

Polygenic Score Workshop

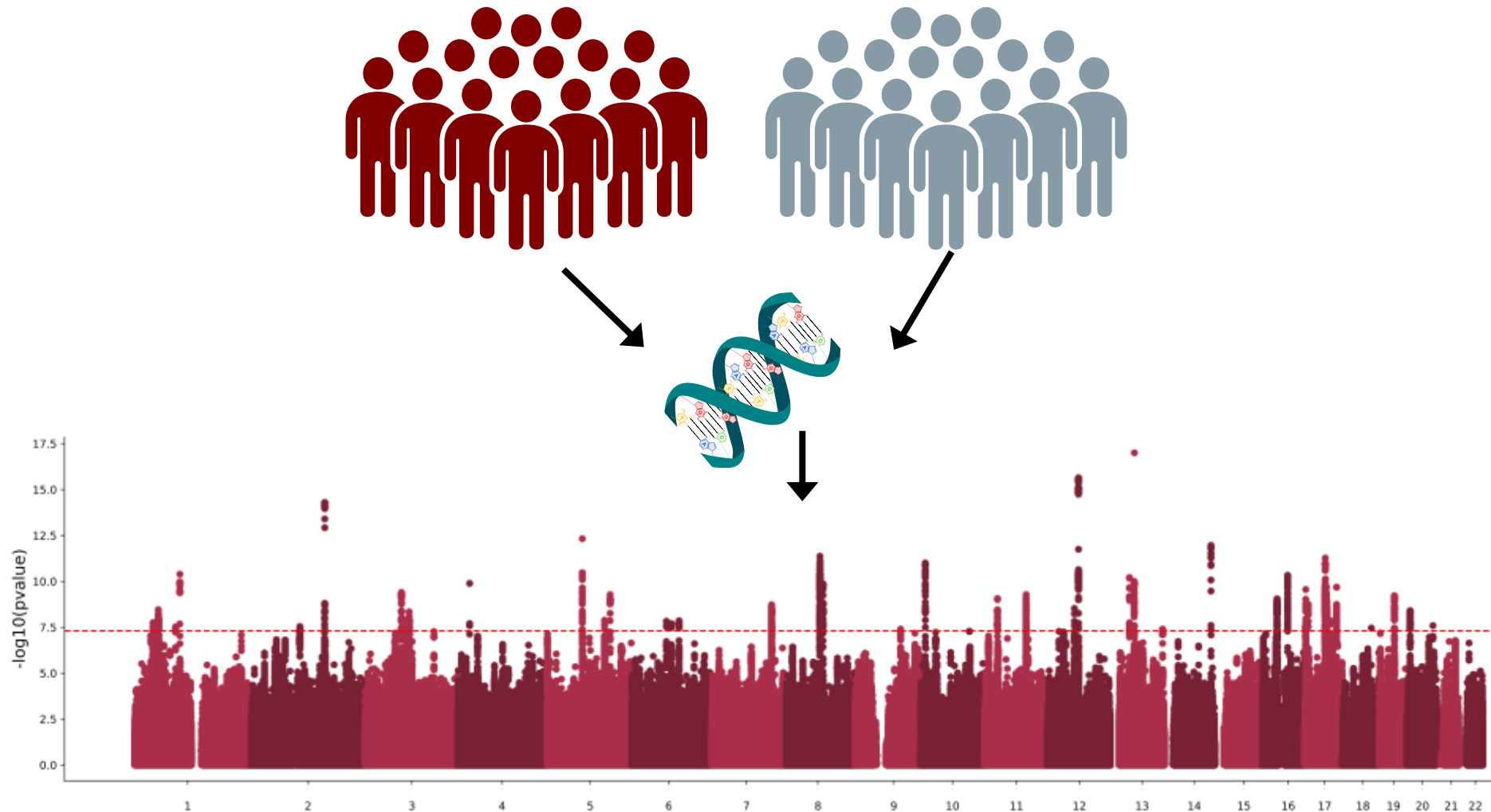
Part 1: PGS Intro and Construction

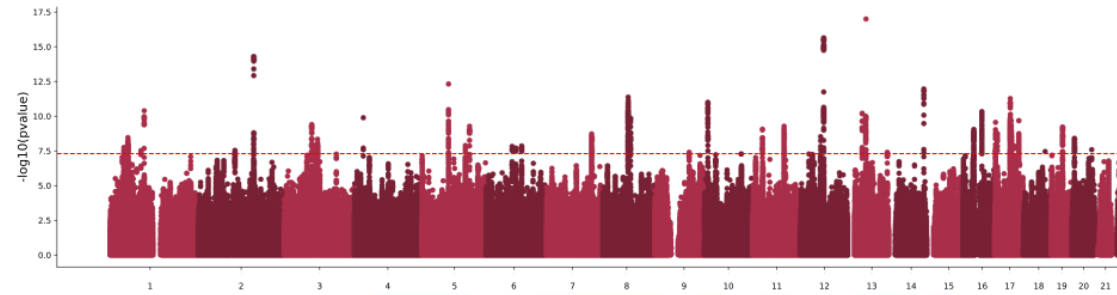
Brittany Mitchell

24th July 2025

Adelaide Health and Medical Sciences Building

Genome-Wide Association Study





Identification of susceptibility variants

Novel biological insights

Improved measures of individual aetiological processes

Clinical advances

Personalized medicine

Therapeutic targets

Biomarkers

Prevention

Diagnostics

Prognostics

Therapeutic optimization

- SNP annotation (FUMA)
- Fine-mapping
- Gene-based tests (MAGMA)
- Genetic Correlations (LDSC)
- Mendelian Randomization

- Polygenic Scores
- Multi-omic scores
- PGx profiles

Polygenic score
(PGS)

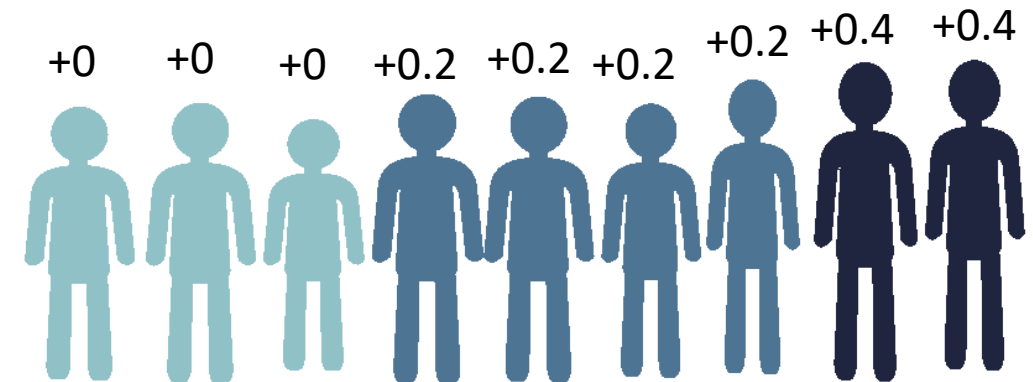
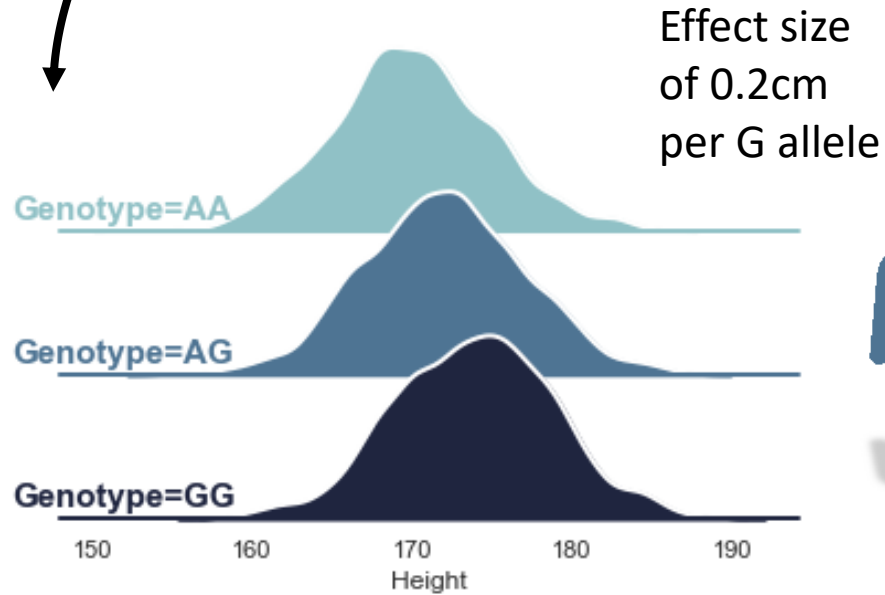
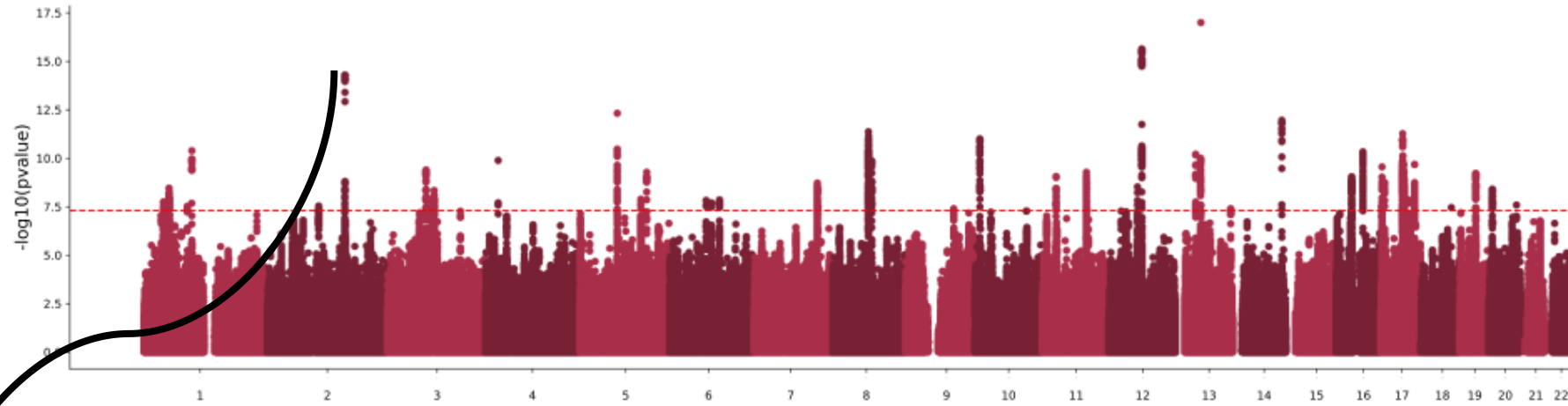
Polygenic index
(PGI)

