THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Introduction to Polygenic Prediction Methods

Jian Zeng

The University of Queensland

j.zeng@uq.edu.au

Outline

- Introduction to polygenic scores (PGS)
- Evaluation and application of polygenic scores
- Basic method to predict polygenic scores (C+PT)
- More advanced methods
- Using GWAS summary statistics and functional annotations

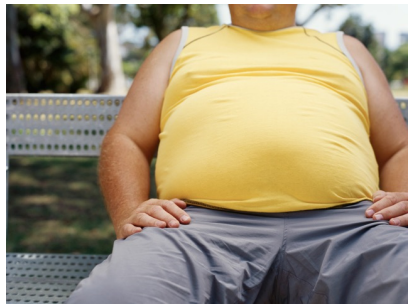
Polygenic scores (PGS)

Polygenic scores (PGS) predict individual genetic values of complex traits using genome variations.

Polygenic risk scores (PRS) are predictors of the genetic susceptibilities of individuals to diseases.

Height

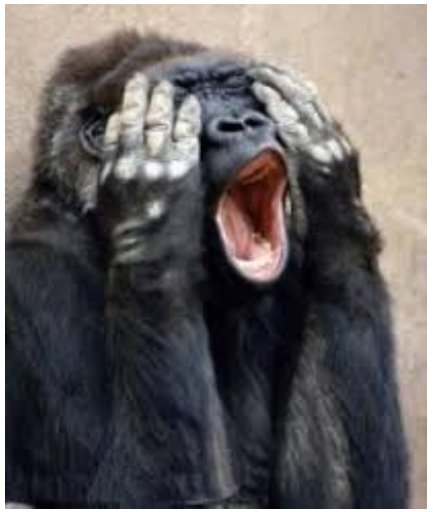
Obesity



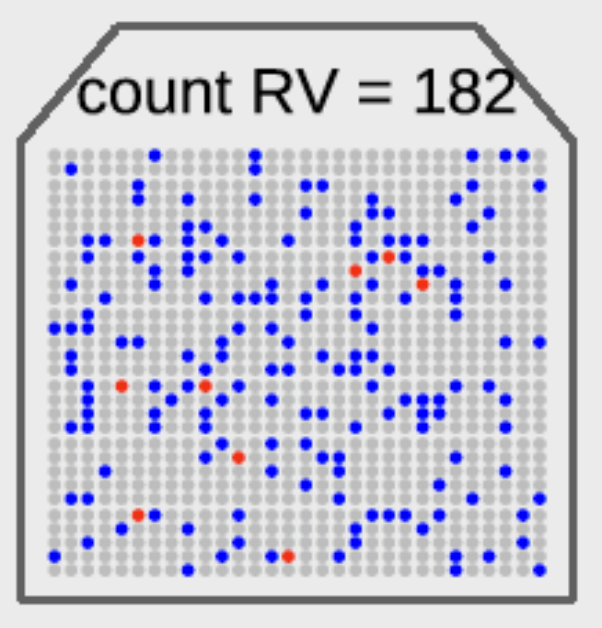
Schizophrenia

What's in a name?

- **PRS**- Polygenic risk score
- **GPRS**- Genomic or genetic profile risk score
- **PGS** -Polygenic score
- **GRS** - Genetic risk score
- **rsPS** – restricted to significant polygenic score
- **gePS** – global extended polygenic score
- **Multi-SNP score** (usually this uses only single nucleotide polymorphisms (SNPs) that are genome-wide significant, hence the same as gePS)
- **MetaGRS** – a PRS constructed from genetic data for the disease/trait of interest plus from other correlated traits
- **MTAG-GRS/PRS** a PRS constructed from GWAS data from multiple correlated traits
- **Genetic score**
- **Genotypic score**
- **Allele score**
- **Profile score**
- **Linear predictor** (this of course is a generic term, but has been used to describe PRS when risk alleles are the only predictors)



Polygenic disease for an individual



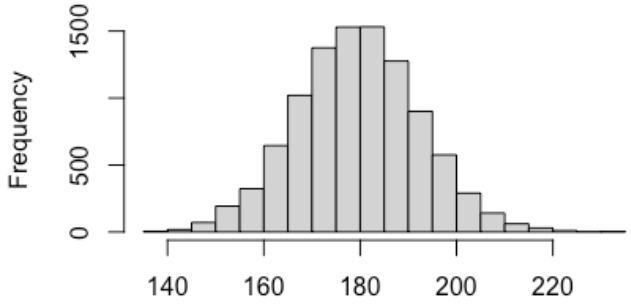
900 DNA polymorphic sites

RV =risk variant

Frequency of risk variant at each site: 0.1 (p)

