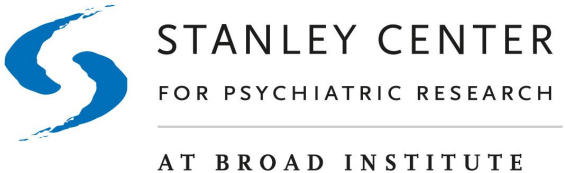
Rare Variants

Daniel Howrigan, PhD
Senior Group Leader - Neale lab

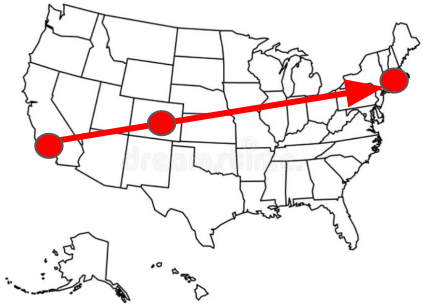
European Social Science Genetics Network (ESSGN) 4th year training

24-26 March 2026



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



METHODOLOGY ARTICLE Open Access

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

Daniel P Howrigan

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

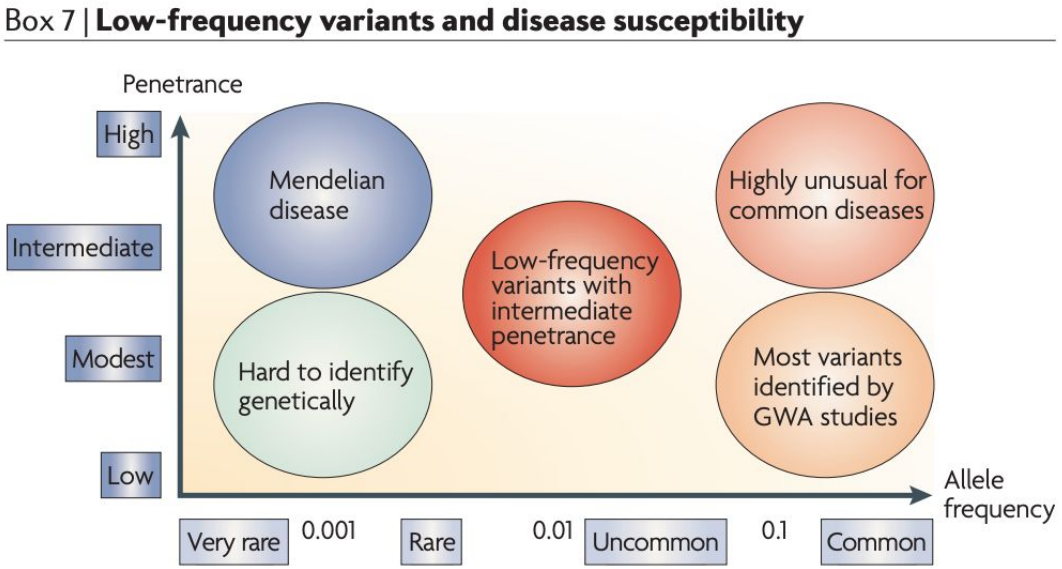
When should we considered genetic variant “rare”?

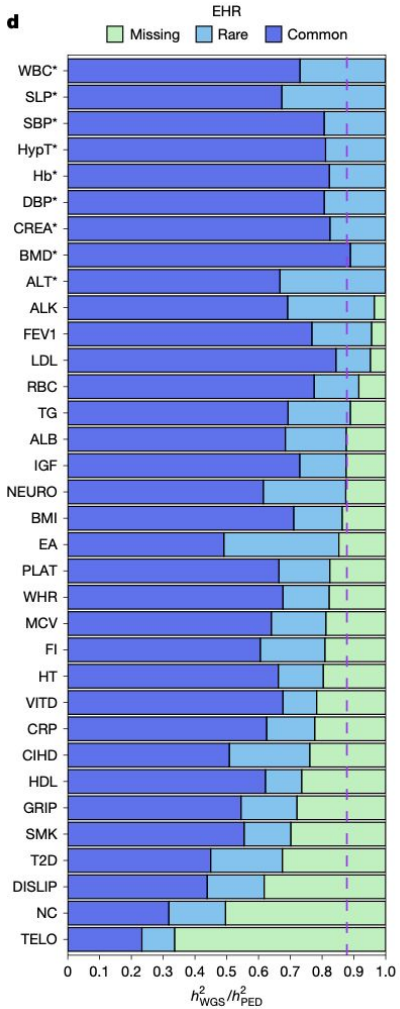
Common variant minor allele frequency (MAF) > 5%

Low frequency = 1-5%

Rare = 0.01 - 1%

Very or ultra-rare < .01%





Article

Estimation and mapping of the missing heritability of human phenotypes

https://doi.org/10.1038/s41586-025-09720-6

Received: 12 January 2025

Accepted: 8 October 2025

Published online: 12 November 2025

Pierrick Wainschtein[1,2]✉, Yuanxiang Zhang[2], Jeremy Schwartzentruber[1], Irfahan Kassam[1], Julia Sidorenko[2], Petko P. Fiziev[1], Huanwei Wang[2,3], Jeremy McRae[1], Richard Border[4], Noah Zaitlen[5,6], Sriram Sankararaman[5,7,8], Michael E. Goddard[9,10], Jian Zeng[2], Peter M. Visscher[2,11], Kyle Kai-How Farh[1,12] & Loic Yengo[2,12]✉

Overview of study design

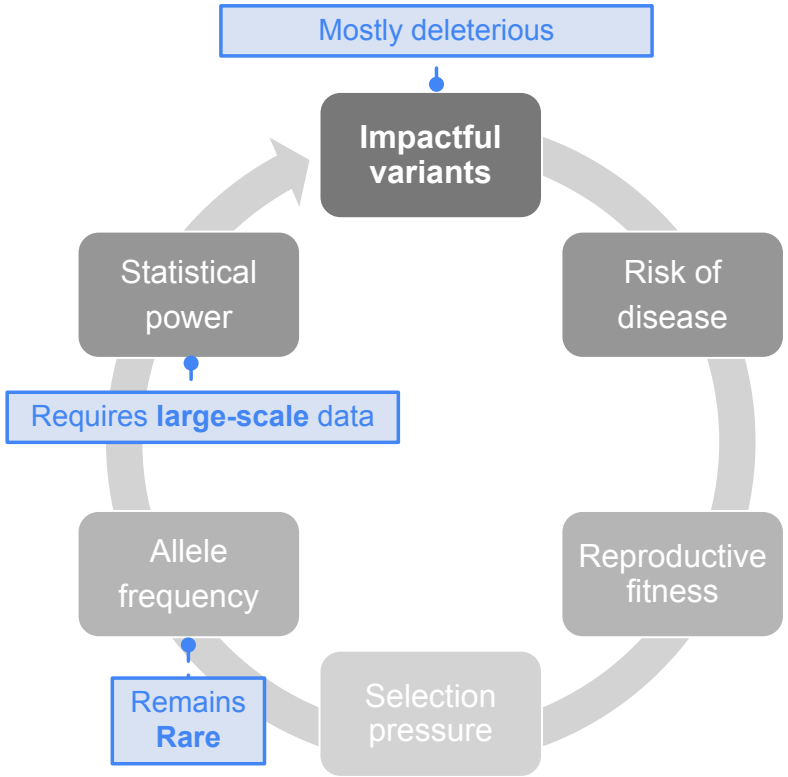
We analysed 490,542 genomes included in the second tranche of WGS data released by the UKB in December 2023[3]. We focused our main analyses on 40,575,204 autosomal sequence variants (including bi-allelic and multi-allelic SNPs and indels) with a MAF > 0.01% (Supplementary Tables 1 and 2) in a genetically homogeneous sample of 347,630 conventionally unrelated individuals (that is, with a genomic relationship coefficient lower than 0.05) sampled from a larger subgroup of 452,618 UKB participants with European ancestry (Methods). We selected 41 complex phenotypes spanning a wide range of human traits and common diseases (Supplementary Table 3) and showing a marginally significant estimate of h^2_{PED} from 171,446 pairs of relatives in the UKB. We then estimated h^2_{WGS} for these 41 traits using the GREML-LDMS method[17] implemented in MPH v.0.53.2 (ref. 18). Heritability estimates for all 41 phenotypes are reported in Supplementary Table 4.

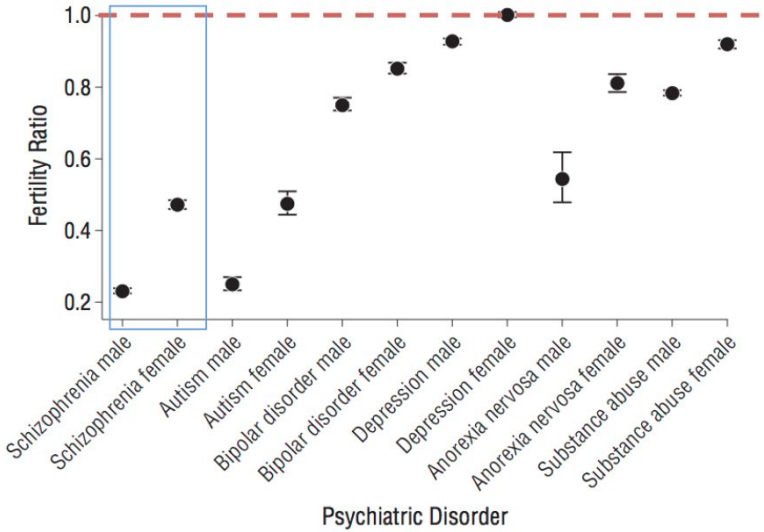
What is the minimum minor allele count (MAC) tested?

Common vs. Ultra-rare variants

	Common variation	Ultra-rare variation
Abundance	~ 10M	1B?
Population specific?	No	Often
Correlation	High	Low
Selection	Unlikely	Likely
Default technology	Array genotyping + imputation	Sequencing

Challenges of RVAS



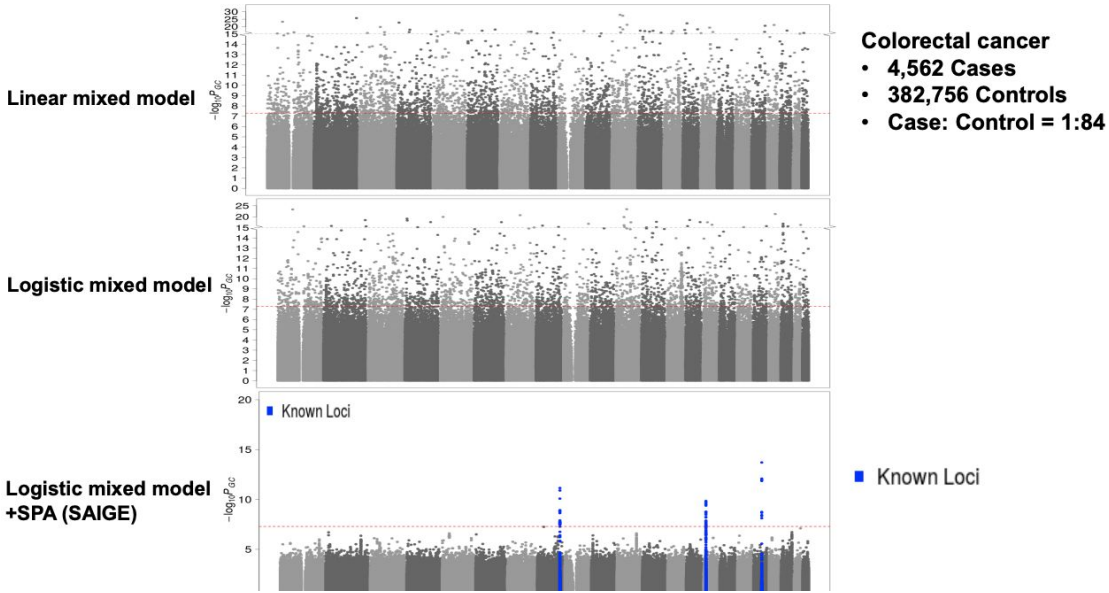


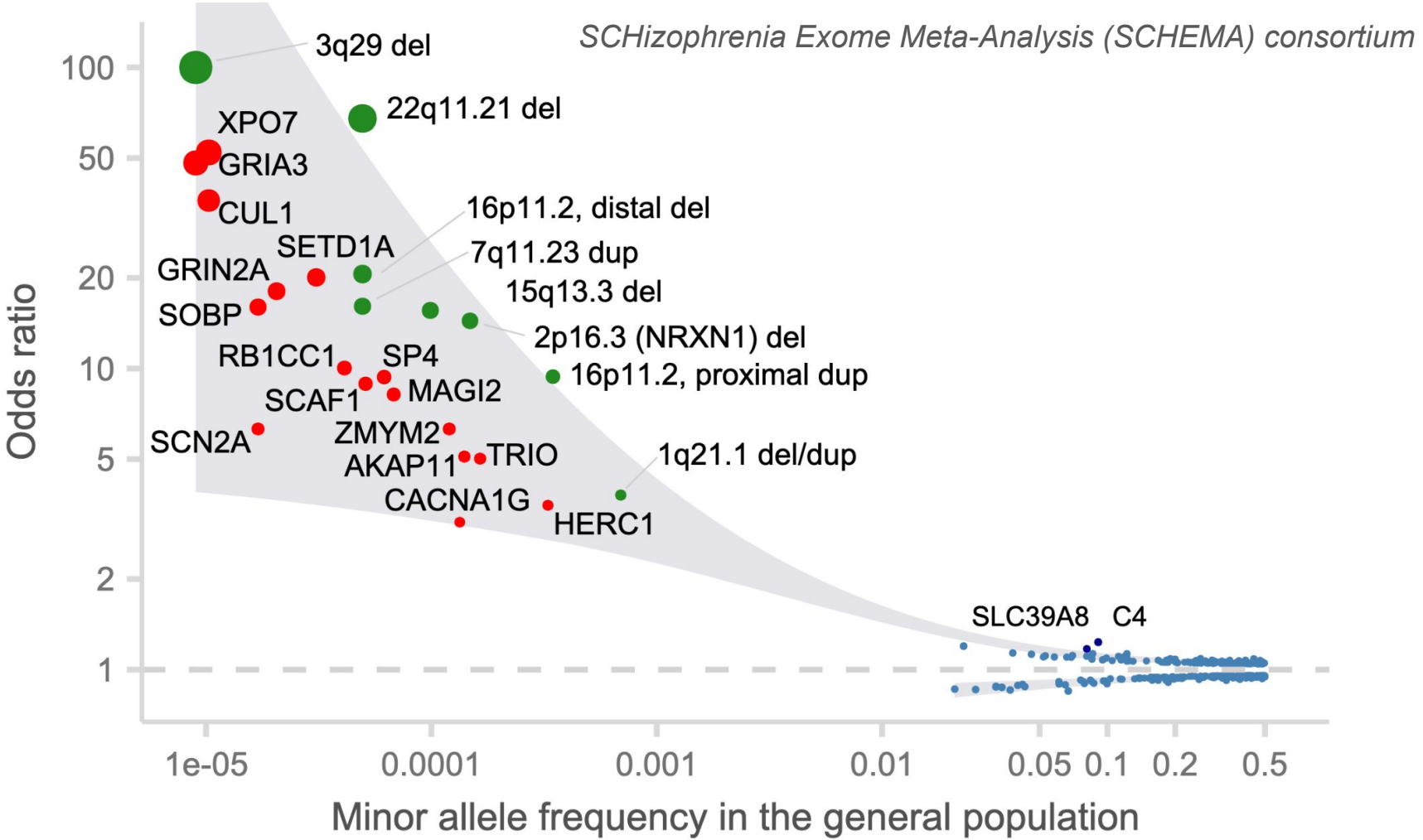
Power, JAMA Psychiatry, 2013

Single SNP association models breakdown at low allele counts, not frequencies

- For case/control studies, low allele counts leads to overdispersion of effect sizes, particularly in unbalanced case/control ratios
- Below MAC ~25, even “fully penetrant” variants unlikely to surpass multiple-testing correction in balanced case/control ratios