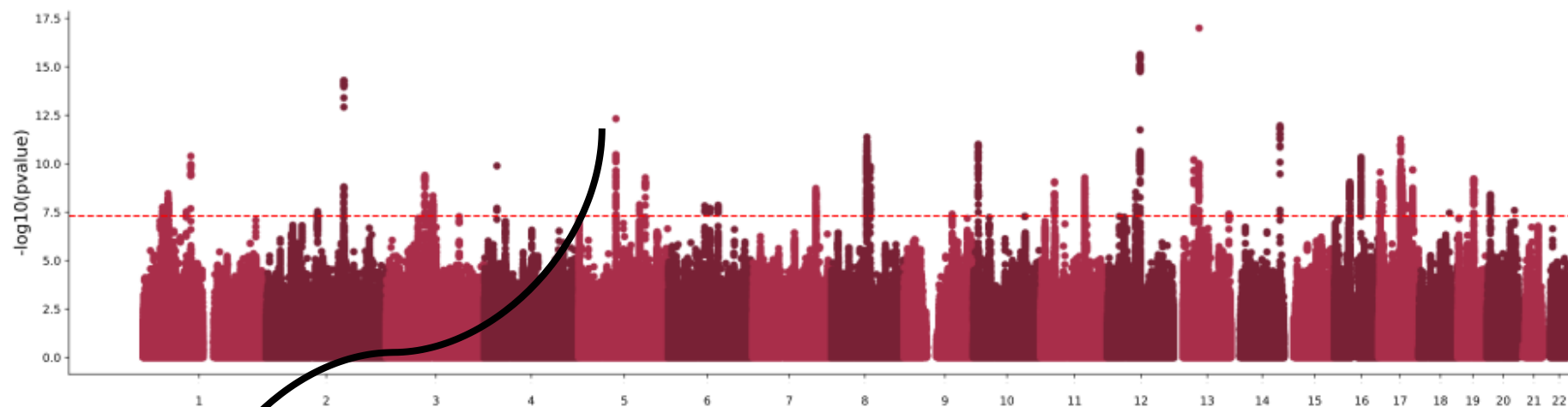
Polygenic risk
score (PRS)

Genetic risk
score (GRS)

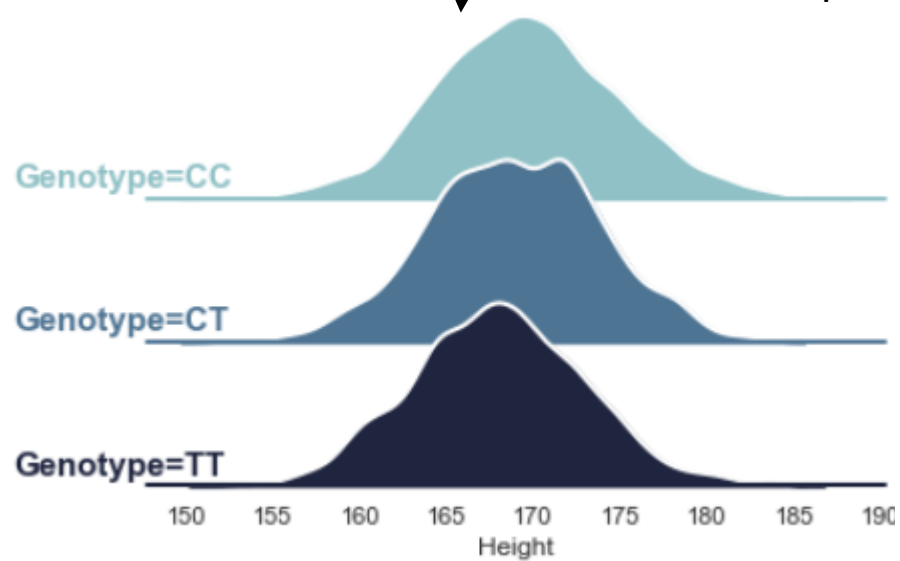
Genome-wide
score (GWS)

What is a polygenic Score?





Effect size of -0.1 per T allele



What is a Polygenic Score?

- An index that linearly aggregates the estimated effects of individual SNPs on the trait of interest.
- Can be considered a measure of an individual's **genetic propensity** towards a trait.
- Defined as a **weighted sum of a persons genotypes at K loci**.
- Start with additive model using measured SNPs:

$$y_i = \underbrace{A_{SNP,i}(x_i)}_{\text{additive SNP factor}} + \epsilon_{i,SNP} = \sum_{j=1}^K \underbrace{\beta_j}_{\text{Effect}} \underbrace{x_{ij}}_{\text{Genotype/dosage}} + \epsilon_{i,SNP}$$

Predictive power of a PGS

If we regress y on \hat{A}_{SNP} we get an Ordinary Least Squares coefficient of

$$\begin{aligned} b &= \frac{Cov(\hat{A}_{SNP}, y)}{Var(\hat{A}_{SNP})} \\ &= \frac{Cov(A_{SNP} + U_i, A_{SNP} + \epsilon_{SNP})}{Var(A_{SNP} + U)} \\ &= \frac{Var(A_{SNP})}{Var(A_{SNP}) + Var(U)} \end{aligned}$$

\hat{A}_{SNP} = **genetic value predicted from SNPs** (e.g., PGS or linear predictor)

U = **residual component**, capturing everything **not explained** by A_{SNP} including:

- environmental factors
- genetic effects not tagged by SNPs (e.g., rare variants)
- measurement error

The proportion of variance in a trait that is explained by genetic effects captured by SNPs (SNP-based heritability)

OLS:

$$y_i = a + bx_i + \epsilon_i$$
$$b = \frac{Cov(x,y)}{Var(x)}, R^2 = \frac{b^2 Var(x)}{Var(y)}$$

Predictive power of a PGS

How much of the **phenotypic variance** is explained by the **predicted genetic value** \hat{A}_{SNP} — such as a polygenic score.

b = **regression coefficient** obtained from regressing the phenotype y on the predicted genetic value

$Var(\hat{A}_{SNP})$ = **variance of the predicted genetic scores** across individuals

$Var(y)$ = **Total variance in the observed phenotype** across individuals

And the expected predictive power is:

$$R^2 \approx \frac{\text{Genetic Variance}}{\text{Total variance}} = R^2 \approx \frac{b^2 Var(\hat{A}_{SNP})}{Var(y)}$$

$$= \left(\frac{Var(A_{SNP})}{Var(A_{SNP}) + Var(U)} \right)^2 \frac{Var(\hat{A}_{SNP})}{Var(y)}$$

