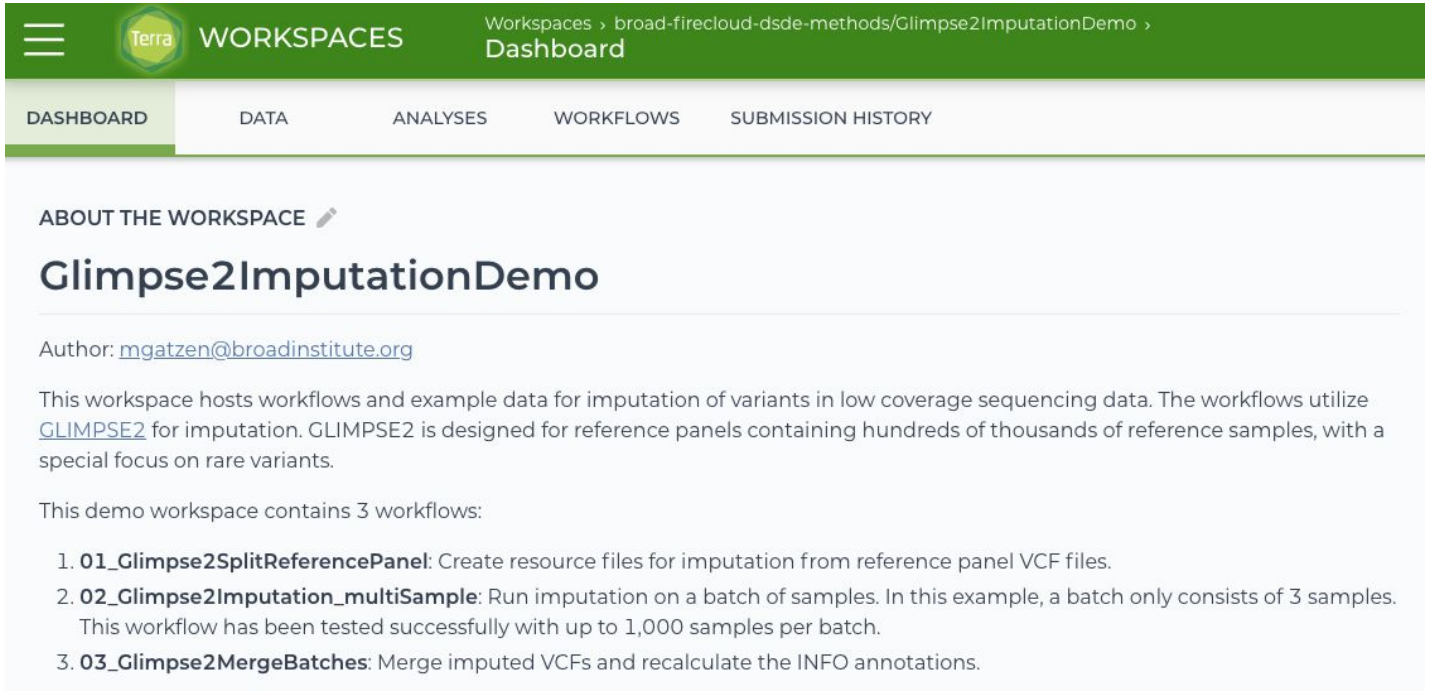
FINBB
Total = 12.6K

KFF
Total = 10K

IBD
Total = 10K

Running GLIMPSE via Terra Workflows

https://app.terra.bio/#workspaces/broad-firecloud-dsde-methods/Glimpse2ImputationDemo



Bigger reference panel in the works!


Franjo Ivankovic, PhD

https://allofus-anvil-imputation.terra.bio/

All of Us + AnVIL Imputation Service

Imputation can help complete your datasets **efficiently and accurately**

Our imputation service leverages Terra and uses a large and diverse reference panel that combines genomes from both the *All of Us* Research Program and AnVIL Centers for Common Disease Genomics.

BGE toolkit (in the works!)

https://atgu.github.io/bge-toolkit/index.html

Jackie Goldstein
BGE pipeline

BGE Toolkit

Search docs

CONTENTS:

- Command-Line Interface
  - bge-toolkit
  - bge-toolkit qc
  - bge-toolkit qc concordance
- Python API Reference

bge-toolkit qc

```
qc

Usage: bge-toolkit qc [OPTIONS] COMMAND [ARGS]...

Run QC-related tools on BGE datasets.

Options
--install-completion        Install completion for the current shell.
--show-completion           Show completion for the current shell, to copy it or customize the
                            installation.
--help                      Show this message and exit.

Commands
concordance   Run concordance on BGE exome and imputation datasets.
```

bge-toolkit qc concordance

```
concordance

Usage: bge-toolkit qc concordance [OPTIONS] COMMAND [ARGS]...

Run concordance on BGE exome and imputation datasets.

Options
*  --exome                  TEXT    Exome dataset to compare to. [default: None]
                                     [required]
*  --imputation             TEXT    Imputation dataset to compare to. [default: None]
                                     [required]
*  --output-dir             TEXT    Output directory. [default: None] [required]
   --EXOME-MAF                      Bin concordance counts by Exome Minor Allele
                                    Frequency. (Global+Variant)
   --EXOME-MAC                      Bin concordance counts by Exome Minor Allele
                                    Counts. (Global+Variant)
```

Thanks everyone!

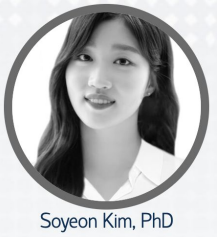
Christiana Liu
BGE pipeline

PUMAS Phenotypes
Reference panel

Jackie
Goldstein
BGE pipeline

NeuroGAP
analysts

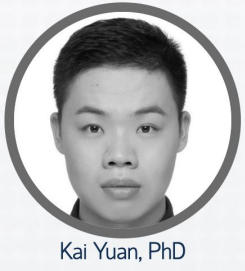
BipEx analyst



BioX
analyst

SCHEMA
analyst

Pipeline Ops



IBD analyst

Pipeline Ops

BipEx + SV QC
pipeline