SAIGE



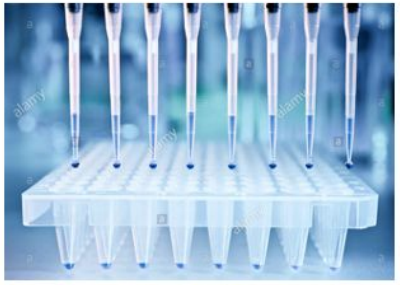

SCHizophrenia Exome Meta-Analysis (SCHEMA) consortium
3q29 del
22q11.21 del
XPO7
GRIA3
CUL1
16p11.2, distal del
SETD1A
7q11.23 dup
GRIN2A
15q13.3 del
SOBP
2p16.3 (NRXN1) del
RB1CC1
SP4
16p11.2, proximal dup
MAGI2
SCAF1
ZMYM2
SCN2A
TRIO
AKAP11
1q21.1 del/dup
CACNA1G
HERC1
SLC39A8
C4
Odds ratio
100
50
20
10
5
2
1
1e-05
0.0001
0.001
0.01
0.05
0.1
0.2
0.5
Minor allele frequency in the general population

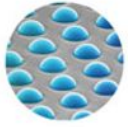
Section Overview

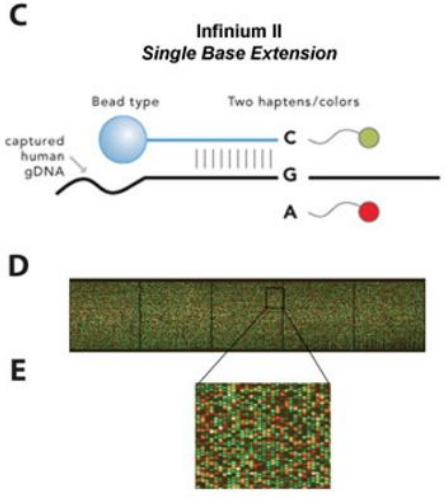
- Sequencing data and formats
  - Exomes and genomes
- What's in an "annotation"?
  - Constraint and functional annotation
- Family-based vs case-control designs
- Gene-based RVAS tests
- Burden heritability

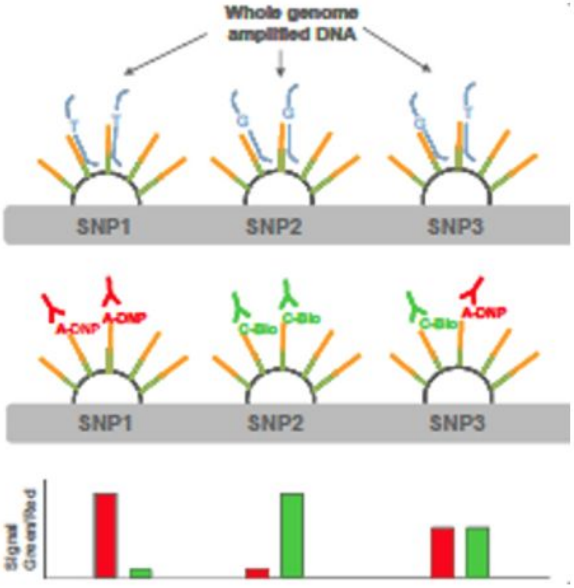
Sequencing data and formats

From sample to GWAS array

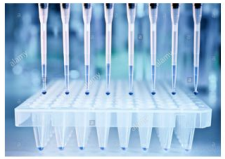


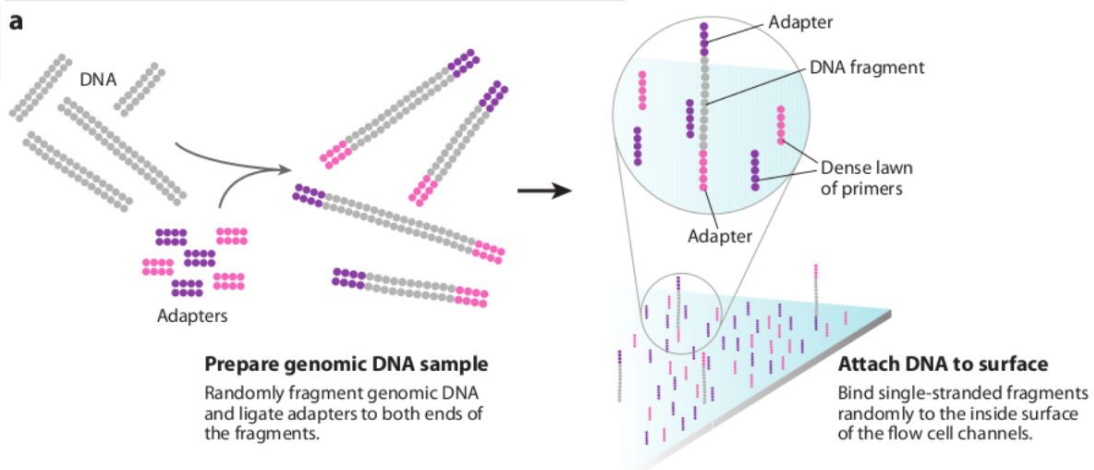


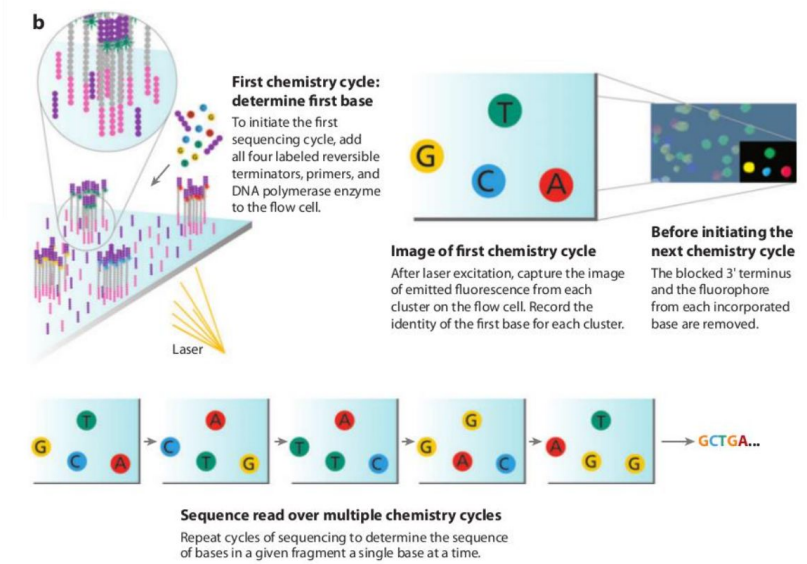




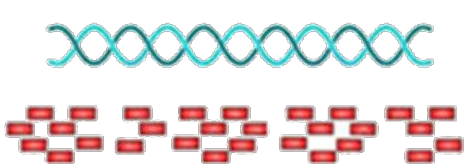
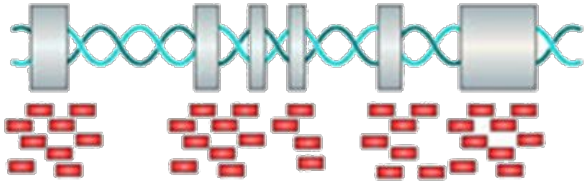
From sample to short-read sequencing







Exomes vs. Genomes

	Whole genome sequencing	Whole exome sequencing
Short reads		
Region	Untargeted	Only protein-coding regions
Features	Better sensitivity towards all variants, but especially indels and structural variants (SVs)	More cost-effective, and captures most of the likely impactful rare variation
	Captures non-coding variation	Misses non-coding variants, less effective at SVs

How is the exome “captured”?

“POND” = Human DNA fragments

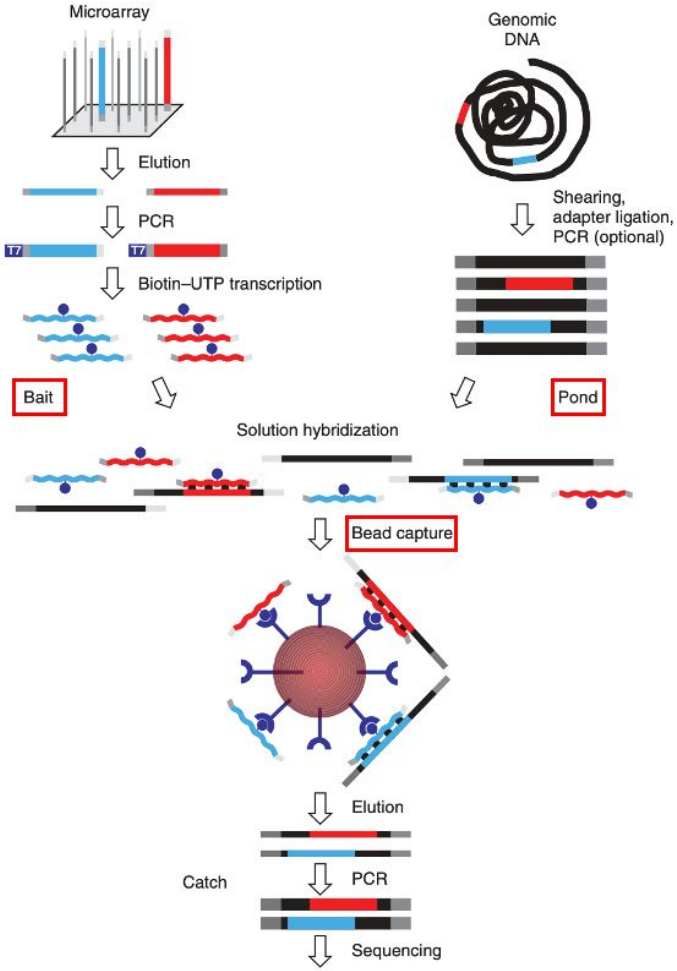
“BAIT” = Biotinylated RNA probes from microarray

“CAPTURE” = selecting DNA fragments with beads for sequencing

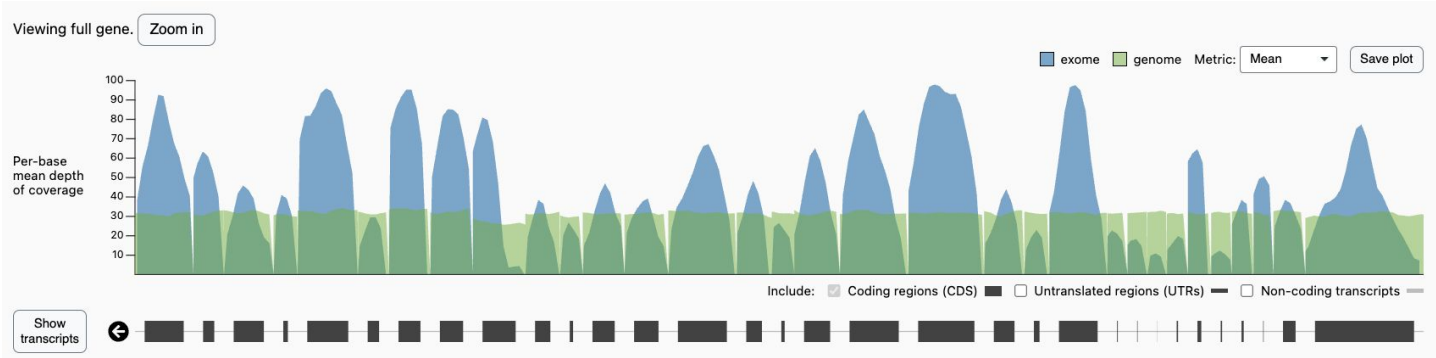
nature biotechnology

Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing

Andreas Gnirke[1], Alexandre Melnikov[1], Jared Maguire[1], Peter Rogov[1], Emily M LeProust[2], William Brockman[1,5], Timothy Fennell[1], Georgia Giannoukos[1], Sheila Fisher[1], Carsten Russ[1], Stacey Gabriel[1], David B Jaffe[1], Eric S Lander[1,3,4] & Chad Nusbaum[1]



NRXN1 gene



Key concepts

- Targeted vs untargeted coverage
- Read alignment
- Paired-end reading
- Insert size

FASTA vs FASTQ

FASTA	FASTQ
Stores nucleotide or protein sequences	Stores sequences and quality scores

`>header`
`AGCTTGA`
`AGCTTGA`
	`@header`
	`AGCTTGA`
	`+ !?5+ *`

Key Differences: • Stores sequences • 2 lines per record • Smaller size • Reference genomes	**Key Differences:** • Stores sequences + scores • 4 lines per record • Larger size • NGS data

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
      CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
      TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G     DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
      GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C     DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
      0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
    FORMAT      NA00001        NA00002        NA00003

20     14370   rs6054257 G       A       29    PASS    DP=14;AF=0.5;DB
    GT:GQ:DP 0|0:48:1 1|0:48:8 1/1:43:5
20     1110696 rs6040355 A       G,T     67    PASS    DP=10;AF=0.333,0.667;DB
    GT:GQ:DP 1|2:21:6 2|1:2:0   2/2:35:4
20     1230237 .         T       .       47    PASS    DP=13
    GT:GQ:DP 0|0:54:7 0|0:48:4 0/0:61:2
20     1234567 microsat1 GTCT    G,GTACT 50    PASS    DP=9
    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Header

Variant records

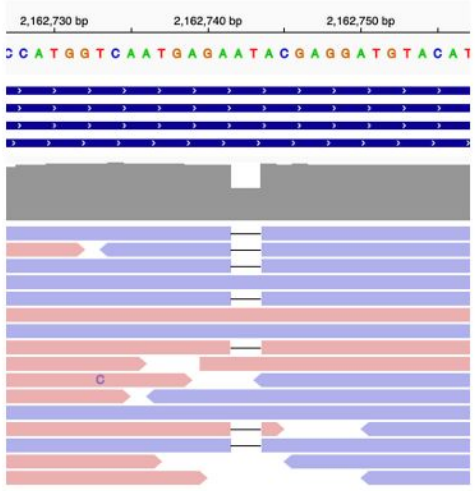
Sequencing data file formats

- FASTA/Q
  - unaligned sequence reads
- SAM / BAM / CRAM
  - Aligned reads to reference genome
- VCF
  - Variant positions only
  - Flexible file structure