⋮

$$\approx \frac{h_{SNP}^2}{h_{SNP}^2 + \frac{M_e}{N}}$$

Sometimes called the Daetwyler formula (Daetwyler et al. 2008)

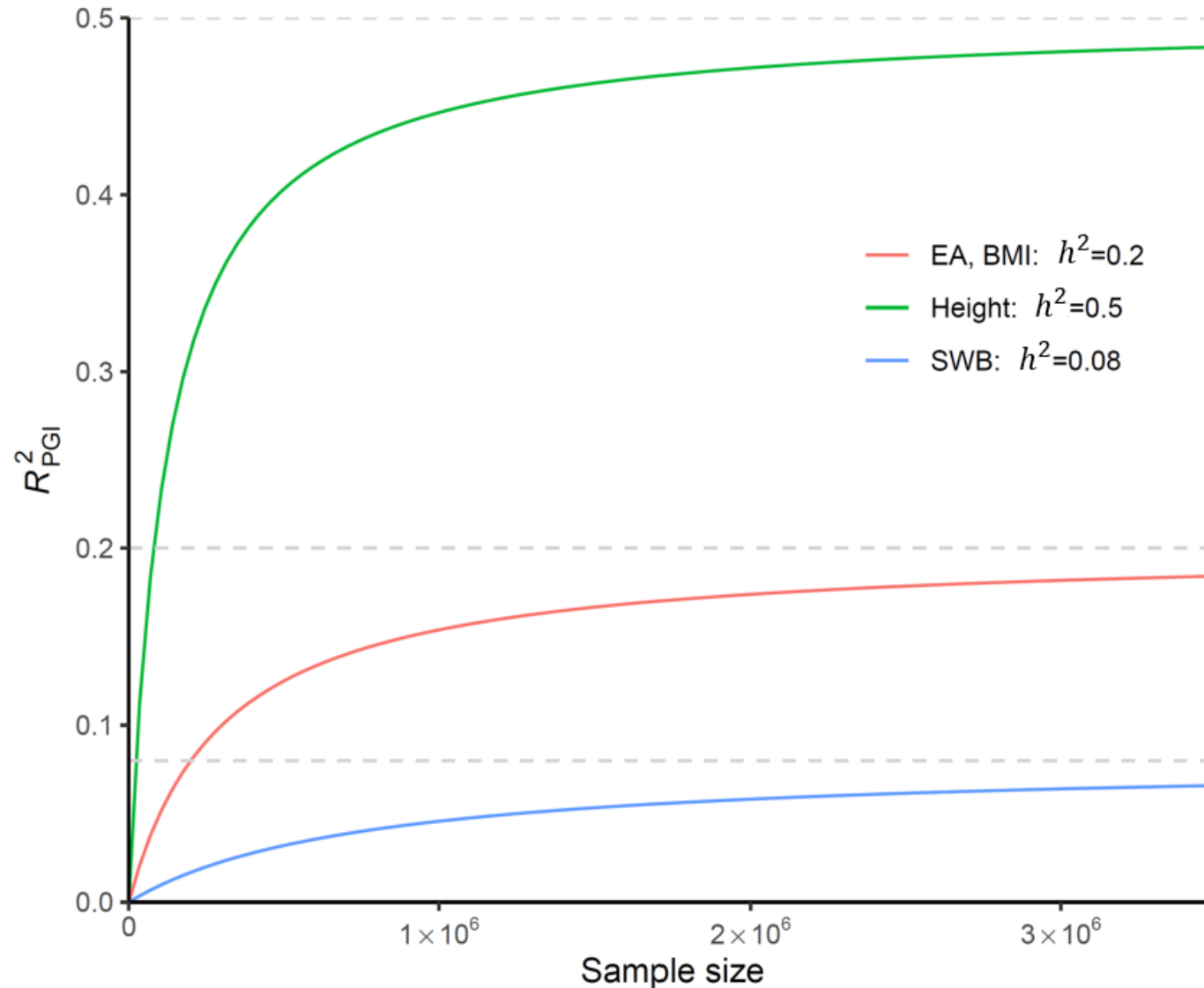
OLS:

$$y_i = a + bx_i + \epsilon_i$$

$$b = \frac{Cov(x,y)}{Var(x)}, R^2 = \frac{b^2 Var(x)}{Var(y)}$$

Effective number of SNPs in the PGS, estimated to be between 50k-70k in genome-wide data for EUR ancestry (Wray et al. 2013)

Theoretical projections for R_{PGS}^2



- SNP-heritability
- Number effective SNPs = GWAS sample size

Constructing PGS

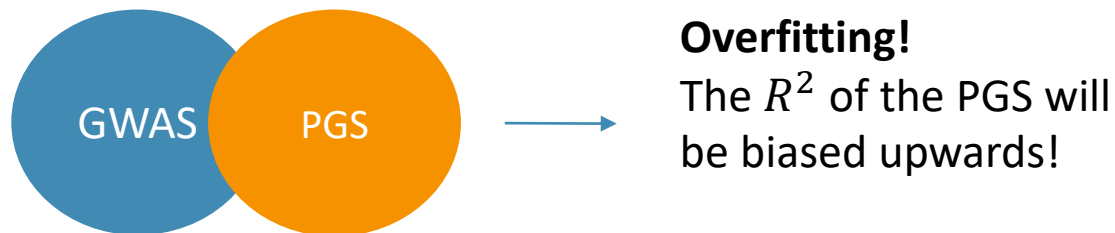
What do you need?

- Individual-level genotype data from a prediction sample (Imputed)
- Weights: GWAS summary statistics from a discovery sample
- Reference genotypes to estimate LD

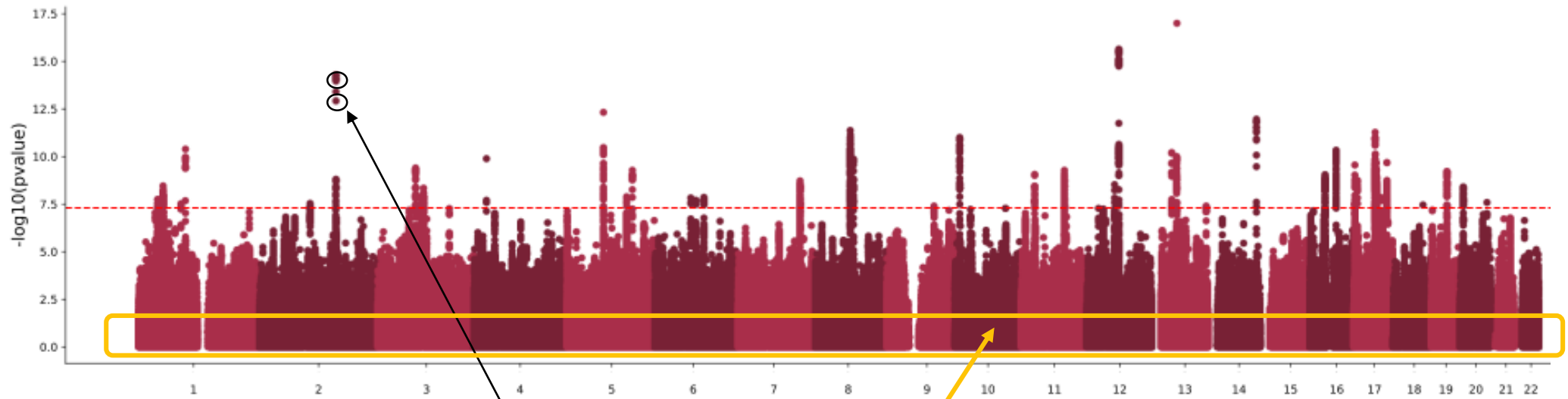
Caution:

The prediction sample should not overlap with the discovery sample!

The prediction sample should be of similar ancestry to discovery sample!



Which SNPs? What weights?



GWAS results give us $\hat{\beta}_j^{GWAS}$, not β_j . Two issues to consider when constructing $\sum_{j=1}^K \hat{\beta}_j^{GWAS} x_{ij}$:

1. For some SNPs, $\hat{\beta}_j^{GWAS}$ may be a very noisy estimate of β_j and/or β_j may be close to 0, so adding those SNPs will add more noise than signal
2. If we include all SNPs, we will overweight (“double-count”) SNPs with high LD scores

Two solutions

Clumping and thresholding

Include only the most strongly associated SNP from each LD block (Purcell et al., 2009)

Weights: Set equal to GWAS coefficients.