Average person 900*2*0.1 = 180 risk variant

Mean +/- 3SD: 142 to 218



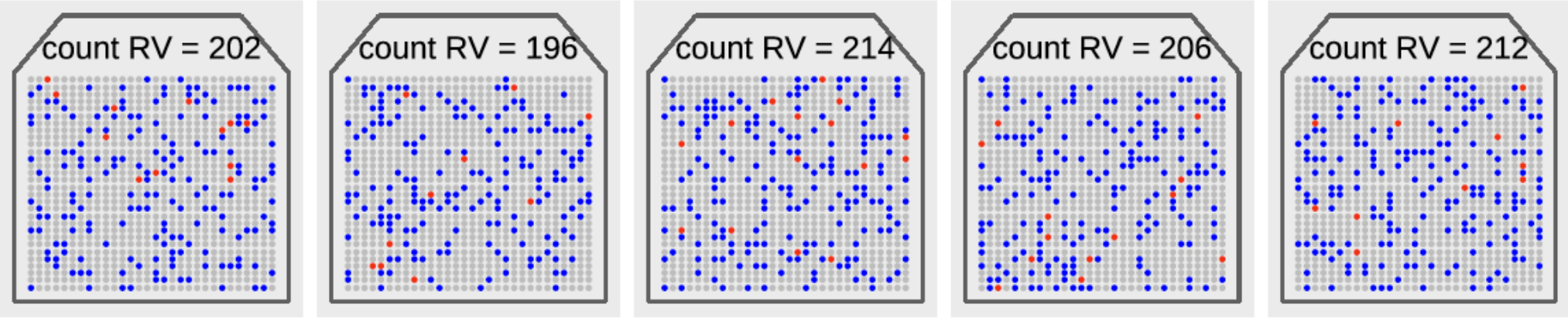
Count of RV in population

0 Grey: Homozygote no risk alleles (or equivalently 2 protective alleles)
1 Blue : Heterozygote one risk allele (and one non-risk/protective allele)
2 Red: Homozygote two risk alleles