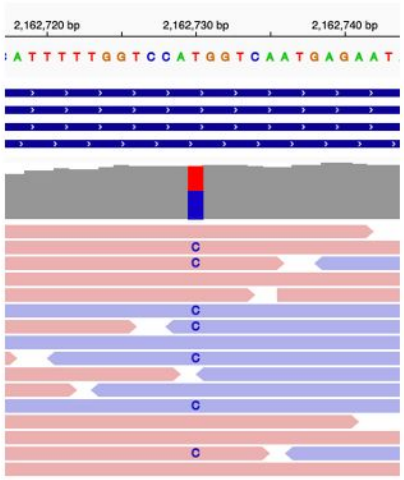
Sequence reads to variant calls

- From sequence reads to variant calls
  - FASTA/FASTQ = raw sequence “reads”
  - BAM/CRAM = Reads evaluated and aligned to reference genome
  - VCF = All genotypes where at least one sample does not match the reference

2 BP deletion



SNV



Variant Call Format (VCF)

Key concepts

- Header
- INFO field
- FORMAT description
- CHROM / POS line
- Genotype format
- GQ in Phred Likelihood scale

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

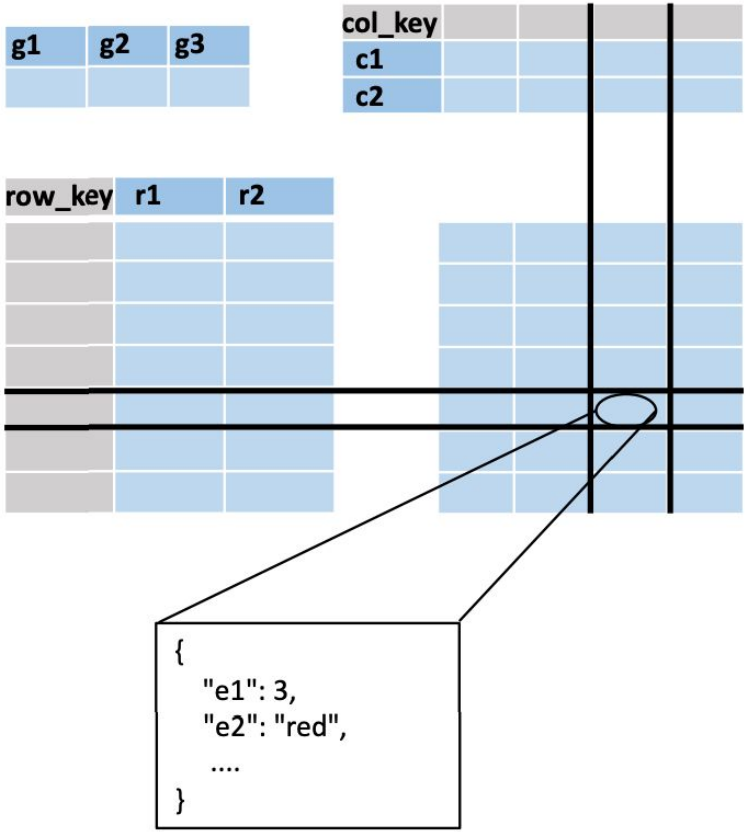
Links:

VCF spec pdf

NRXN1 variant level data

Hail Matrix table format (.mt)

	Hail Matrix Table	VCF
g	“global” annotations	FORMAT field
c	“column” annotations	Sample level fields
r	”row” annotations	variant level fields
e	“entry” fields	genotype level data
key	assigned field for sorting	Assumes chrom / position order

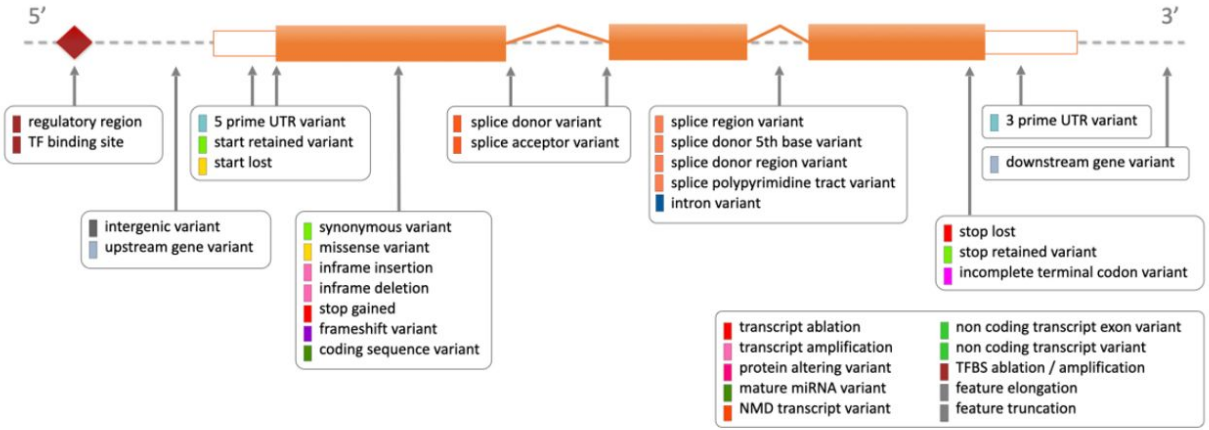


[Hail Matrix table cheat sheet]

Variant Annotation

What does “annotating” a variant mean?

- Annotation – **any** additional description of the genomic position or variant genotype
  - Tissue expression
  - Conservation across species
  - Chromatin accessibility
  - GWAS trait association
  - ***Functional effect on transcription***



Non-protein coding annotations

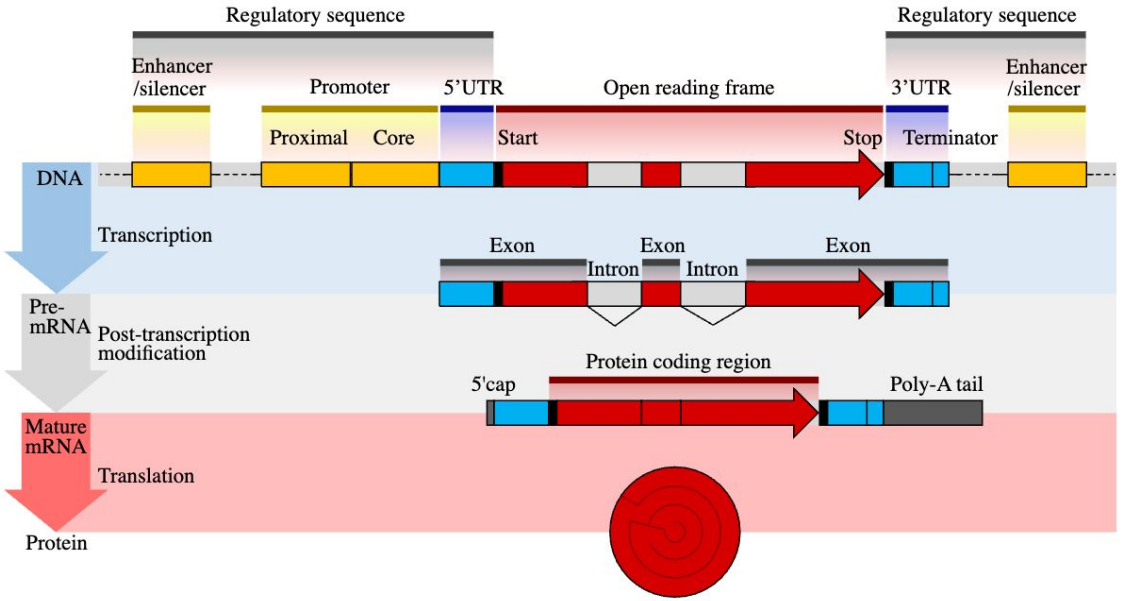
- Promoter
- Enhancer/silencer/repressor
- Transcription factor
  - Binding site
  - Modifier

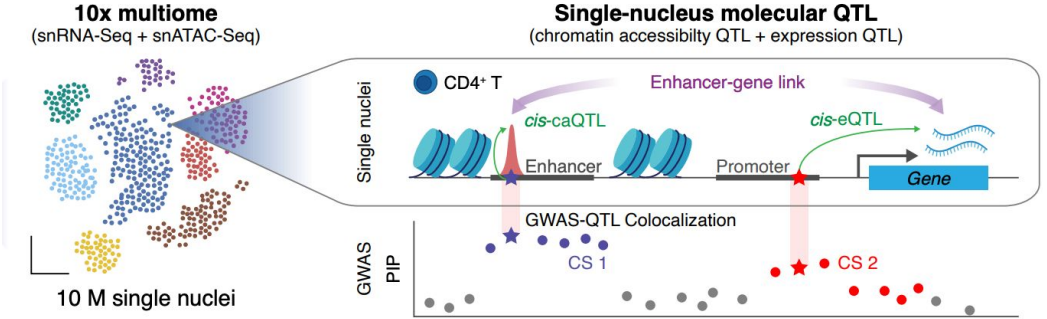
Epigenetic effects

- Histone modifiers
- Methylation

Quantifying non-coding variant effects

- QTL (quantitative trait locus)
- Gene Expression (eQTL)
- Alternative Splicing (sQTL)
- Chromatin Accessibility (caQTL)





Variant Effect Predictor (VEP) consequence terms

PTV
LoF
LGD

Missense

Splice region

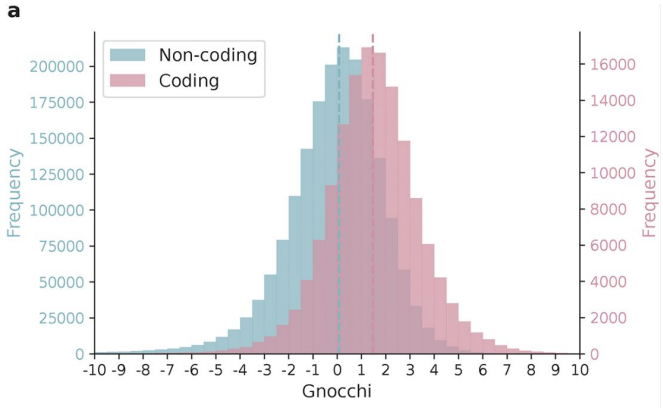
Synonymous

* SO term	SO description	SO accession	Display term	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	Inframe insertion	MODERATE
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	Inframe deletion	MODERATE
missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
splice_donor_5th_base_variant	A sequence variant that causes a change at the 5th base pair after the start of the intron in the orientation of the transcript	SO:0001787	Splice donor 5th base variant	LOW
splice_donor_region_variant	A sequence variant that falls in the region between the 3rd and 6th base after splice junction (5' end of intron)	SO:0002170	Splice donor region variant	LOW
splice_polypyrimidine_tract_variant	A sequence variant that falls in the polypyrimidine tract at 3' end of intron between 17 and 3 bases from the end (acceptor -3 to acceptor -17)	SO:0002169	Splice polypyrimidine tract variant	LOW
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	Start retained variant	LOW
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous variant	LOW

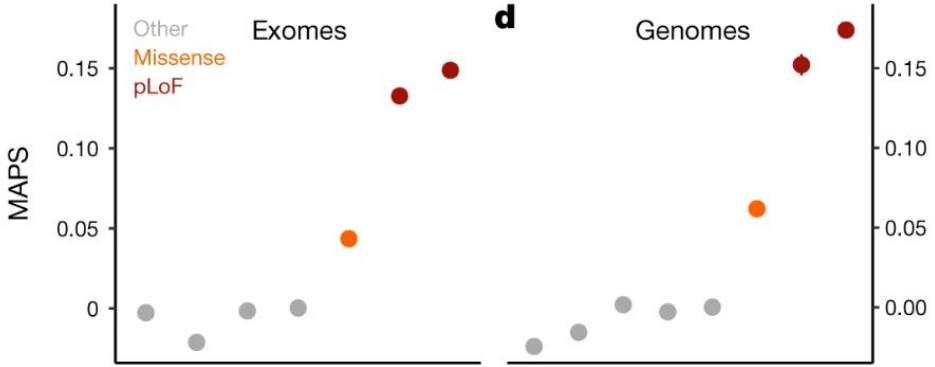
Annotation via constraint and predicted “pathogenicity”

- Cross-species conservation
  - phyloP
  - GERP
- Missense deleteriousness
  - Polyphen-2
  - REVEL
  - SIFT
- Within-species constraint
  - MPC
  - LOEUF / pLOF
- Machine learning aggregate scores
  - CADD
- Deep neural networks
  - AlphaMissense
  - popEVE
  - PrimateAI