Loci: Selected by

1. using a **clumping** algorithm that ensures the included markers are all approximately independent of each other
2. omitting SNPs whose P value for association with the phenotype is above a certain **threshold**

$$\sum_{j=1}^K \hat{\beta}_j^{GWAS} x_{ij}$$

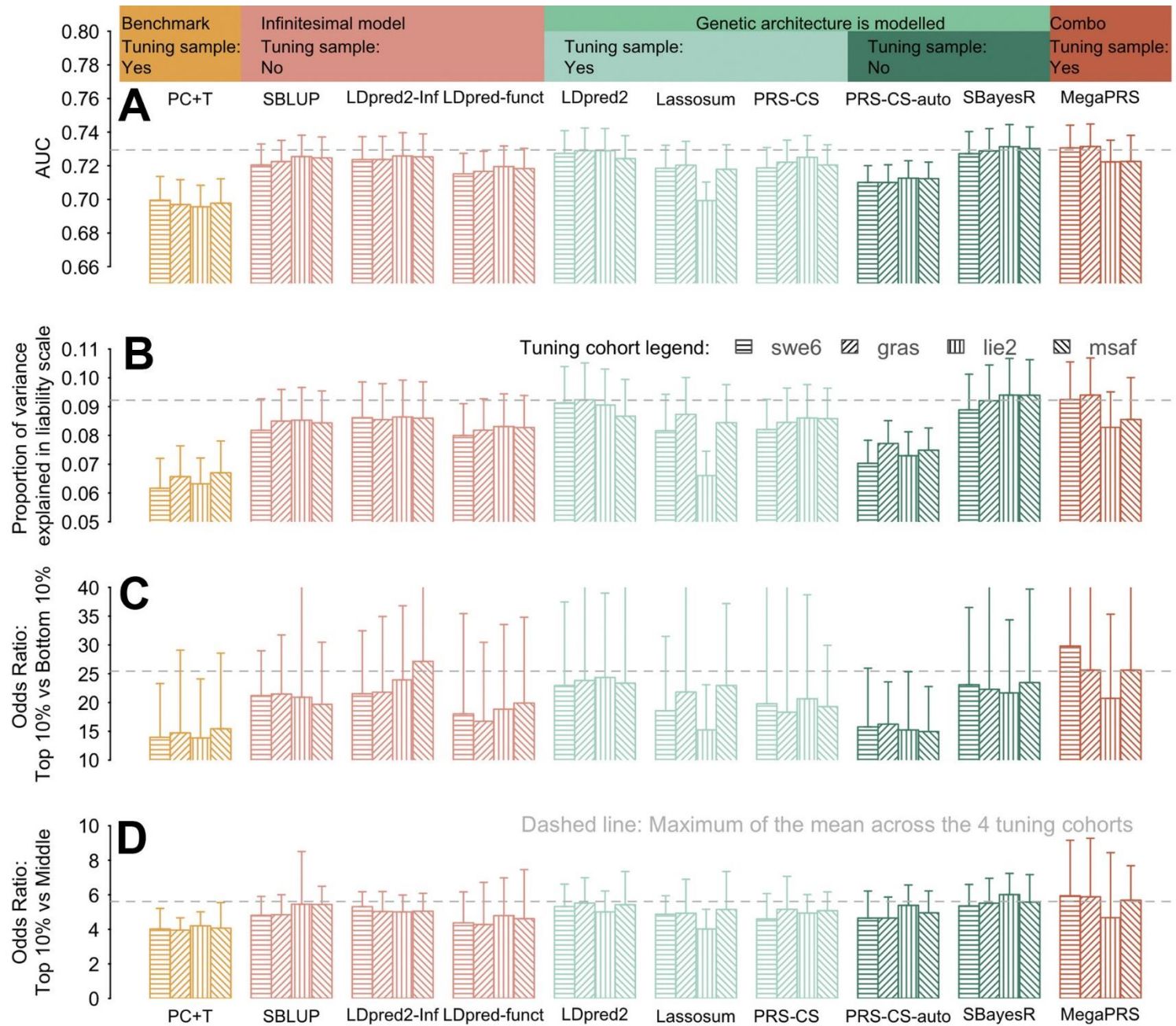
Bayesian approaches

Include all SNPs but adjust the effect sizes for LD

Weights: Set to GWAS coefficients **adjusted for LD and/or Genetic architecture** → approximate results from a theoretical multiple regression of the phenotype on all SNPs

Loci: Include **all SNPs**, no LD-based pruning

Examples: LDpred (Vilhjalmsson et al. 2015, Prive et al. 2020), PRS-CS (Ge et al. 2019), SBayesR (Lloyd-Jones et al. 2019), **SBayesRC (Zhang 2024)**



Archival Report

A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts

Guiyan Ni ^a, Jian Zeng ^a, Joana A. Revez ^a, Ying Wang ^a, Zhili Zheng ^a, Tian Ge ^d, Restuadi Restuadi ^a, Jacqueline Kiewa ^a, Dale R. Nyholt ^c, Jonathan R.I. Coleman ^g, Jordan W. Smoller ^{d e f}
Schizophrenia Working Group of the Psychiatric Genomics Consortium

[Show more](#) ▾

If the purpose is to maximize predictive power, then Bayesian approaches clearly do better

C+T vs Bayesian approaches

Clumping and thresholding

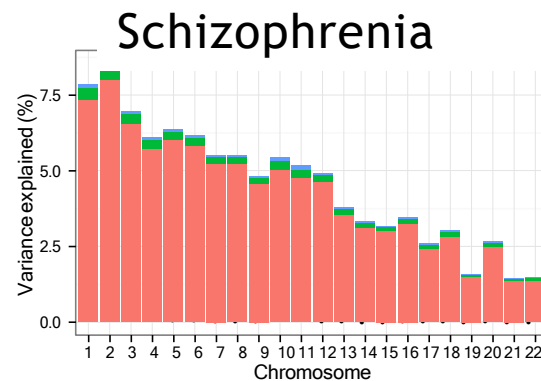
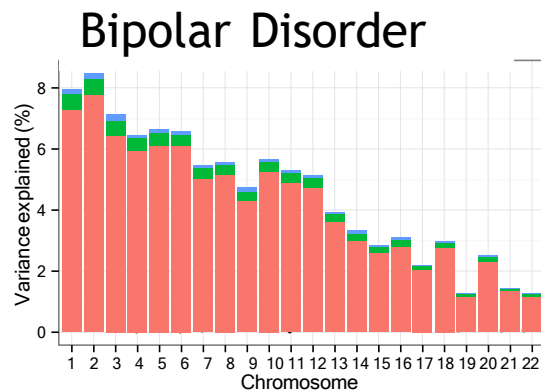
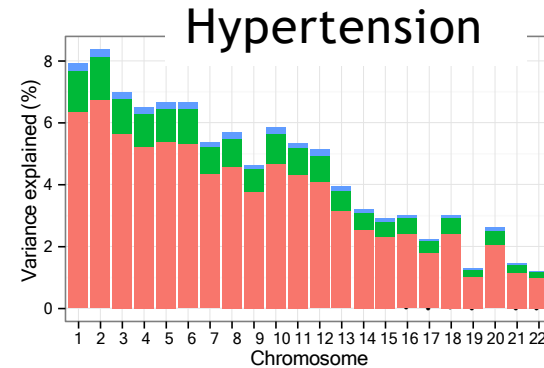
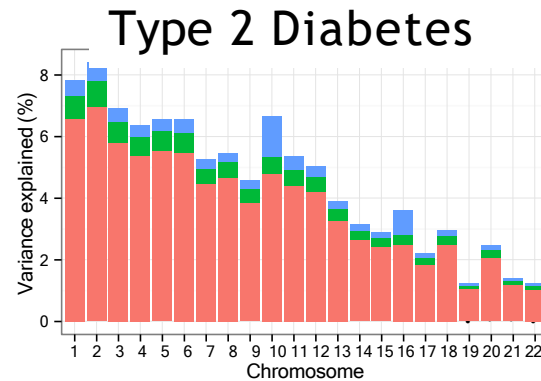
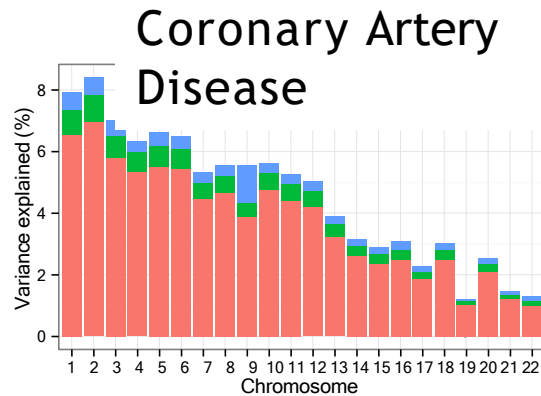
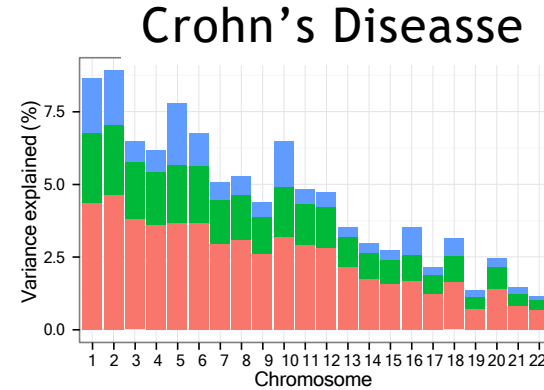
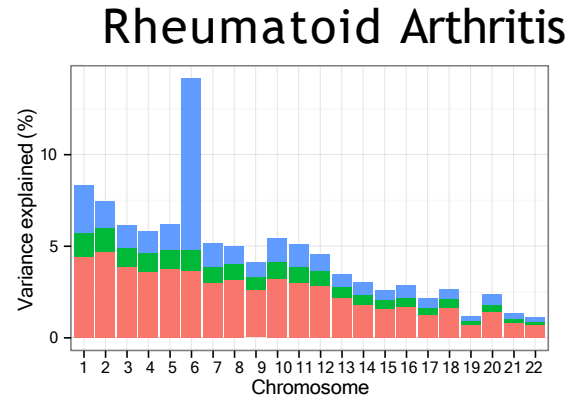
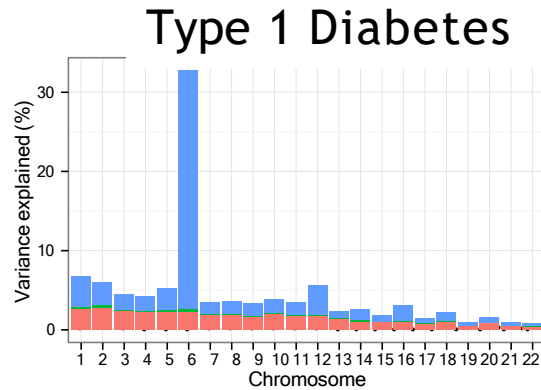
Faster and easier, but too black & white?