NRXN1 gene



“MAPS” = Mutability adjusted proportion of singletons



Coding annotation broadly enriched for rare variant h^2 in UKB phenotypes

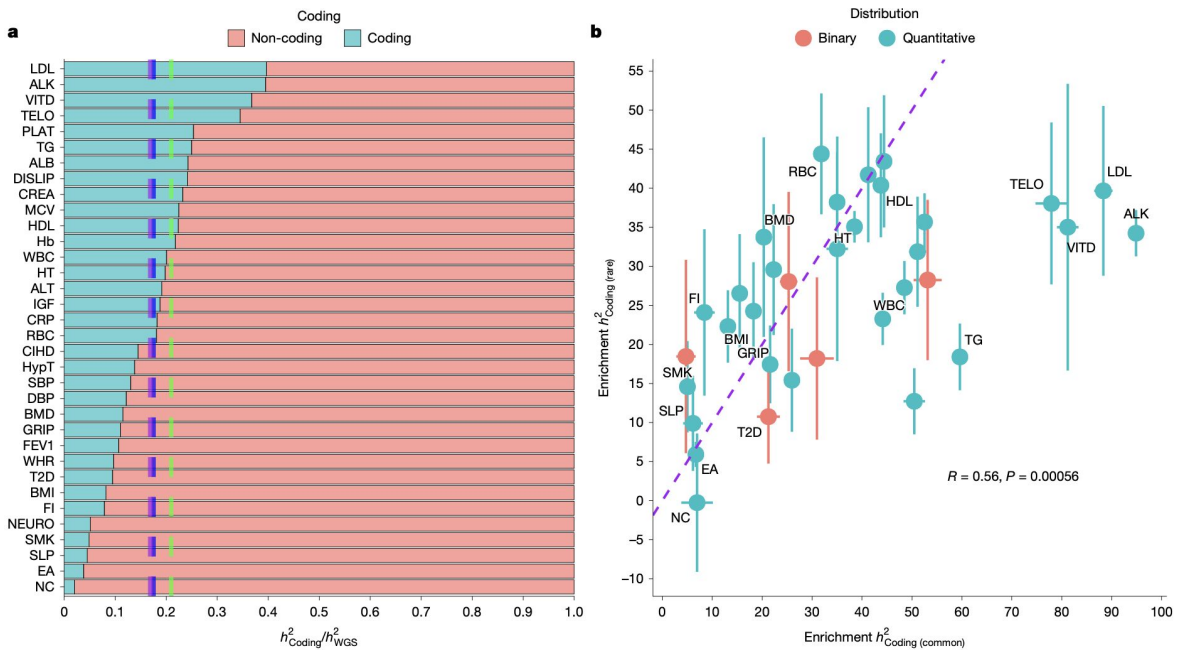


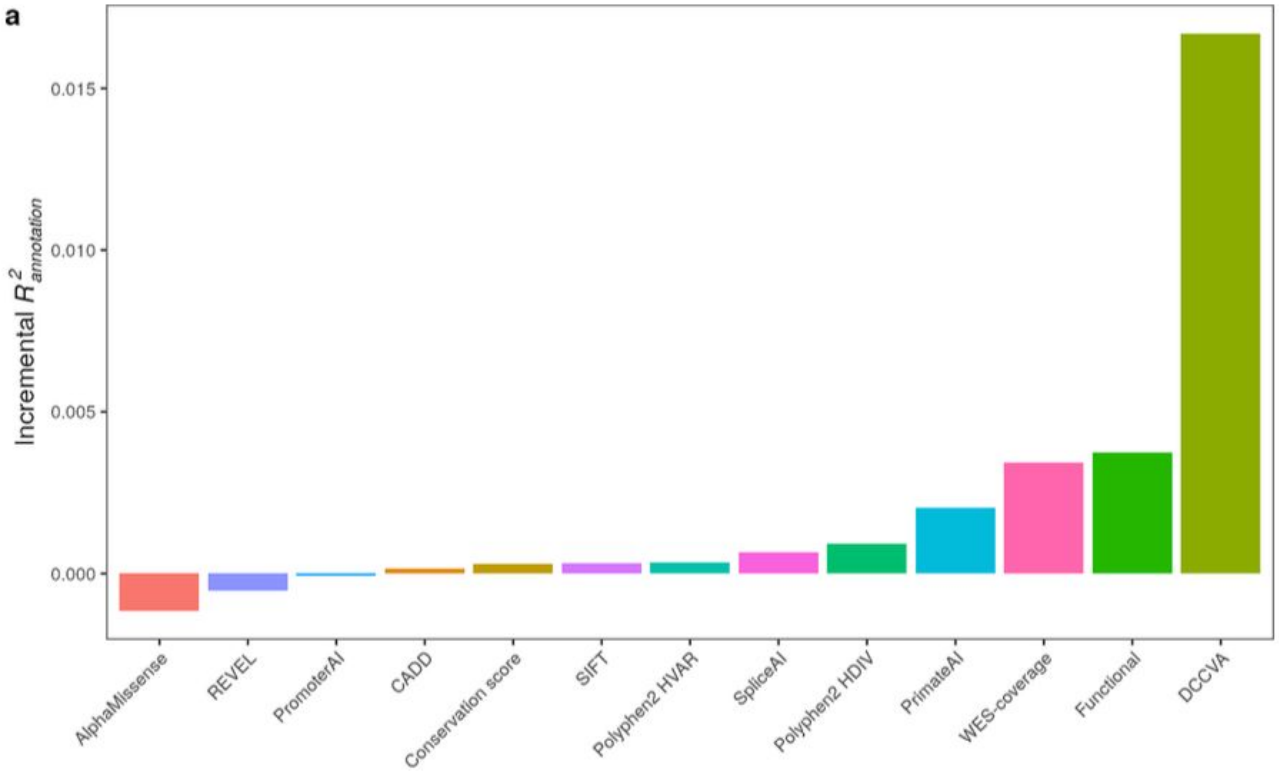
Fig. 2 | Relative contribution of coding and non-coding variants to WGS-based heritability. a, This panel represents, across 34 phenotypes, the ratio of proportion of phenotypic variance explained by coding variants (h^2_{coding}) over that explained by all WGS variants (h^2_{WGS}). The contribution to h^2_{WGS} from coding and non-coding variants were estimated jointly (Methods). The blue (and dashed) vertical line represents the mean of h^2_{Coding}/h^2_{WGS} across phenotypes. h^2_{Coding} was further partitioned into jointly estimated contributions from rare and common coding variants. The purple vertical line represents the mean across phenotypes of the ratio of phenotypic variance explained by common coding variants over that explained by all common variants. The green vertical line represents the mean across phenotypes of the ratio of phenotypic variance explained by rare coding variants over that explained by all rare variants. **b**, This panel compares heritability enrichment in coding variants between common variants (*x* axis) and rare variants (*y* axis). Error bars represent s.e.s. The correlation between heritability enrichment for common and rare variants was calculated using a Pearson's correlation coefficient (R) over $n = 34$ traits. The P value measuring the significance of that correlation is denoted as P in the bottom-right corner of the panel and is based on a two-sided Pearson's correlation test.

Heritability enrichment in coding variants

We partitioned $\hat{h}^2_{WGS}$ to assess the relative contribution of coding and non-coding variants to trait heritability. Specifically, we focused on coding variants within loci covered by WES technologies. We identified these WES loci using the Resource field 3803 (based on IDT xGen Exome Research Panel v.1.0 and 100 bp flanking region upstream and downstream of each capture target). In total, 408,096 (that is, 1% of all WGS variants with MAF > 0.01%) variants were included in the WES-covered regions and 40,167,108 variants were not. We used the Nirvana pipeline

Wainschtein et al., 2025

Yet constraint annotations less likely than GWAS hits explains rare variant h^2 in UKB phenotypes



DCCVA

Distance to Closest Common Variant Association

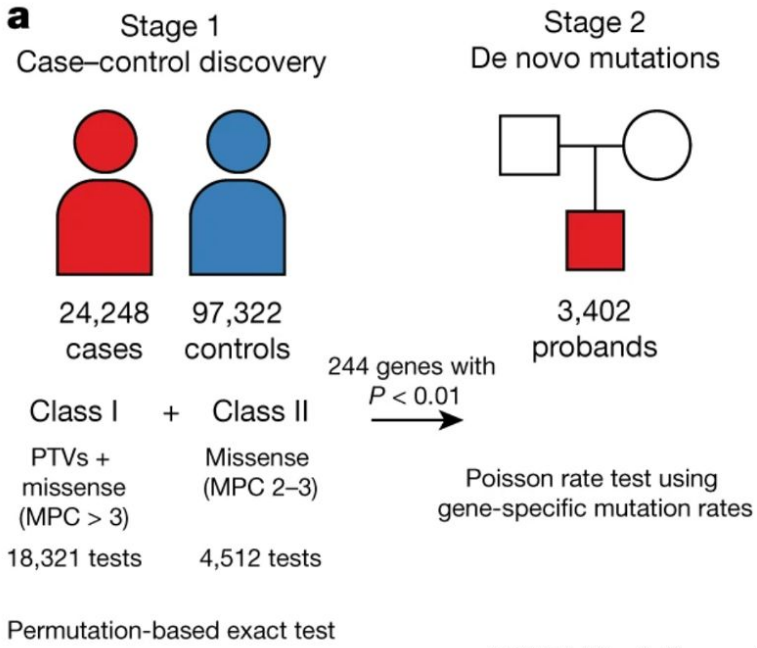
Wainschtein et al., 2025

Case-control vs family-based designs

Case-control vs family-based designs

Case-control design

- Common and rare variants
- Complex traits
- Population-based modeling



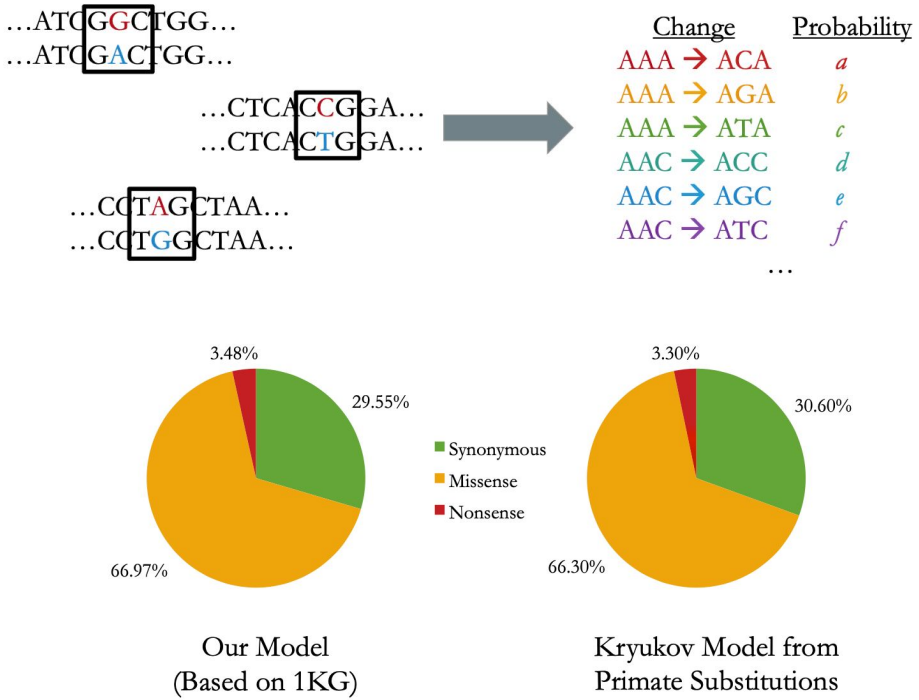
Family-based design

- *De novo* mutation
- Mendelian disease
- Transmission tests
- Robust to population stratification

SCHEMA v1 figure by TJ Singh

Using the *de novo* mutation rate as our expectation

- Modeling a fixed expectation for each annotation class
- More powerful than control rates, particular among rare observations
- Sensitive for testing gene recurrence as well (2 or more *de novo* mutations)

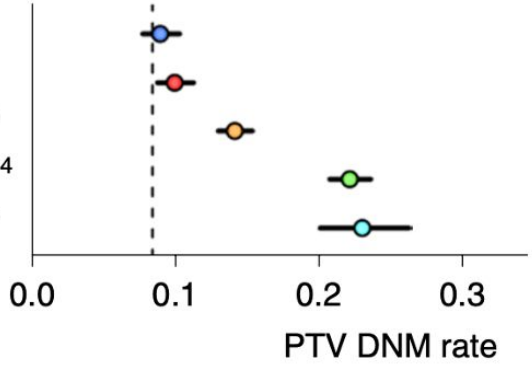


Family-based testing for an exome-wide burden of *de novo* PTVs

Exome-wide burden of PTVs in disease-affected probands

- For trios, testing PTV rate against the mutational model (as opposed to control trios)
- Poisson-rate test (enrichment is the rate-ratio)

Disease	Trios	DNM rate	Enrichment	*P* value
Control	2,182	0.09	1.06	0.378
Schizophrenia	2,541	0.1	1.18	0.01
Autism spectrum disorder	3,982	0.14	1.67	1.4×10^{-29}
Developmental delay	4,293	0.22	2.61	2.5×10^{-144}
Intellectual disability	971	0.23	2.72	7.9×10^{-38}

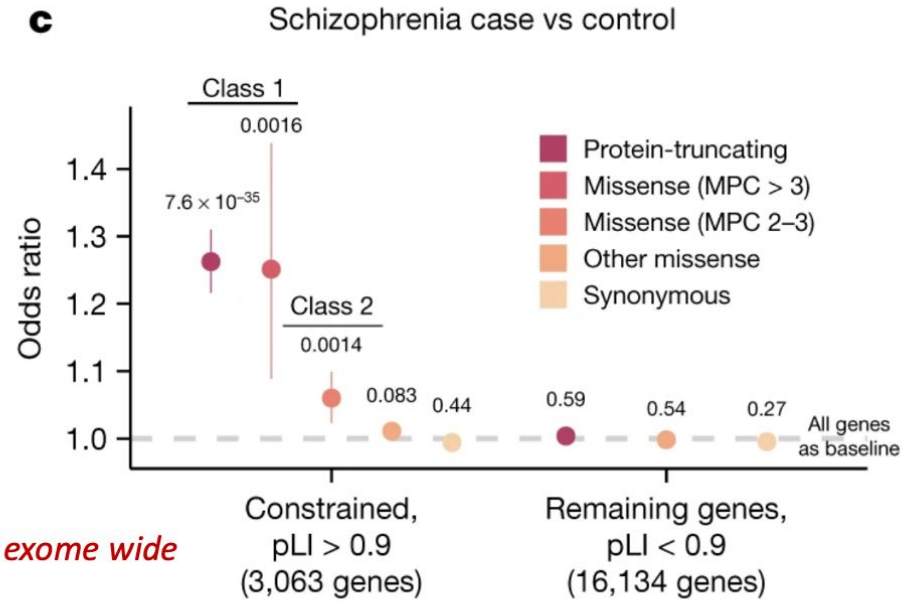


Both family and case-control designs test for burden

Exome-wide burden of ultra-rare PTVs in 24k SCZ cases

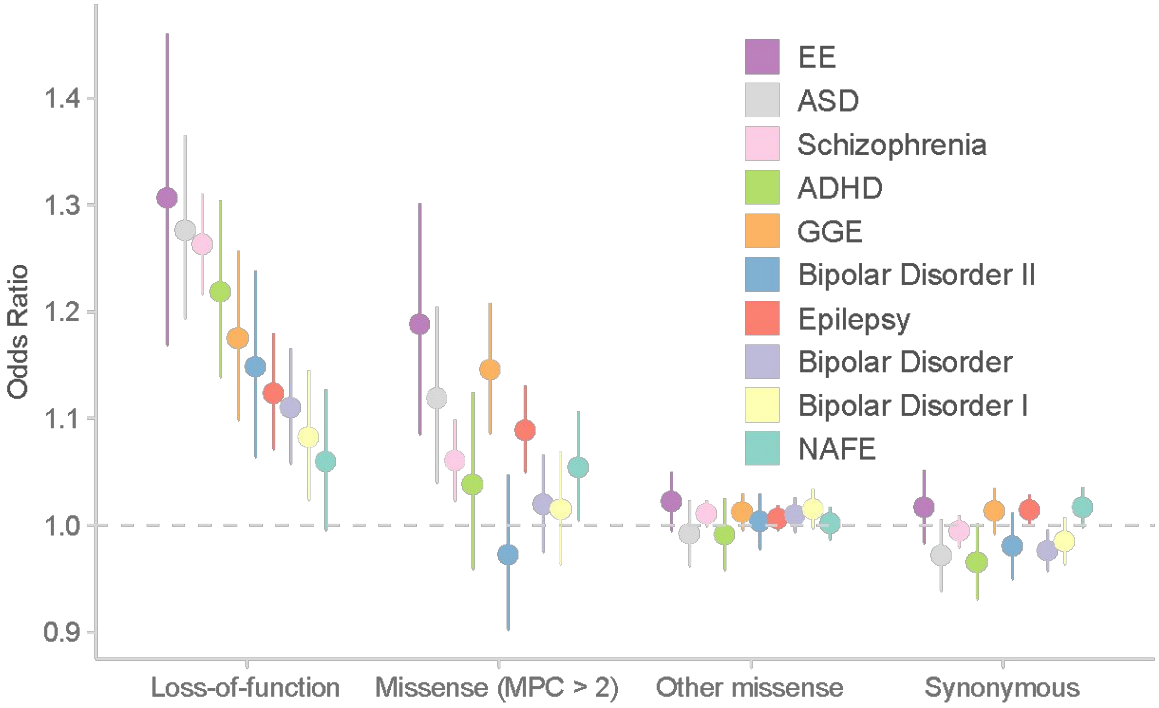
- Testing cases against controls
- "Ultra-rare" = not variant in gnomAD*
- Logistic regression model with covariates
  - Principal components
  - Total rare variant count
- Meta-analyzed across continental ancestries
- Synonymous burden as "quasi" control rate