- If clumping r^2 or P -value cutoffs too strict, it drops potentially causal SNPs.
- If clumping r^2 and P -value cutoffs too relaxed, there is a lot of double-counting and noise

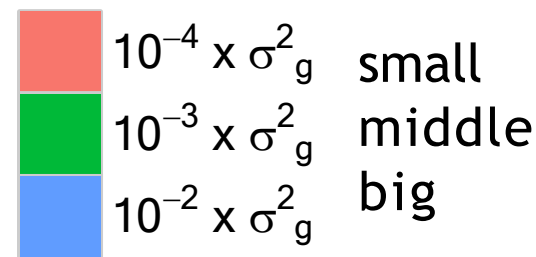
Bayesian approaches

- Utilize information from all SNPs by adjusting SNP weights for LD, but
 - If the reference panel is not a good match for the population from which summary statistics were obtained, prediction accuracy might be compromised
 - The assumed prior distribution might not accurately model the true genetic architecture

Polygenic Traits


















Mixture component



Most variation is explained by small effects

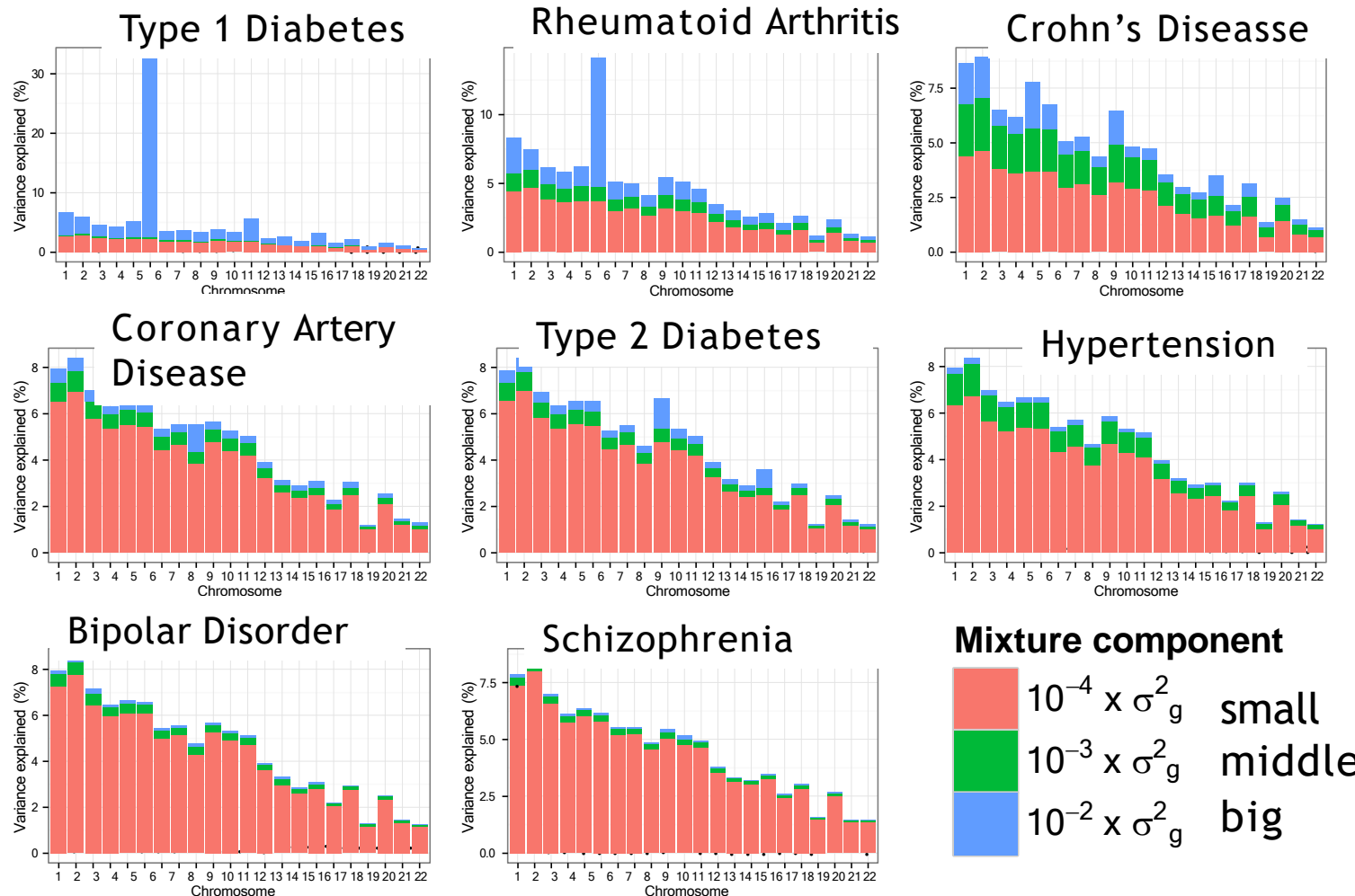
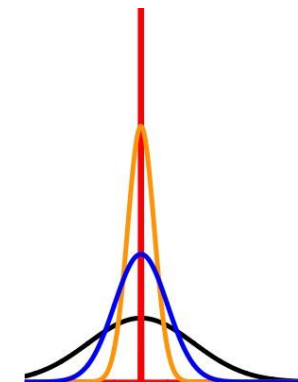
The method that best describes the distribution of SNP effects gives the most accurate PRS

Improved polygenic prediction by Bayesian multiple regression on summary statistics

[Luke R. Lloyd-Jones](#) , [Jian Zeng](#) , [Julia Sidorenko](#) , [Loïc Yengo](#) , [Gerhard Moser](#) , [Kathryn E. Kemper](#) ,
[Huanwei Wang](#) , [Zhili Zheng](#) , [Reedik Magi](#) , [Tõnu Esko](#) , [Andres Metspalu](#) , [Naomi R. Wray](#) , [Michael E. Goddard](#) , [Jian Yang](#)  & [Peter M. Visscher](#) 

SBayesR:

- Assumes SNP effect sizes follow a **mixture of normal distributions**, including a spike at zero (to allow many SNPs to have no effect).
- Estimates the posterior effect sizes of SNPs genome-wide.

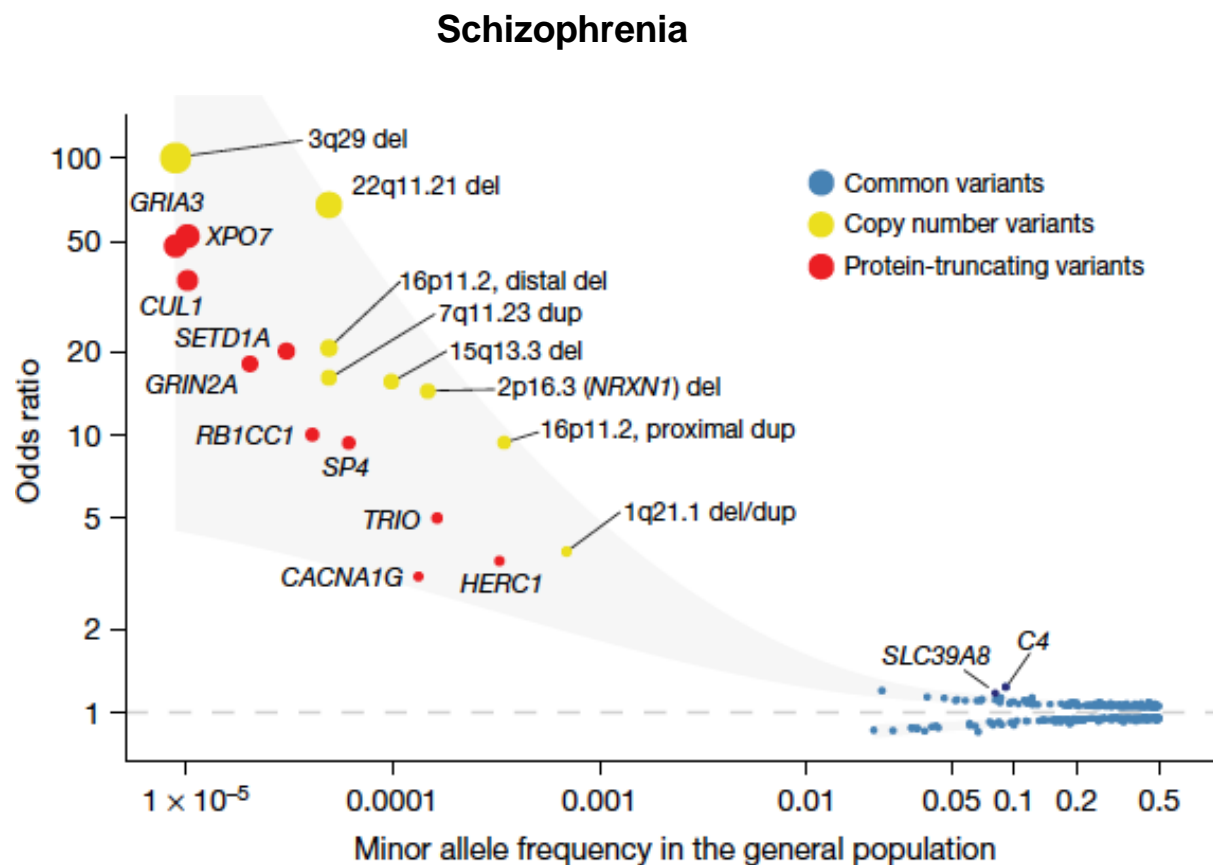


SBayesRC

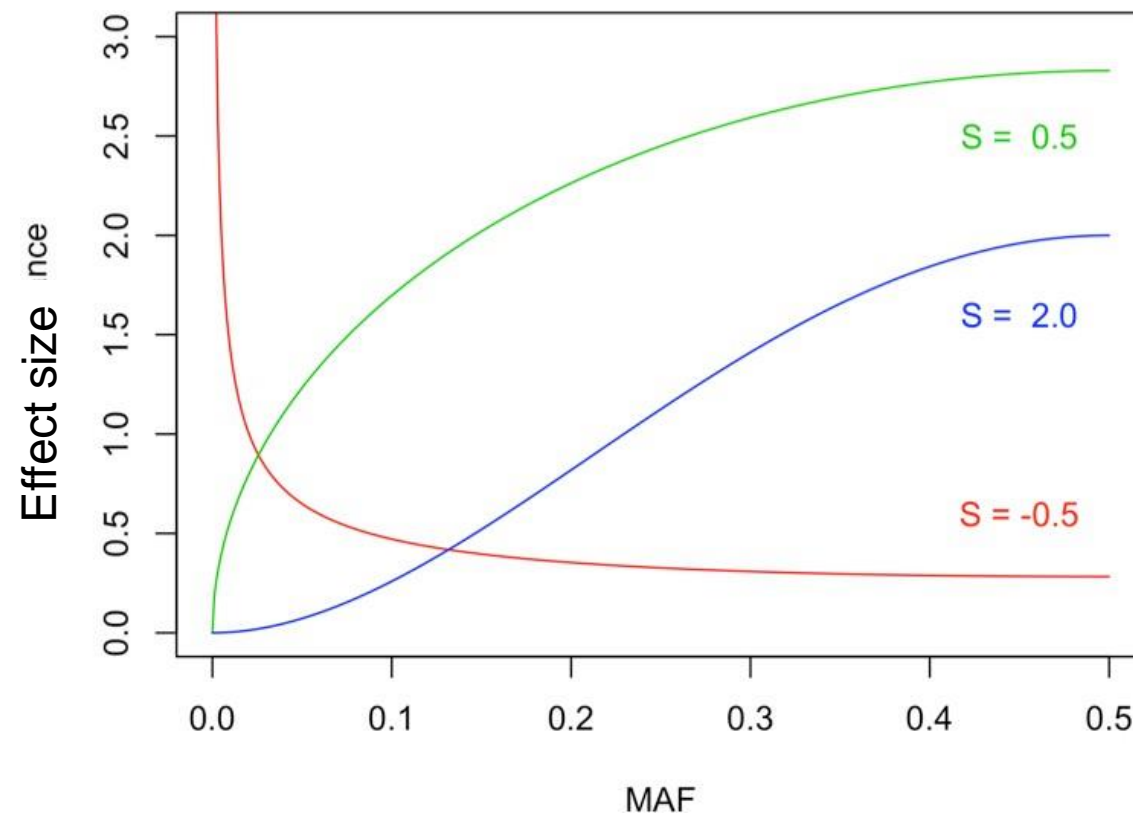
Signatures of negative selection

Negative Selection: Reverse relationship between effect size and allele frequency

$$\beta_j \begin{cases} \sim N(0, [2p_j q_j]^S \sigma_\beta^2), & \pi \\ = 0, & 1 - \pi \end{cases}$$



Singh et al. 2022 Nature



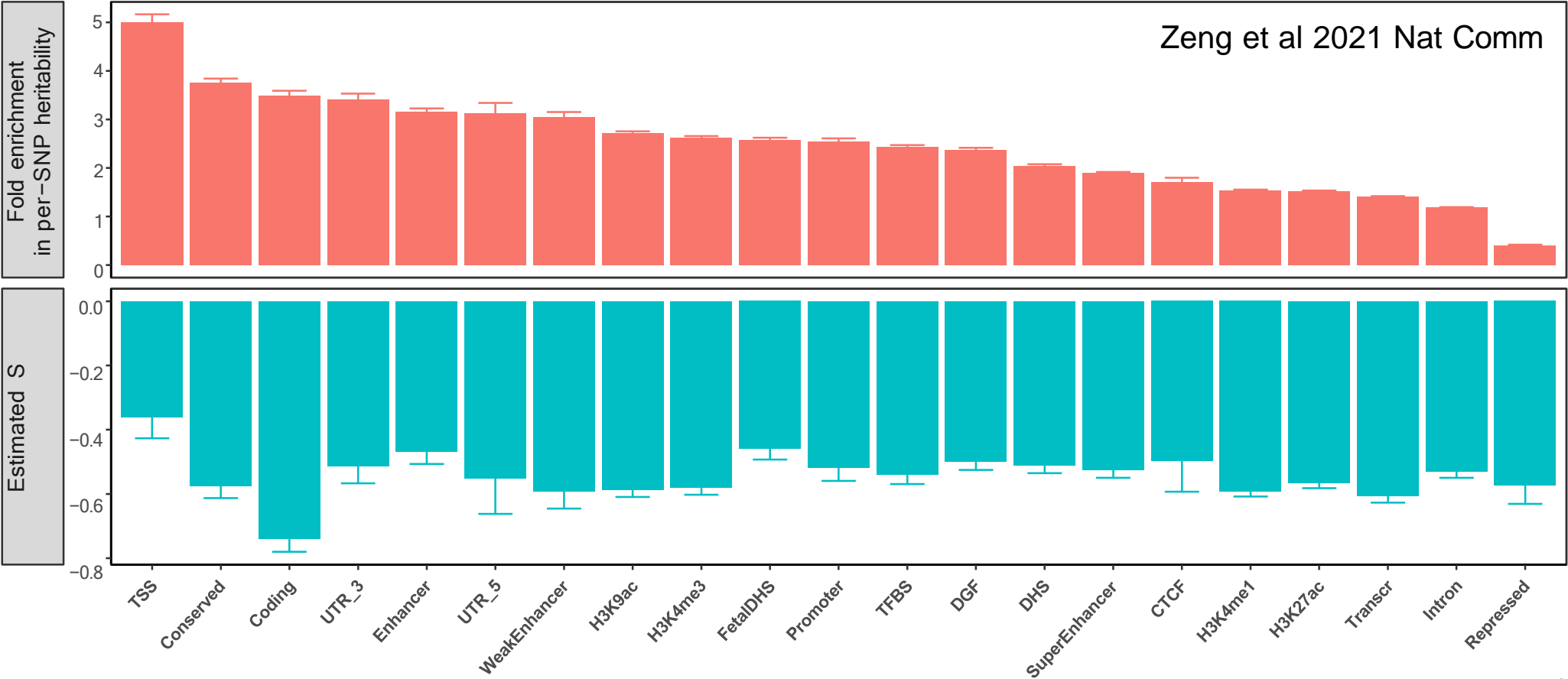
Minor allele frequency (MAF)

Zeng et al 2018 Nat Genet

Functional genetic architecture

Genetic architecture and selection signatures vary across functional annotation categories.