* *gnomAD excluding psych disease participants*



Technically a gene set test, not exome wide

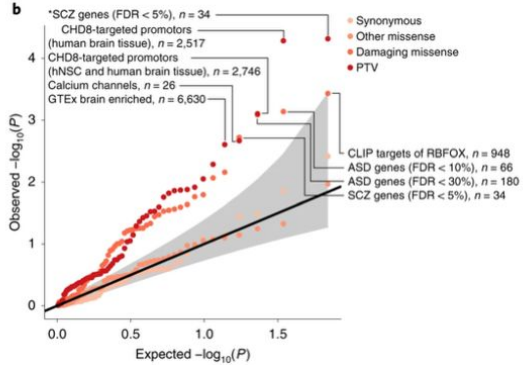
Enrichment of variants in constrained genes across a variety of neuropsychiatric illnesses



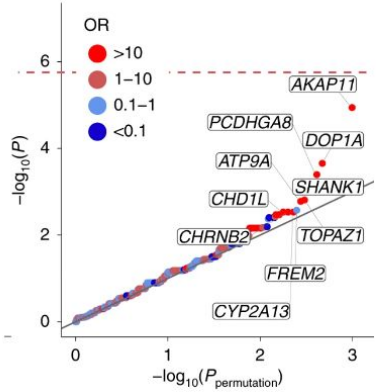
Excess burden applies across all levels of inquiry

- Exome-wide burden
  - Stratify by allele frequency/count
  - Stratify by functional annotation
  - **AIM: narrow the search space of signal**
- Gene-set burden
  - Replicating candidate gene sets vs open search (GO terms)
  - Permutation methods account for gene size/redundancy
- Single gene burden
  - If ultra-rare variation, often not enough variants to run regression model
  - Fishers exact test

Bipolar exome gene sets



Bipolar exome genes

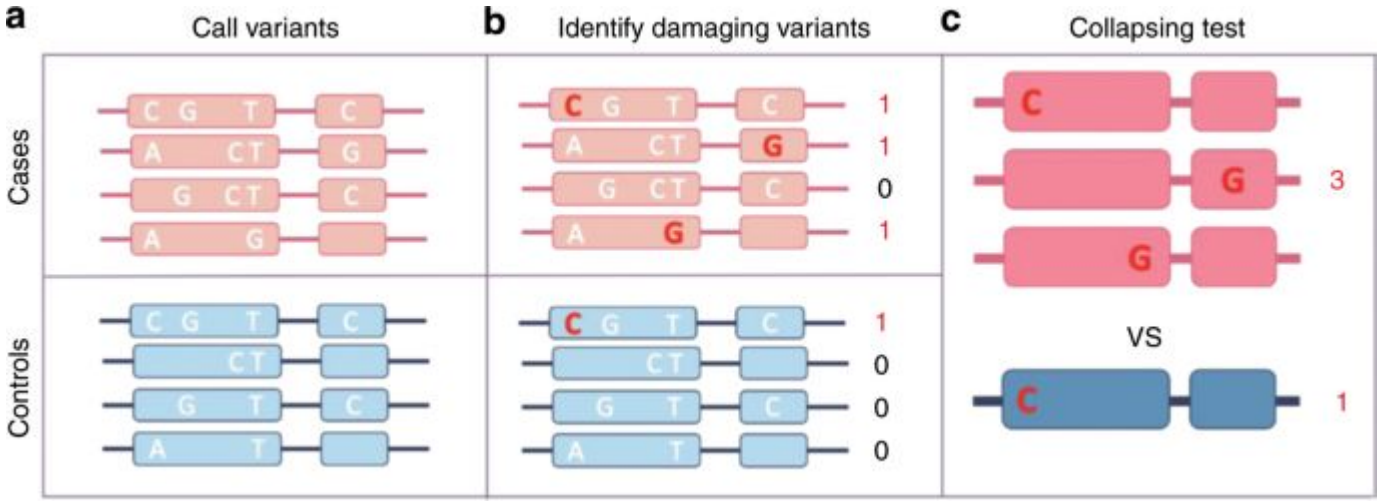


Gene-based RVAS tests

RVAS = Rare Variant Association Tests

Gene-based burden testing

Often we determine which variant annotations are “damaging” on the exome-wide / constraint gene-set burden



* Cirulli et al. 2020

GRIN2A gene in SCHEMA

24.2K cases
97.3K controls
3.4K trios

Gene Result

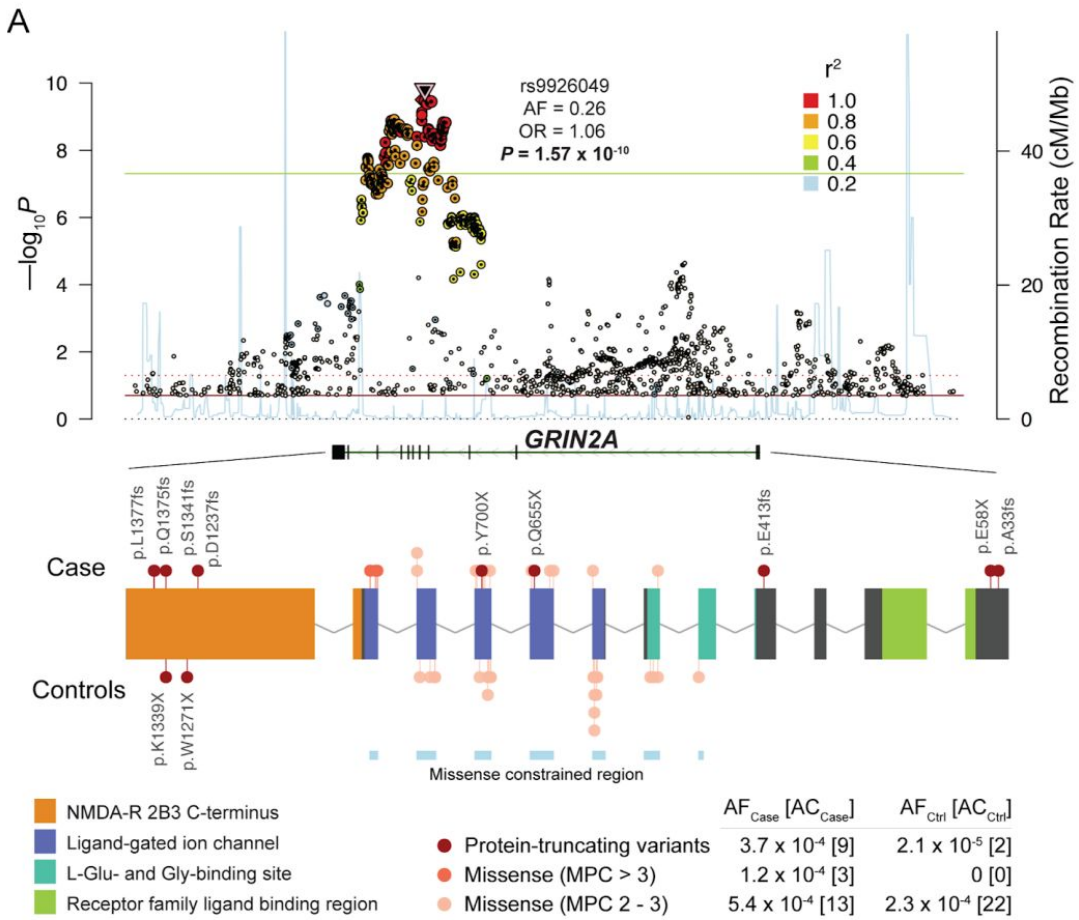
Class	Consequence	Cases	Controls	Odds Ratio	Case/Control *P*-value
I	PTV	9	2	24.1	0.00000718
	Missense (MPC ≥ 3)	3	0		
II	Missense (3 > MPC ≥ 2)	13	22	2.37	0.0142

Meta-analysis *P*-value: 7.37e-7

Meta-analysis *Q*-value: 0.00167

Burden test used

- Fisher's Exact Test (2 x 2)
- Cochran-Mantel-Haenzel (CMH) Test (2 x 2 x N)



Many collapsing tests proposed
Basic 2 x 2 burden test is still most widely used

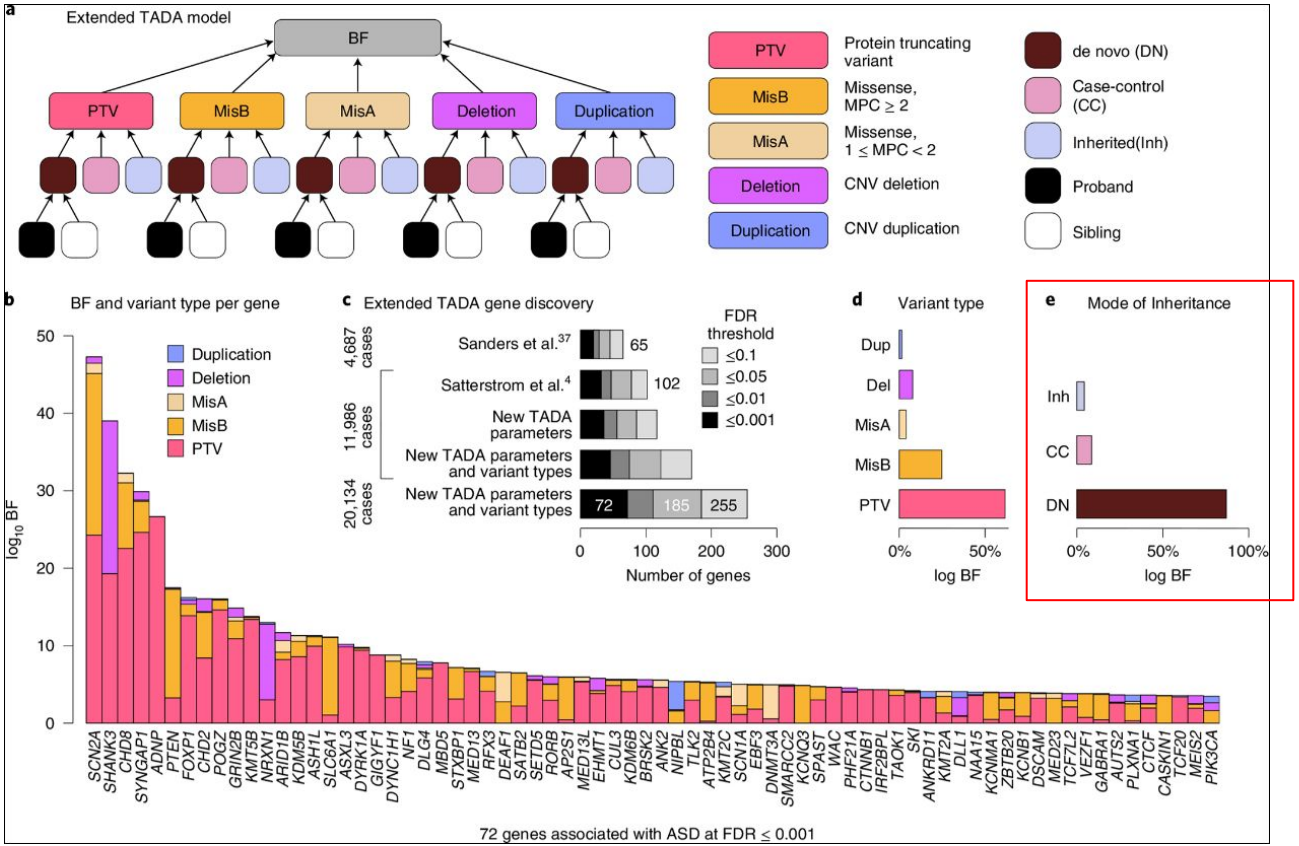
Type	Examples	Description	Strengths	Weaknesses
Burden test	ARIEL test, CAST, CMC method, MZ test, WSS	Collapse rare variants into genetic scores	Powerful when a large proportion of variants are causal and effects are in the same direction	Lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants
Adaptive burden test	aSum, Step-up, ER EC test, VT, KBAC method, RBT	Use data-adaptive weights or thresholds	More robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	Computationally intensive; VT requires the same assumptions as burden tests
Variance components test	SKAT, SSU test, C-alpha test	Test variance of genetic effects	Powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	Less powerful than burden tests when most variants are causal and effects are in the same direction
Combined test	SKAT-O, Fisher method, MiST	Combine burden and variance-component tests	More robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	Slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive
Other	RUNNER, DeepRVAT	Not test on region-based aggregated score	More flexible than traditional aggregated tests	Haven't been widely tested, less straightforward to interpret

*Lee et al.

Using a bayesian framework (*TADA*) with well-powered sample sizes and reliably specified parameters

Autism gene discovery

Fu, Nat Gen 2022



Burden Tests

Compare the difference in the number of individuals carrying variants in a given gene/region between disease case and control cohorts.

- Demonstrate more power in the presence of a high proportion of causal variants with the effects in the same direction

$$h(\mu_i) = \alpha_0 + \alpha' X_i + \beta' G_i \Rightarrow H_0: \beta = 0$$

μ_i - continuous traits
$logit(\mu_i)$ - binary traits

$$Q_B = \left(\sum_{j=1}^{m} w_j \sum_{i=1}^{n} G_{ij} (y_i - \hat{\mu}_i) \right)^2 \sim \chi^2_{df=1}$$

A threshold indicator or weight for variant j

Allele count of variant j in sample i

Adaptive Burden Tests

Use data-adaptive weights or thresholds on variants to adjust for traditional burden tests.

	VT test*	EREC*
Feature	Find the optimal MAF threshold of rare variants by varying the threshold	Estimates a regression coefficient of each variant and uses them as the weight
Pre-requisites	Requires similar assumptions to those of the burden tests; i.e., it requires a majority of rare variants under the optimal threshold to be causal and have effects in the same direction	Requires estimation of regression coefficients, which are difficult to estimate stably for rare variants
Strengths	Improve the power compared to the burden tests	
Weaknesses	Computationally intensive when applied to large-scale sequencing studies Because they rely on a large number of permutation or bootstrap samples to compute p values and are difficult to control for covariates, such as population stratification	

* Price et al. * Lin et al.

Variant Component Tests / Kernel-based Tests

Use the **variance** of genetic effects and demonstrate more power in the presence of variants with **diverse** functional directions or a smaller proportion of causal variants

C-alpha*

- The C-alpha test statistic T contrasts the variance of each observed count with the expected variance, assuming the binomial distribution.
- It is sensitive to risk and protective variants in the same gene or pathway.

$$T = \sum_{i=1}^{m} [(y_i - n_i p_0)^2 - n_i p_0 (1 - p_0)]$$

The proportion of cases out of total samples, e.g. 0.5

$y_i \sim binomial(n_i, p_i)$
$H_0: p_i = p_0$
$H_1: p_i \sim \{p_i < p_0$ (protective) $| \; p_i > p_0$ (detrimental) $| \; p_i = p_0$(neutral)$\}$

* Neale et al.

SKAT*

- SKAT aggregates the associations between variants and the phenotype through a kernel matrix and can allow for SNP-SNP interactions, i.e., epistatic effects.
- SKAT is especially powerful when a genetic region has both protective and deleterious variants or many noncausal variants.

$$Q_s = \sum_{j=1}^{m} w_j^2 \left\{ \sum_{i=1}^{n} g_{ij} (y_i - \hat{\pi}_i) \right\}^2$$

The score statistics for testing $H_0: \beta_j = 0$ in the single-variant test for SNP j

* Wu et al.