Suggesting different effect size distributions across functional annotations



SBayesRC (Zheng et al 2024 Nat Genet)

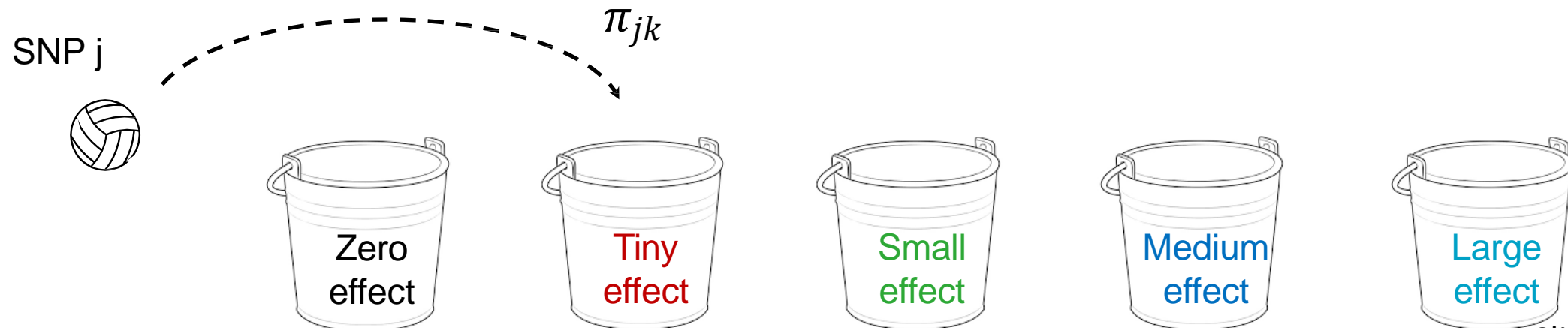
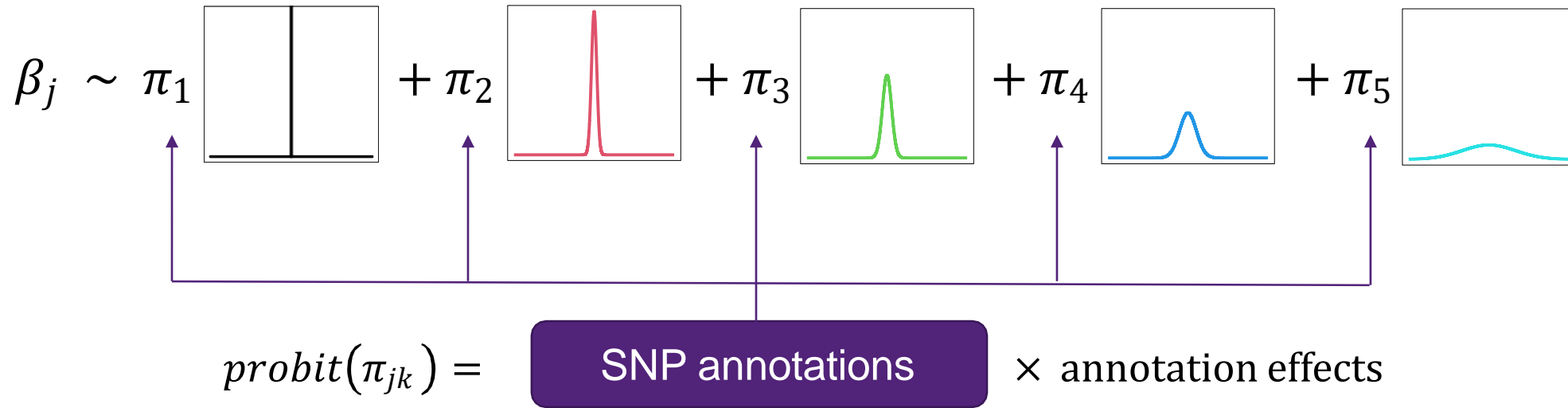
Incorporate functional annotations through a hierarchical prior:



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

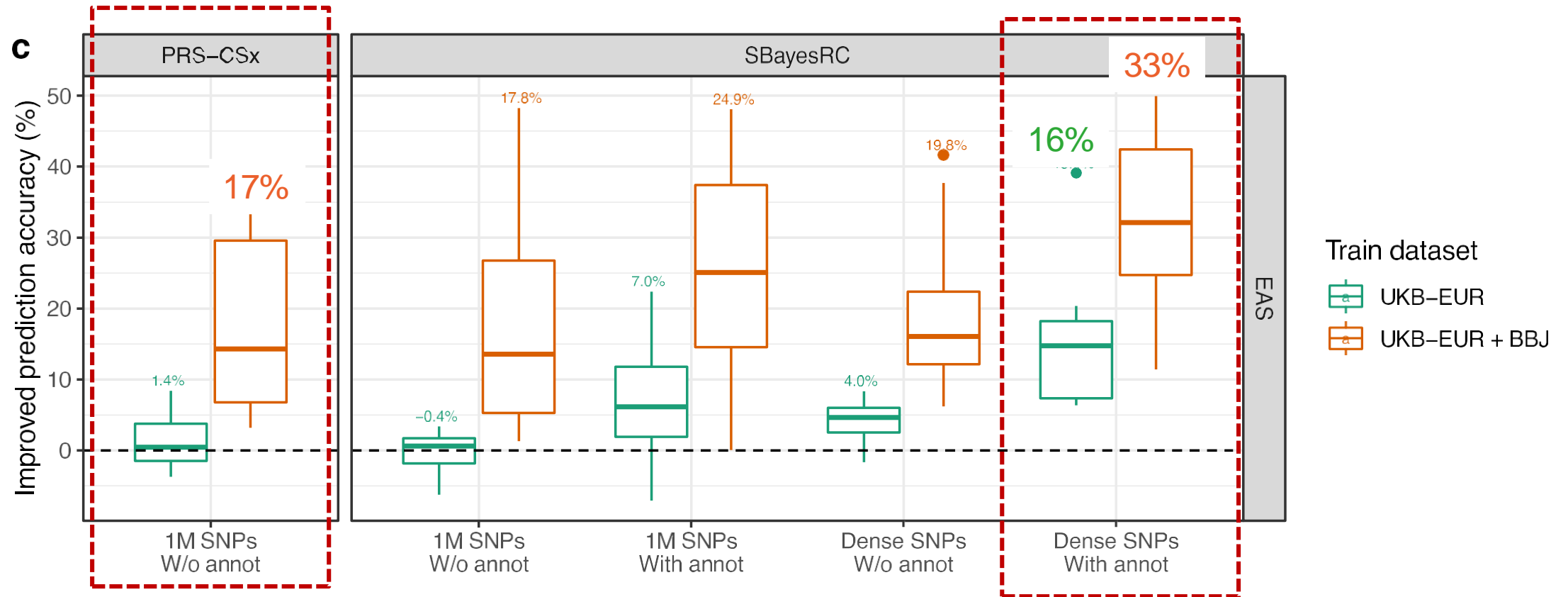


Zhili Zheng




Trans-ancestry prediction

Use GWAS data from UKB EUR and BBJ EAS to predict UKB EAS










Practical Steps

- https://github.com/jodithea/Polygenic_score_workshop.git


Polygenic_score_workshop
Private
Watch 0

main
2 Branches
0 Tags

Add file
Code

	Update traffic log: 2025-07-21T01:22:18Z	b91350f · 4 hours ago	24 Commits
	.github/workflows	Add GitHub Action to log traffic stats	5 days ago
	01_Create_PGS	add evaluate PGS in R and update readme	5 days ago
	02_Evaluate_PGS	add evaluate PGS in R and update readme	5 days ago
	.gitignore	update gitignore	last week
	README.md	Add GitHub Actions workflow to log traffic	5 days ago
	traffic_log.csv	Update traffic log: 2025-07-21T01:22:18Z	4 hours ago

README
✎
☰

Polygenic Scores Workshop

Practical Steps

1. Choosing your LD reference

- SBayesRC and other software (e.g. SBayesR, PRS-CS) provide calculated LD estimates
- These are usually limited to a set of good quality SNPs (e.g. HapMap3) to reduce errors in LD estimation and computational burden while ensuring sufficient coverage
- SBayesRC can be run with 1M, **7M** or 13M LD panel – more density increases prediction
- But you can also estimate your own! Why would you?
 - You may want to include more SNPs
 - The available LD reference data may not be a good match for the ancestry of your GWAS
- If you decide to obtain your own LD estimates, you should make sure that the quality of your data is good. Bayesian approaches are very sensitive to errors in LD estimates!
- The SNPs included in the PGS will be limited to the intersection of GWAS, LD reference sample and target data.

Practical Steps

1. Choosing your LD reference

- Both the LD and Annotation files needed are provided in the tutorial

Eigen-decomposition LD Matrix

- Files: `ukbEUR_Imputed/block*.eigen.bin`
- Source: [GCTB website](#)
 - Under 'Eigen-decomposition data of LD matrices' select '7M Imputed SNPs'
- Use the 'ukbEUR_Imputed' directory
- This contains LD matrix eigen-decomposition from unrelated UK Biobank individuals of European ancestry for ~7 million SNPs, used by SBayesRC.

Functional Genomic Annotation

- File: `annot_S-LDSC_BaselineLDv2.2.txt`
- Source: [GCTB website](#)
 - Under "Functional genomic annotations," select "7M SNP annotations"
- This is the formatted data for per-SNP functional annotations for ~7 million SNPs

Practical Steps

2. QC and Formatting of GWAS Summary Statistics

Script: '01_Format_GWAS_sumstats.sh'

- Ensure no sample overlap exists between individuals used in the GWAS and your genotype cohort
 - Request LOO (Leave-One-Out) summary statistics from the authors OR remove overlapping individuals from your analysis
- Confirm that genome builds are consistent:
The LD matrix provided in GCTB is on build 37 so your GWAS summary statistics should also be on build 37
Conduct conversions if needed
- GWAS Summary statistics QC:
 - Filter SNPs with MAF > 0.01
 - Filter on imputation score > 0.6 (if this data is available)
- Reformat the GWAS summary statistics to COJO format (**required by SBayesRC**)
.ma file

Header row: SNP A1 A2 freq b se p N (SNP identifier (rsID))
Ensure A1 and A2 are uppercase

Practical Steps

2. QC and Formatting of GWAS Summary Statistics

Common mistakes:

- Check the scale of your effect column (eg beta, not OR)
- Be very sure which column is the effect allele and non-effect allele. Worth double-checking the Readme!
- Missing required fields – not always provided in summary stats; need to merge from a reference panel (eg EAF)
- Incorrectly formatted p-values (p=0 or wrong scale e.g. Regenie output $\log_{10}(\text{p-values})$)
- Reference genome build mismatch
- Including duplicated or multiallelic SNPs – not all software can handle this
- It is a good idea to limit the SNPs in the GWAS to those available in the target data prior to adjusting for LD, especially if the overlap is rather poor (e.g. if you only have array SNPs in the target data)
- Bayesian software will assume all SNPs in the GWAS will be included in the PGS and make LD adjustments to maximize prediction accuracy. If some SNPs cannot be included because they are not in the target data, the adjustments will be suboptimal.

Practical Steps

3. Run SBayesRC in GCTB to get new weights

3.1 Imputation

Scripts: '02_a_Impute_GWAS_sumstats.sh' and '02_b_Impute_GWAS_sumstats.sh'

- Match alleles between the GWAS summary statistics and the LD matrix
- Remove SNPs with sample size >3 SD from median
- Impute summary statistics for SNPs in the LD matrix but missing in the GWAS summary statistics

This process is performed in parallel over 591 LD matrix blocks'

- Merge results into a single QCed/imputed summary statistics file

3.2 SBayesRC

Script: '03_SBayesRC_PGS_weights.sh'

- Run SBayesRC to calculate the polygenic weights for each SNP

Key output file: *.snpRes

column 2 = SNP (rsID)

column 5 = A1 (effect allele)

column 8 = A1Effect (Posterior mean of the SNP effect size (beta) = PGS weight)

Practical Steps

3. Run SBayesRC in GCTB to get new weights

Key output file: *.snpRes

column 2 = SNP (rsID)

column 5 = A1 (effect allele)

column 8 = A1Effect (Posterior mean of the SNP effect size (beta) = PGS weight)

Index	Name	Chrom	Position	A1	A2	A1Frq	A1Effect	SE	VarExplained	PEP	Pi1	Pi2	Pi3	Pi4	Pi5	PIP
1	rs12132974	1	801661	T	C	0.079275	-0.000019	0.000193	5.466305e-09	0.010000	0.992953	0.007047	0.000000	0.000000	0.000000	0.00704718
2	rs12134490	1	801680	C	A	0.079225	0.000000	0.000000	0.000000e+00	0.000000	0.995023	0.004977	0.000000	0.000000	0.000000	0.00497723
3	rs17276806	1	801858	T	C	0.079300	0.000004	0.000051	3.859203e-10	0.005000	0.995801	0.004199	0.000000	0.000000	0.000000	0.00419903
4	rs139867617	1	802856	T	C	0.079275	0.000015	0.000212	6.569427e-09	0.005000	0.995680	0.004312	0.000009	0.000000	0.000000	0.00432014
5	rs7526310	1	804759	T	C	0.123250	-0.000010	0.000145	4.579786e-09	0.005000	0.995424	0.004514	0.000062	0.000000	0.000000	0.00457627
6	rs72631880	1	805556	A	T	0.085425	0.000000	0.000000	0.000000e+00	0.000000	0.993051	0.006489	0.000460	0.000000	0.000000	0.00649419
7	rs11240779	1	808631	A	G	0.772475	0.000000	0.000000	0.000000e+00	0.000000	0.997112	0.002871	0.000018	0.000000	0.000000	0.00288814
8	rs11240780	1	808928	T	C	0.772600	0.000000	0.000000	0.000000e+00	0.000000	0.998095	0.001891	0.000014	0.000000	0.000000	0.00190496
9	rs57181708	1	809876	G	A	0.101175	0.000003	0.000039	2.789843e-10	0.005000	0.996498	0.003501	0.000001	0.000000	0.000000	0.00350165

Practical Steps

4. QC target data

- Applying some QC to target data to minimize noise and genotyping errors is recommended:
 - sample-level filters: limit to a single genetic ancestry, drop individuals with low genotyping rate
 - SNP-level filters: drop SNPs with low call rate, MAF, HWE P-value (genotyped SNPs), imputation accuracy (imputed SNPs)
 - If you are using imputed data, **use dosages rather than hard calls**. Hard calls don't account for imputation uncertainty!

Practical Steps

5. Matching weights and genotype data

Script: '04_Align_genotype_data_with_SBayesRC.sh'

- Ensure SNPs and alleles in the genotype data align with those in the *.snpRes file
- Ambiguous SNPs (i.e. A/T and C/G) are removed
- Strand flipping is done where required (e.g. A/G vs T/C)
- Allele flipping is done where required (e.g. A/G vs G/A)
- Changes are made to the genotype data. The SBayesRC *.snpRes file remains unchanged

Practical Steps

6. Scoring you PGS using PLINK

Script: '05_Plink_PGS_scores.sh'

- Use PLINK with the PGS weights (SBayesRC *.snpRes file) and the aligned genotype data (.bed/.bim/.fam files) to create PGS scores for the genotyped individuals
- Key output file: *.profile

FID	IID	PHENO	CNT	CNT2	SCORESUM
HG00096	HG00096	-9	961868	136436	-0.008254
HG00097	HG00097	-9	961868	137816	-0.025151
HG00099	HG00099	-9	961868	136648	-0.037584
HG00101	HG00101	-9	961868	137351	-0.005926
HG00102	HG00102	-9	961868	136682	-0.02834
HG00103	HG00103	-9	961868	137204	-0.033321
HG00105	HG00105	-9	961868	136604	-0.036338
HG00107	HG00107	-9	961868	136511	-0.039837
HG00108	HG00108	-9	961868	136579	-0.000696

Thank you