Combined Tests - e.g., SKAT-O*

Take the benefit of both burden and variance-component tests to construct association tests that are more robust with respect to the proportion of causal variants and the functional direction of variants

A unified test that includes burden tests and SKAT in one frame- work. In particular, the test statistic of the proposed unified test is

$$Q_\rho = \rho Q_B + (1-\rho) Q_S, 0 \le \rho \le 1$$

Q_B: *Burden score statistics*

Q_S: *SKAT statistics*

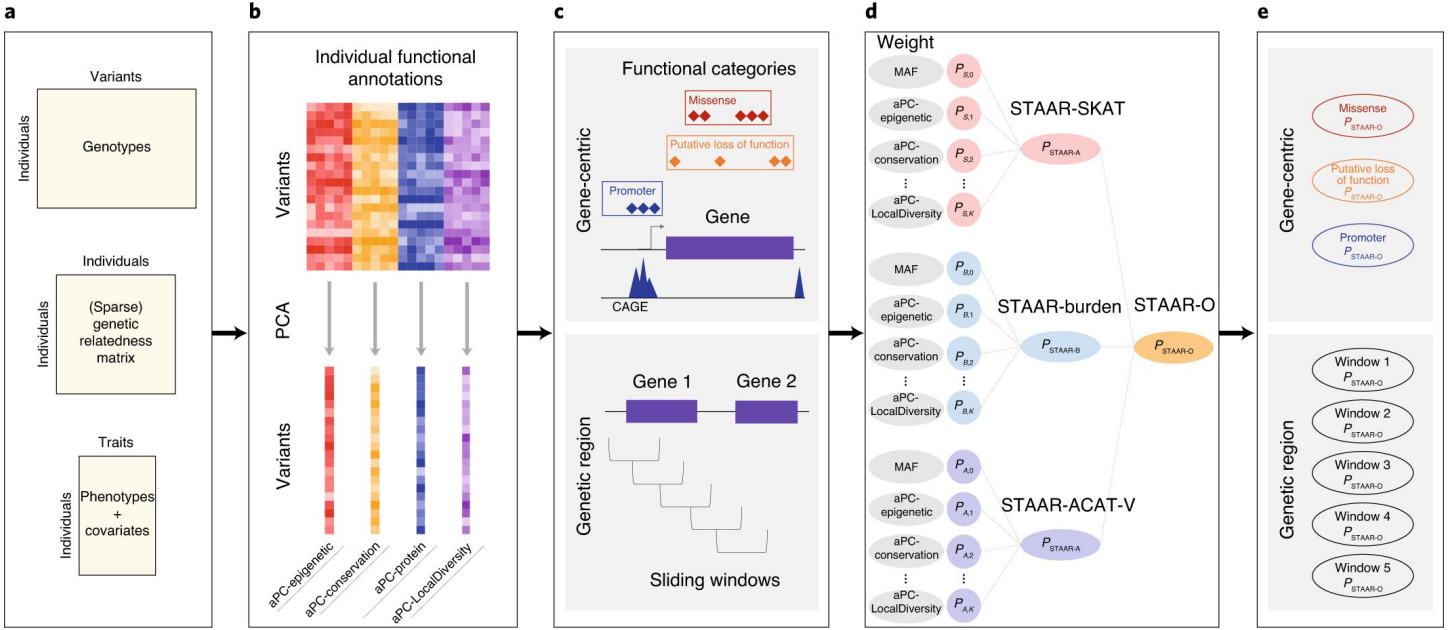
Strengths
Computationally efficient
Easy to adjust for covariates for population stratification, e.g., age, gender, PC
Maximize power
Adjust for small sample size and dichotomous phenotype

* Lee et al.

Novel Test #1: STARR

A scalable and powerful rare variant association test method that dynamically incorporates both qualitative functional categories and quantitative complementary annotation scores using a unified omnibus multidimensional weighting scheme

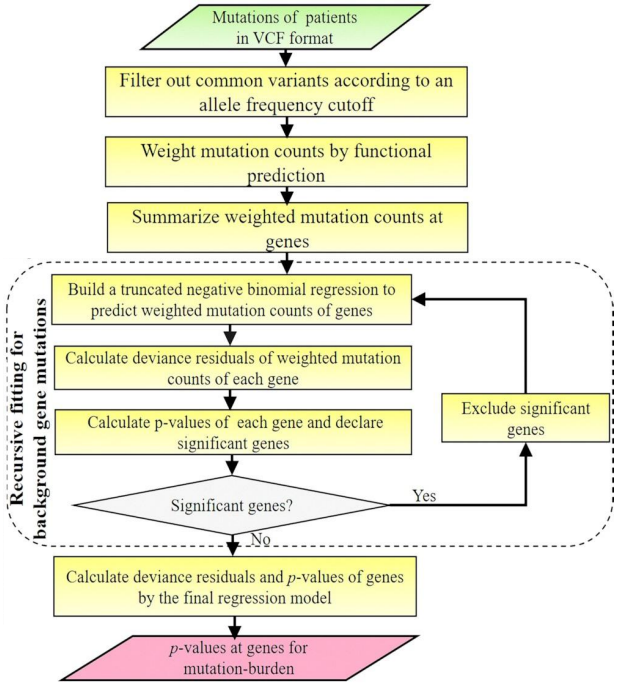
- Functional annotation
- Population structure
- Relatedness
- Scalable



* Li et al.

Novel Test #2: RUNNER*

Rare variants association test based on the of the observed mutation burden of a gene in cases from a baseline predicted by a weighted recursive truncated negative-binomial regression

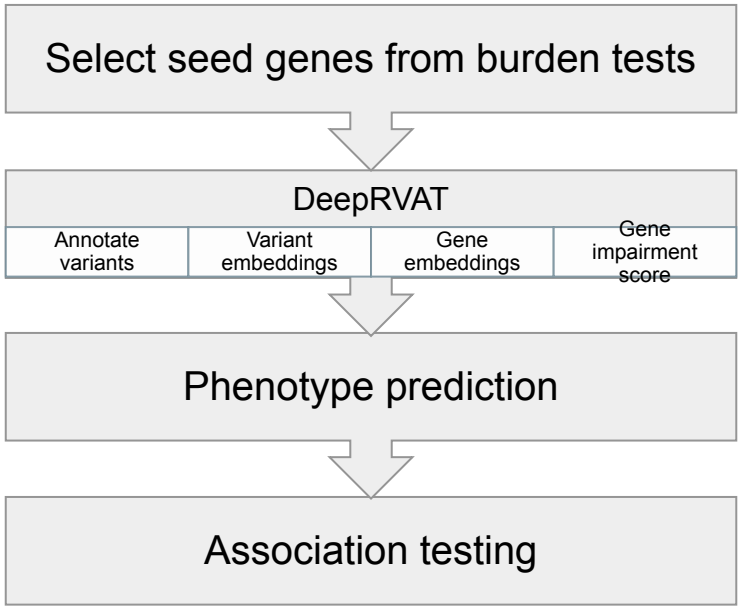


Strengths	Limitations
Insensitive to population structure and relatedness	Not adjusted for confounders (sex, age)
Runtime insensitive to sample size	Difficult to select reference population for admixed pop
Independent of mutation types (coding & splicing)	Not considering non-coding region
More powerful and more accurately reflecting mutation burden	Decreased power for highly polygenic diseases

*Jiang et al.

Novel Test #3: DeepRVAT

Data-driven framework that uses deep neural networks to learn a flexible rare variant aggregation function. DeepSet networks to efficiently model variant effects and interactions



Strengths:

1) Learn variant effects without strong filtering or specifying a kernel
2) Model nonlinear and epistatic effects
3) Efficiently incorporate dozens of multi-modal variant annotations
4) Provide trait-specific burden scores
5) Utilize GPUs for biobank-scale analyses.

Exome-wide significance

Bonferroni correction

$0.05 * N_{genes} * N_{indep.\ annotations}$

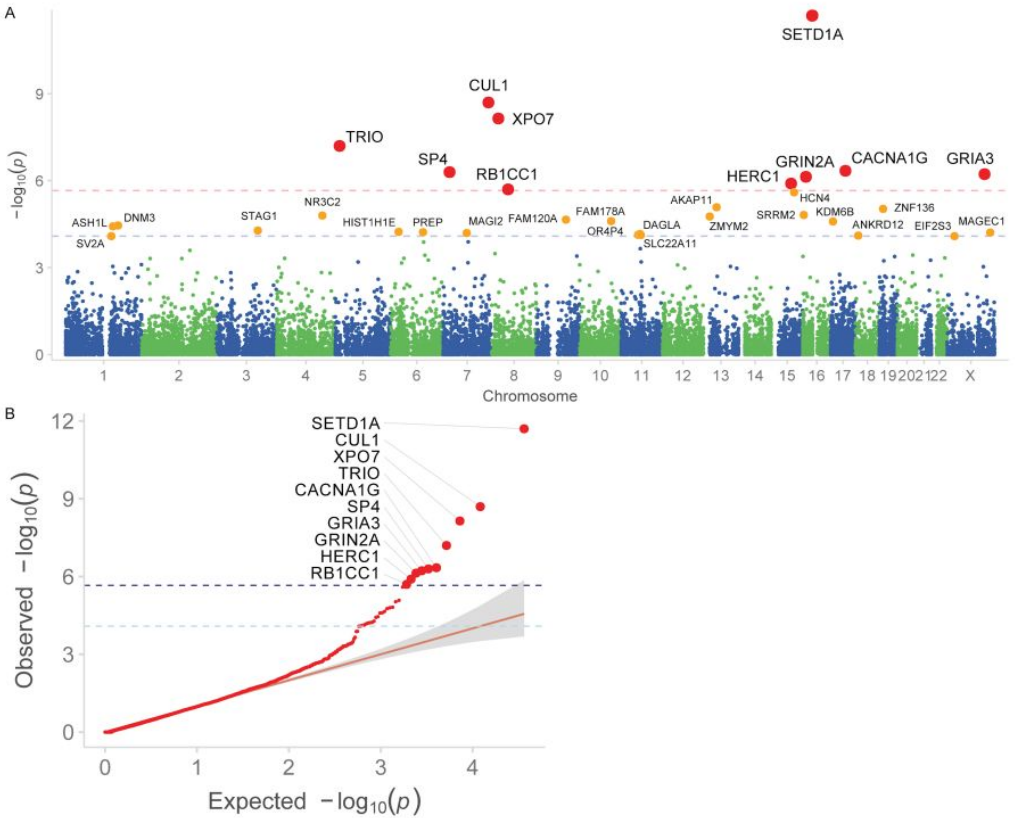
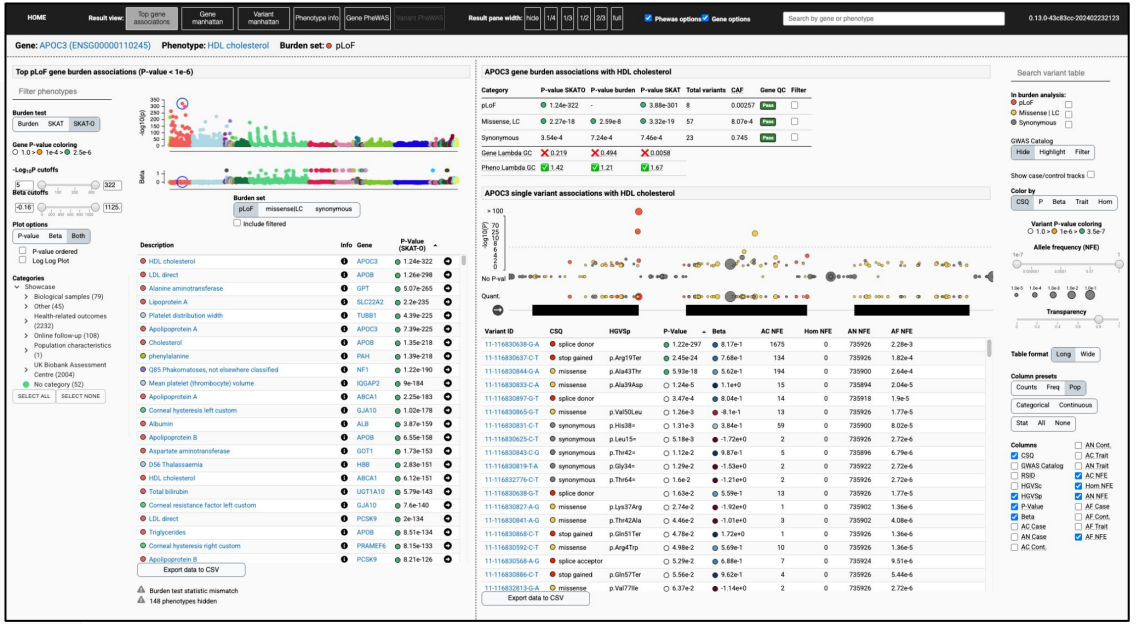


Figure 2. Results from the meta-analysis of ultra-rare coding variants in 3,402 trios, 24,248 cases, and 97,322 controls.

A: Manhattan plot. $-\log_{10}$ *P*-values are plotted against the chromosomal location of each gene. Genes reaching exome-wide significance are in red, and genes significant at FDR < 5% are in orange. Red dashed line: $P = 2.14 \times 10^{-6}$; Blue dashed line: FDR < 5%, or $P = 8.23 \times 10^{-5}$.

B: *Q*-*Q* plot. Observed $-\log_{10}$ *P*-values are plotted against expectation given a uniform distribution. Genes reaching exome-wide significance are plotted with a larger size. The direction of effect is indicated by the color of each point. Dark blue dashed line: $P = 2.14 \times 10^{-6}$; Light blue dashed line: FDR < 5%.

https://app.genebass.org/



genebass

gene-based association summary statistics

Search by gene or phenotype

Browse

Dataset: 394,841 exomes
Release date: June 7, 2022
Reference genome: GRCh38
Browser: 0.13.0-43c83cc-202402232123

Genebass is a resource of exome-based association statistics, made available to the public. The dataset encompasses 4,529 phenotypes with gene-based and single-variant testing across 394,841 individuals with exome sequence data from the UK Biobank. Genebass was developed by the following organizations which provided funding and guidance:

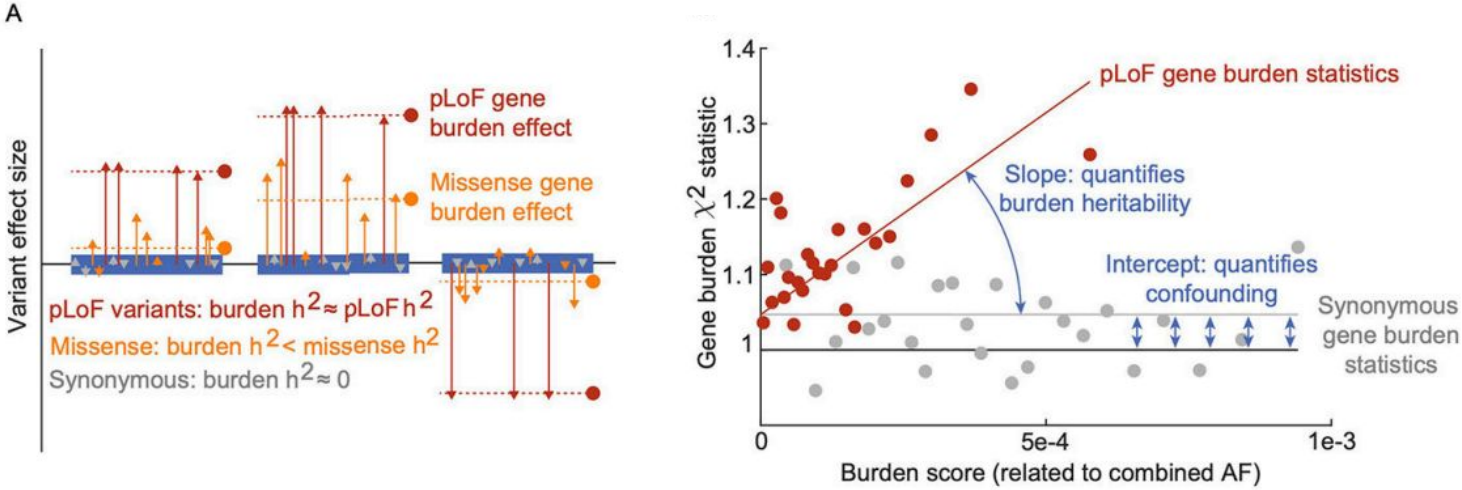
abbvie Biogen Pfizer BROAD INSTITUTE

Burden heritability

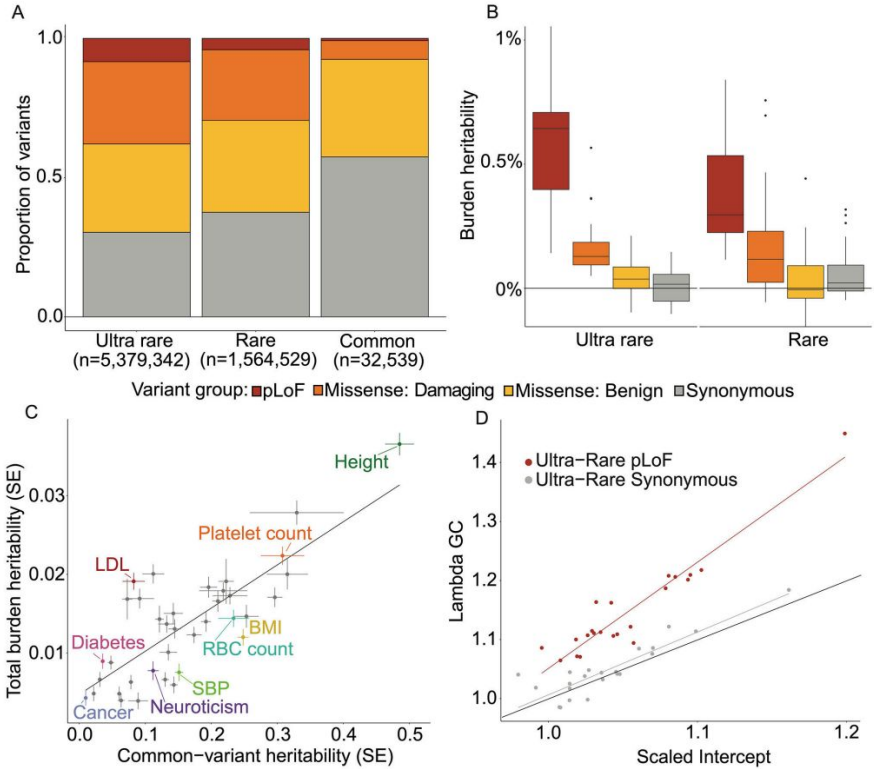
How much trait variance can be explained by ultra-rare variants?

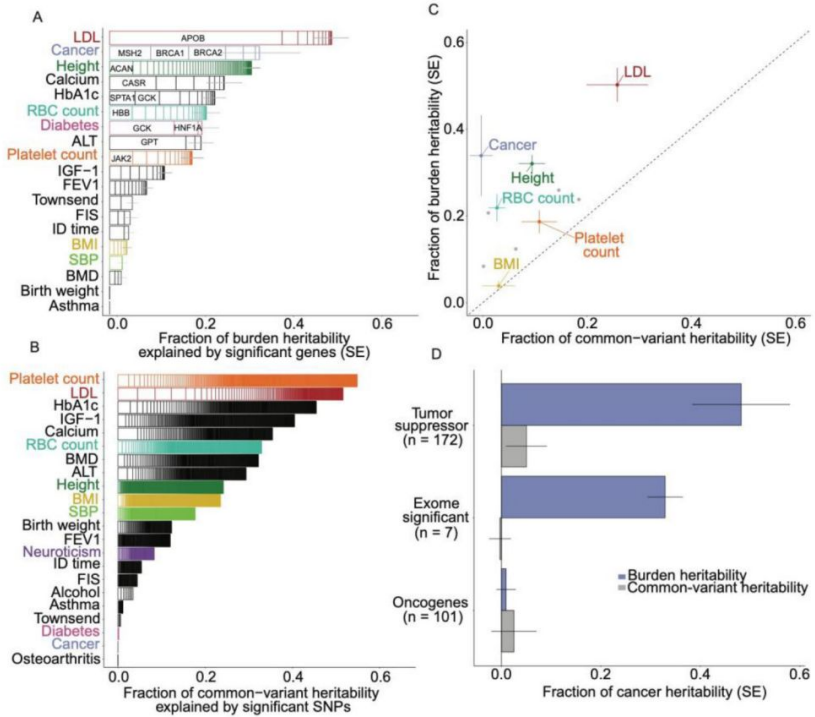
Polygenic architecture of rare coding variation across 394,783 exomes

Daniel J. Weiner, Ajay Nadig, Karthik A. Jagadeesh, Kushal K. Dey, Benjamin M. Neale, Elise B. Robinson, Konrad J. Karczewski & Luke J. O'Connor



UK Biobank burden heritability regression results



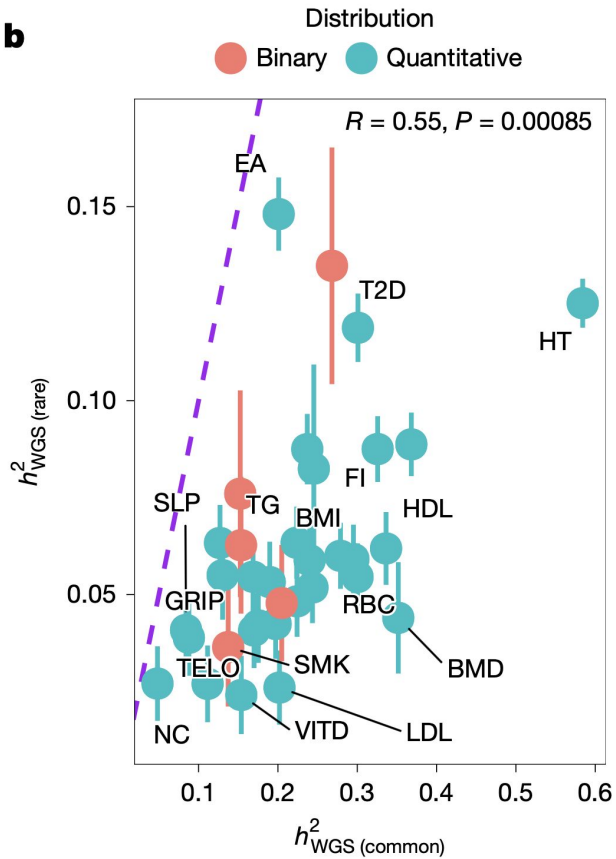


Addressing the difference with other estimates of rare variant heritability

Two recent papers reported that rare variants from whole-genome sequencing data are an important source of heritability for complex traits. Wainschtein et al.[24] reported the heritability explained by rare and low-frequency variants (MAF=1e-4 - 0.01) is 0.3 (SE 0.1) for height and 0.29 (SE 0.25) for BMI; Jang et al.[29] similarly reported that rare variants explain a large fraction of heritability for smoking phenotypes, with large standard errors. Unlike our burden estimates, these estimates include noncoding SNPs and SNPs at intermediate allele frequencies (0.001-0.01), and they do not aggregate variants by gene. Because of these differences, our rare variant heritability estimates are smaller but much better powered: 0.037 (SE=0.001) for height, 0.012 (SE=0.001) for BMI, and 0.006 (SE=0.001) for smoking status, respectively (Supplementary Tables 8 and 11).

Addressing the difference with other estimates of rare variant heritability

Two recent papers reported that rare variants from whole-genome sequencing data are an important source of heritability for complex traits. Wainschtein et al.[24] reported the heritability explained by rare and low-frequency variants (MAF=1e-4 - 0.01) is 0.3 (SE 0.1) for height and 0.29 (SE 0.25) for BMI; Jang et al.[29] similarly reported that rare variants explain a large fraction of heritability for smoking phenotypes, with large standard errors. Unlike our burden estimates, these estimates include noncoding SNPs and SNPs at intermediate allele frequencies (0.001-0.01), and they do not aggregate variants by gene. Because of these differences, our rare variant heritability estimates are smaller but much better powered: 0.037 (SE=0.001) for height, 0.012 (SE=0.001) for BMI, and 0.006 (SE=0.001) for smoking status, respectively (Supplementary Tables 8 and 11).



BurdenNOT ML

Heritability and effect-size distribution of rare protein-coding variation

Wenhan Lu

Senior Computational Associate

Karczewski, Neale and O'Connor Labs

Broad Institute of MIT and Harvard

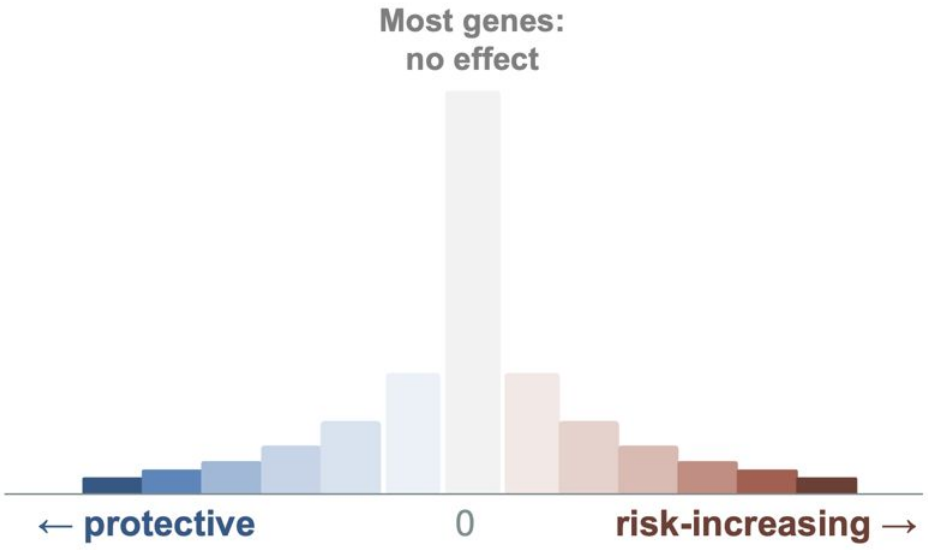
Session: Population and Statistical Genetics I

Probabilistic Modeling of Genomics - March 27, 2026

Burden effect size distribution

How gene-level effect sizes (β) are distributed across all coding genes in the genome.

Addresses key questions about the rare variant genetic architecture of a trait



Heritability
How much trait variation is explained by rare variants?

Polygenicity
Is the trait driven by many genes or a few?

Symmetry
How does selection shape effect direction across traits?

Power & Replication
How large a study do we need to saturate discovery?

Goal: estimate the distribution of burden effect sizes β

Input

Rare variant
(MAF < 1 x 10^{-3})
association statistics

← negative effects | null | positive effects →

-k	...	-2	-1	0	1	2	...	k

Model

Conditional Likelihood

- **Normal**
- **Beta-Binomial**

$$\beta_g \sim w_1 f_1 + \ldots + w_K f_K$$

Mixture distribution

f_k Components - Fixed

w_k Weights - Learned

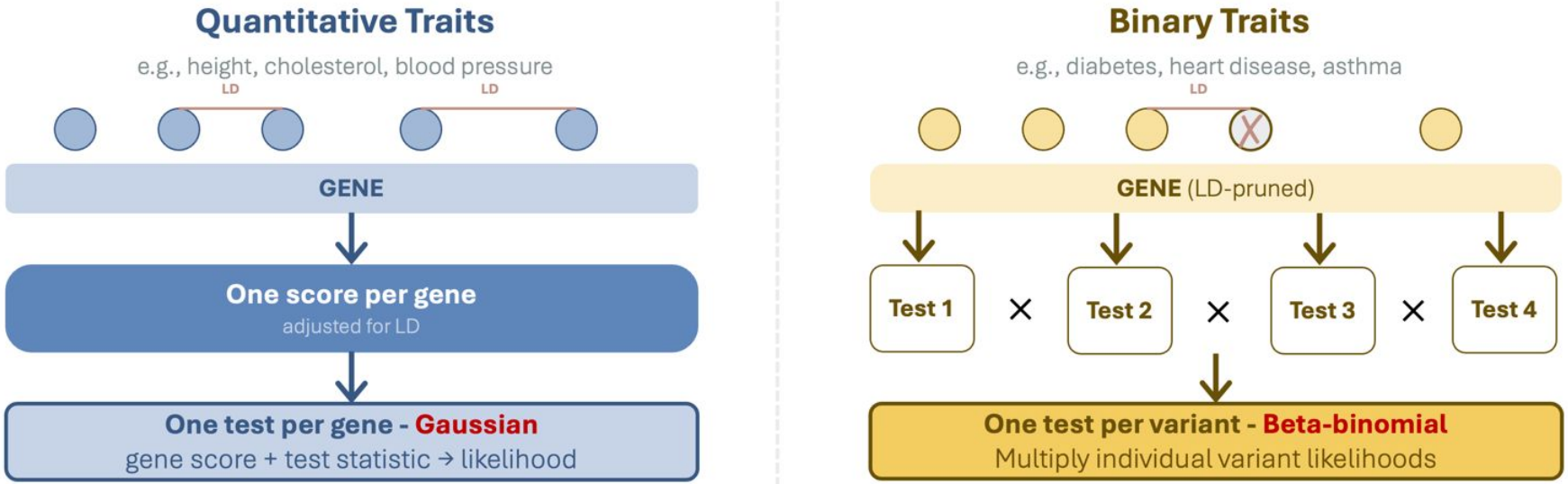
Output

- **Components**
- **Weights**

True burden effect sizes β

⇨ h^2_{Burden} | Polygenicity | Power | ...

Rare variants ~ binary traits association deviate from **Gaussian** assumptions



BurdenEM (Iterations)

Posterior

For each gene, compute **P(component | data)** using current weights & data likelihood

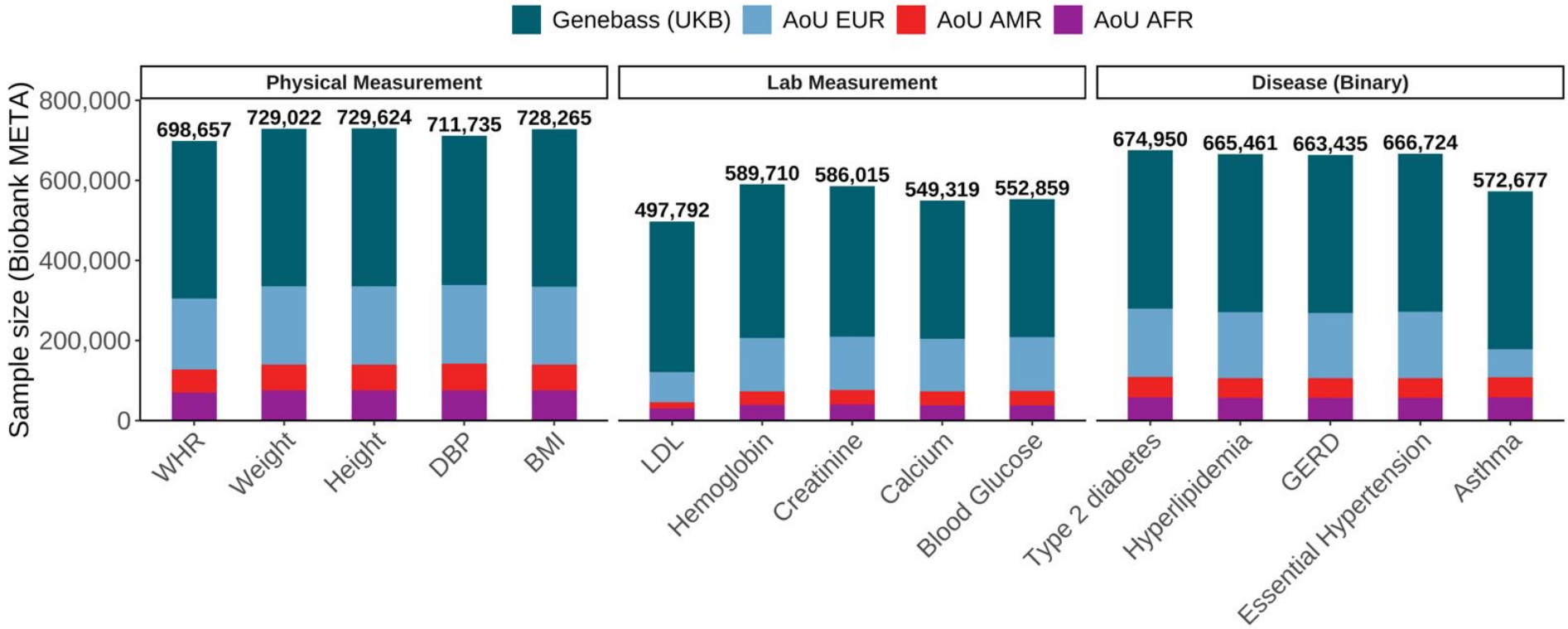
⟷

Update

Refit weights via least-squares

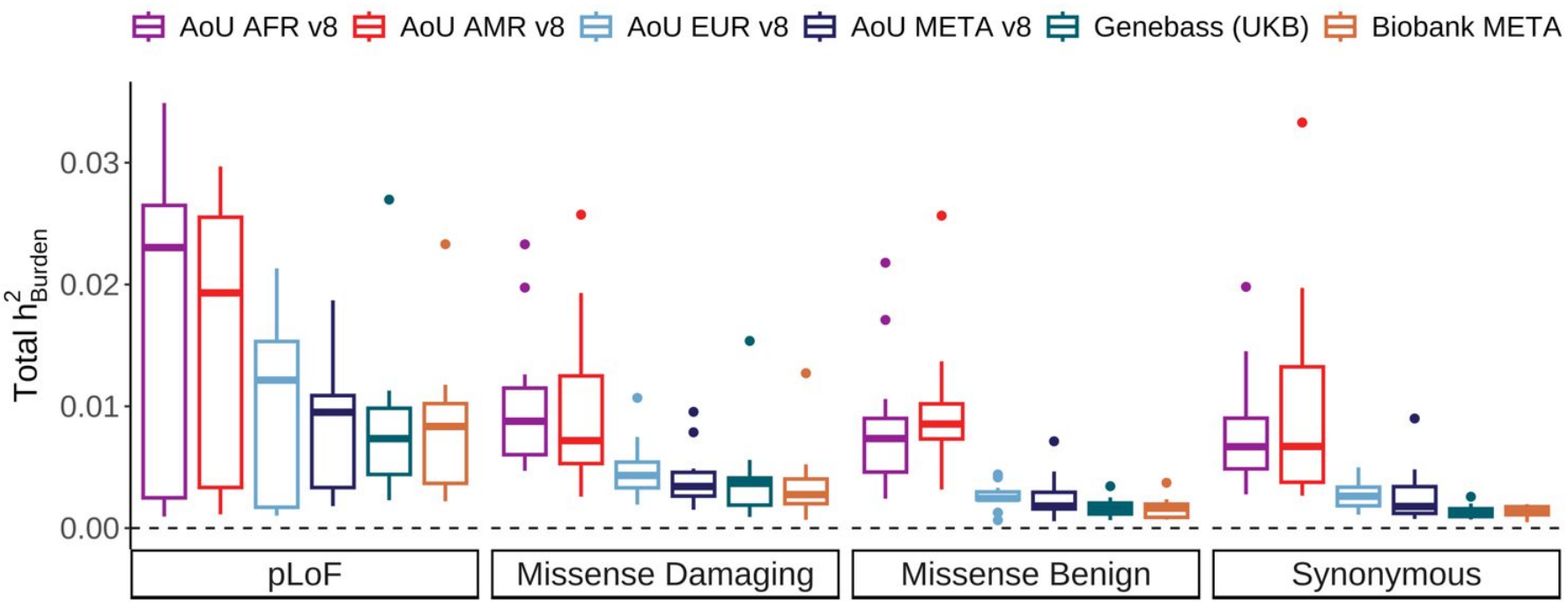
Closed-form regression on posteriors: $\Delta = (X^T X)^{-1} X^T R$

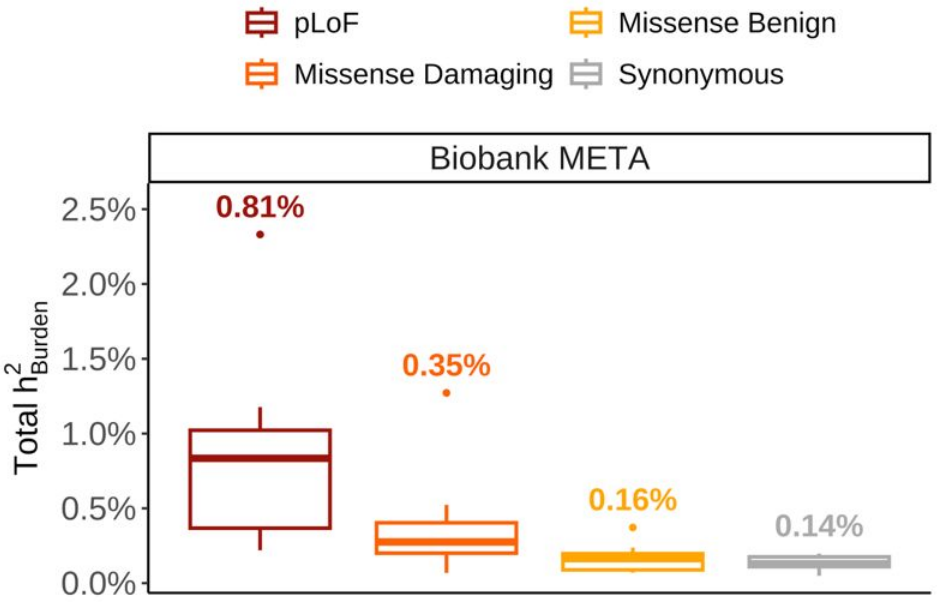
Curated 15 phenotypes shared between two biobanks



AoU AFR v8	AoU AMR v8	AoU EUR v8	AoU META v8	Genebass (UKB)	Biobank META
77,444	71,540	227,273	376,257	394,841	**771,098**

h^2_{Burden}: Proportion of phenotypic variance explained by the burden of alleles in a certain variant class





h^2_{Burden} due to burden of damaging rare coding variants is estimated to be **~1.2%,** highly **concordant** with those from Burden Heritability Regression (BHR, **1.3%**).

👉 Rare coding variants will make a **modest contribution** to heritability.

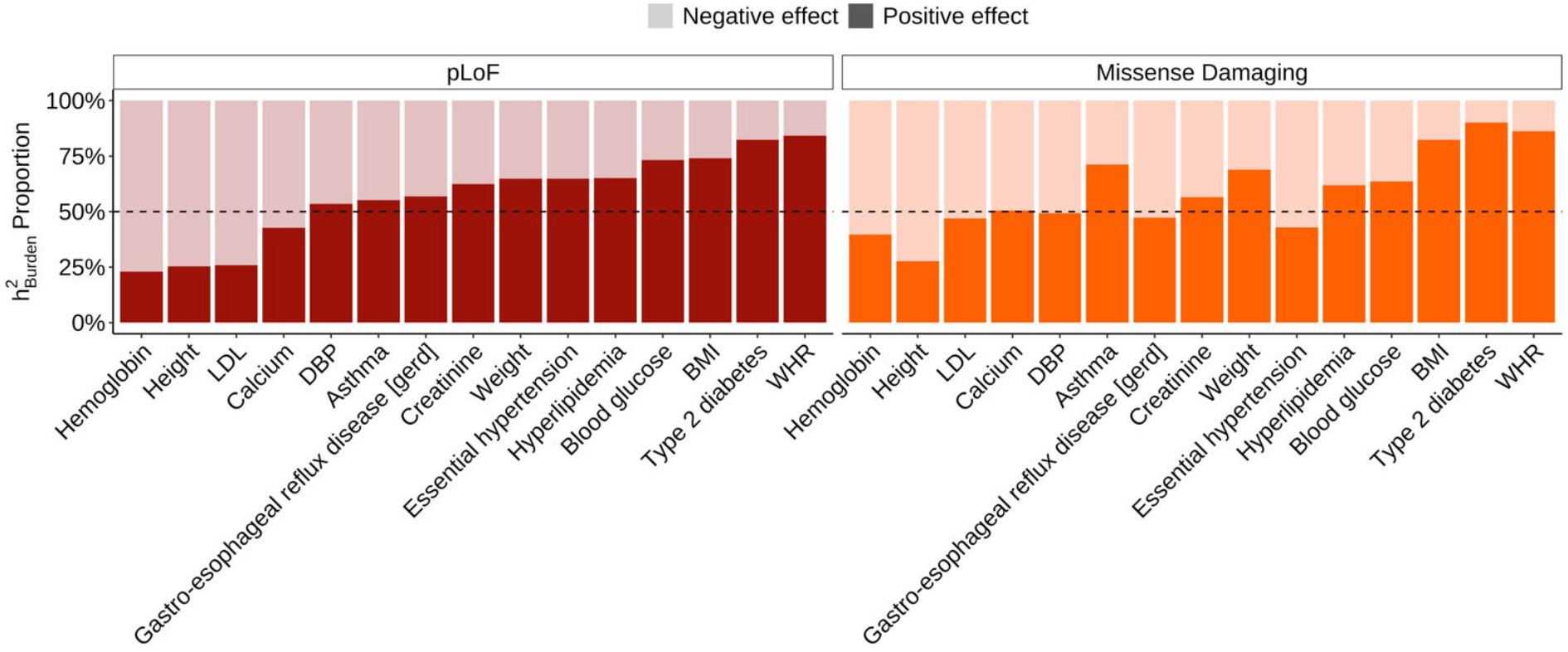
Driven by **different motivations**, BurdenEM quantifies rare variant properties not covered by BHR

Trait-increasing & decreasing

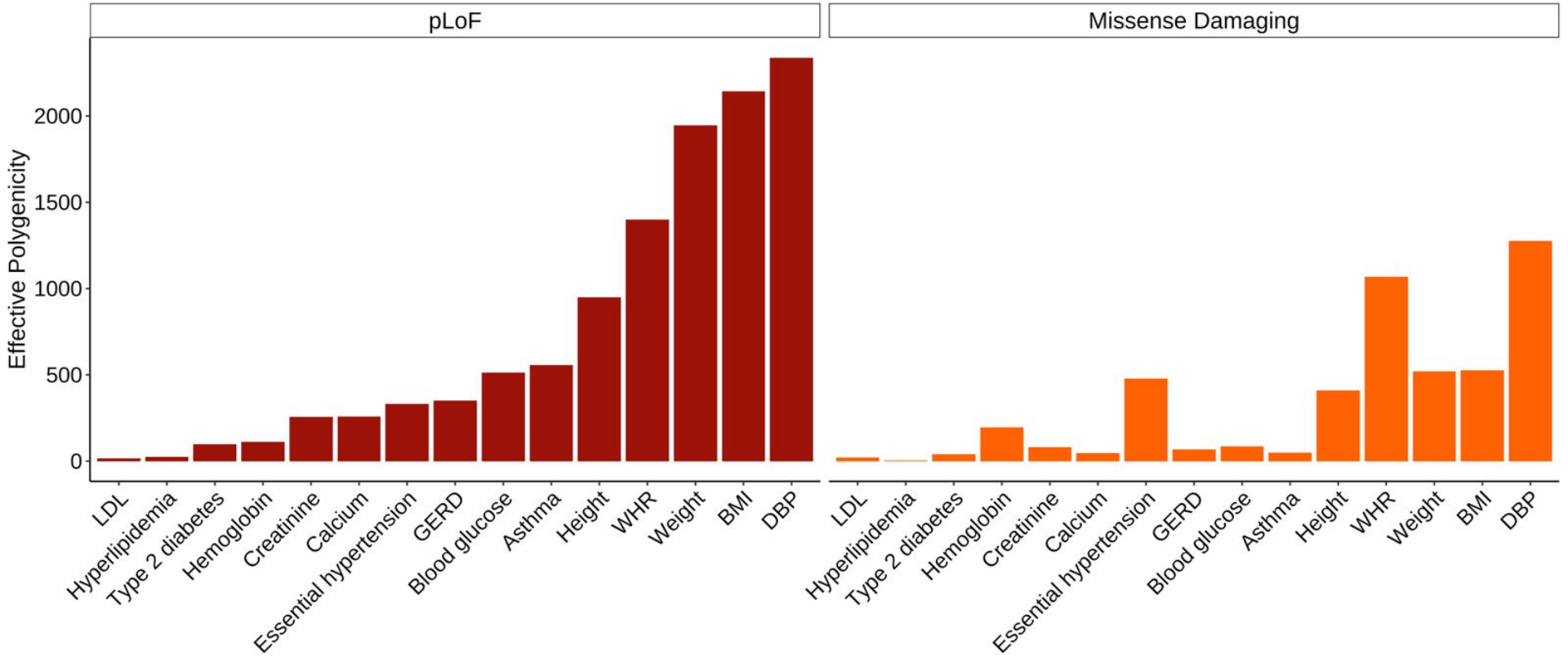
Power analysis

Polygenicity

Proportion of h^2_{Burden} from positive and negative effect sizes varies across traits



Rare variant polygenicity varies across traits



Thank you for listening