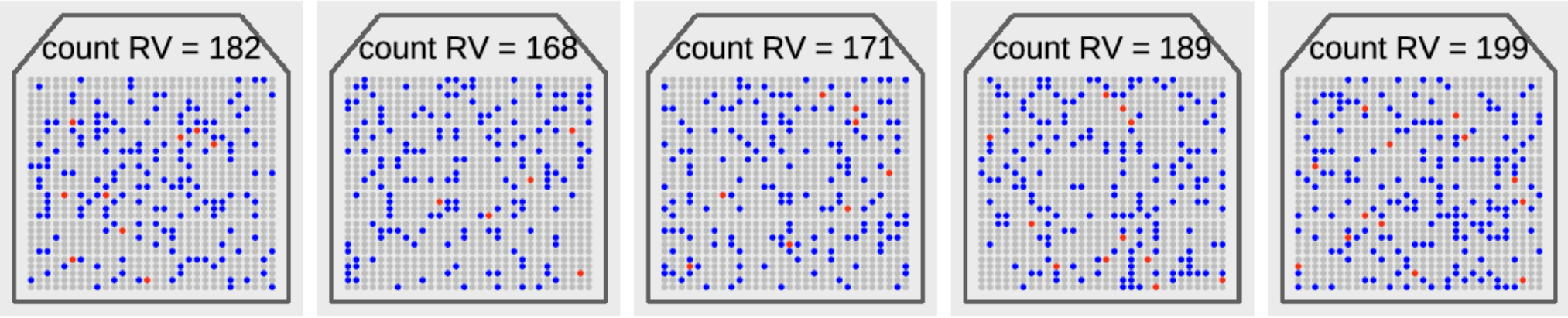
Toy example

Polygenic disease for an individual

Affected over lifetime

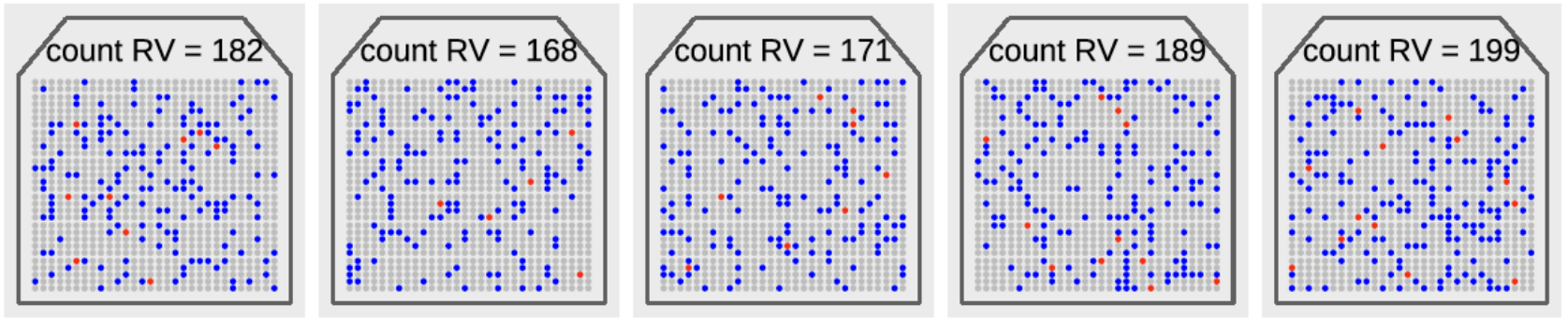


Not affected over lifetime



- We all carry risk variants for all diseases.
- Robustness
- Those affected carry a higher burden.
- Non-genetic factors contribute to risk too
- Each person carries a unique portfolio of risk alleles

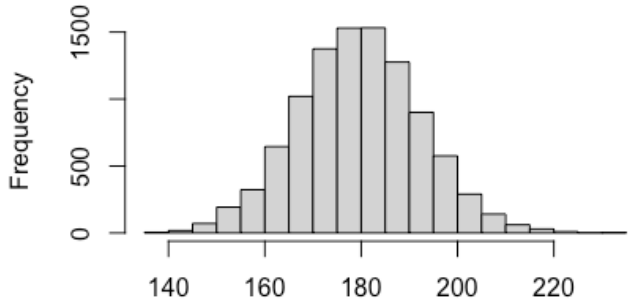
Polygenic score



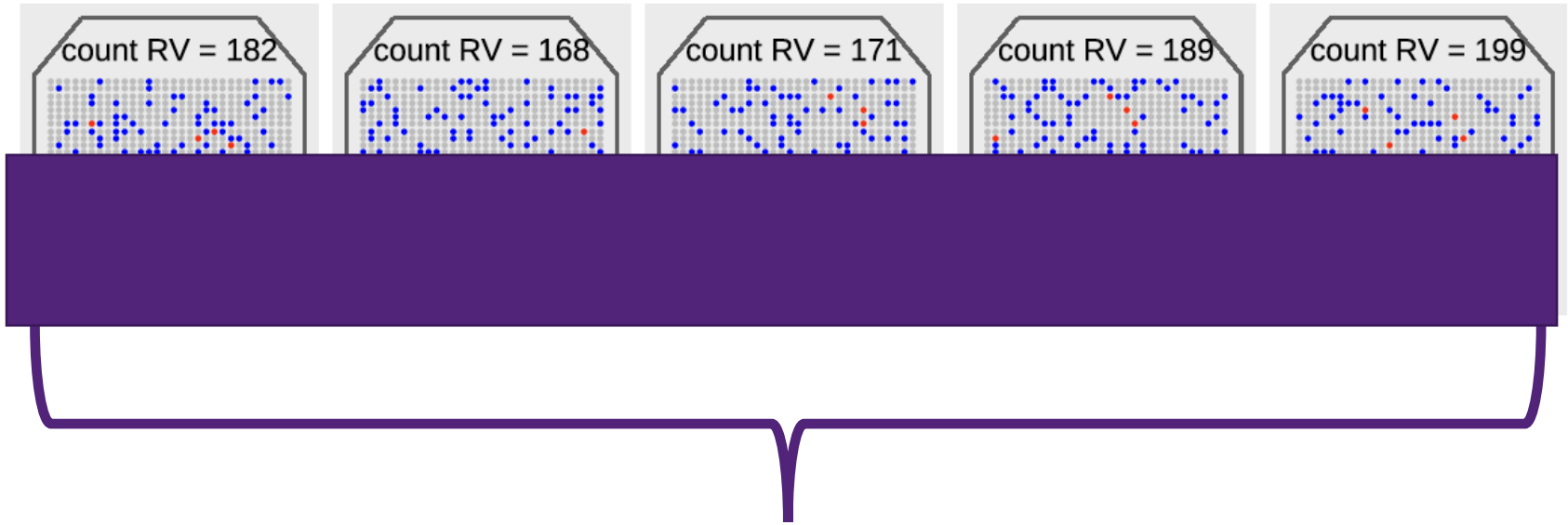
→ "True" polygenic score

Genetic variance between people attributed to all genetic factors V(A)

$h^2 = \frac{V(A)}{V(P)}$ heritability



Polygenic score



→ "True" polygenic score

Not all variants captured on genotyping arrays

Genetic variance between people attributed to all genetic factors V(A)

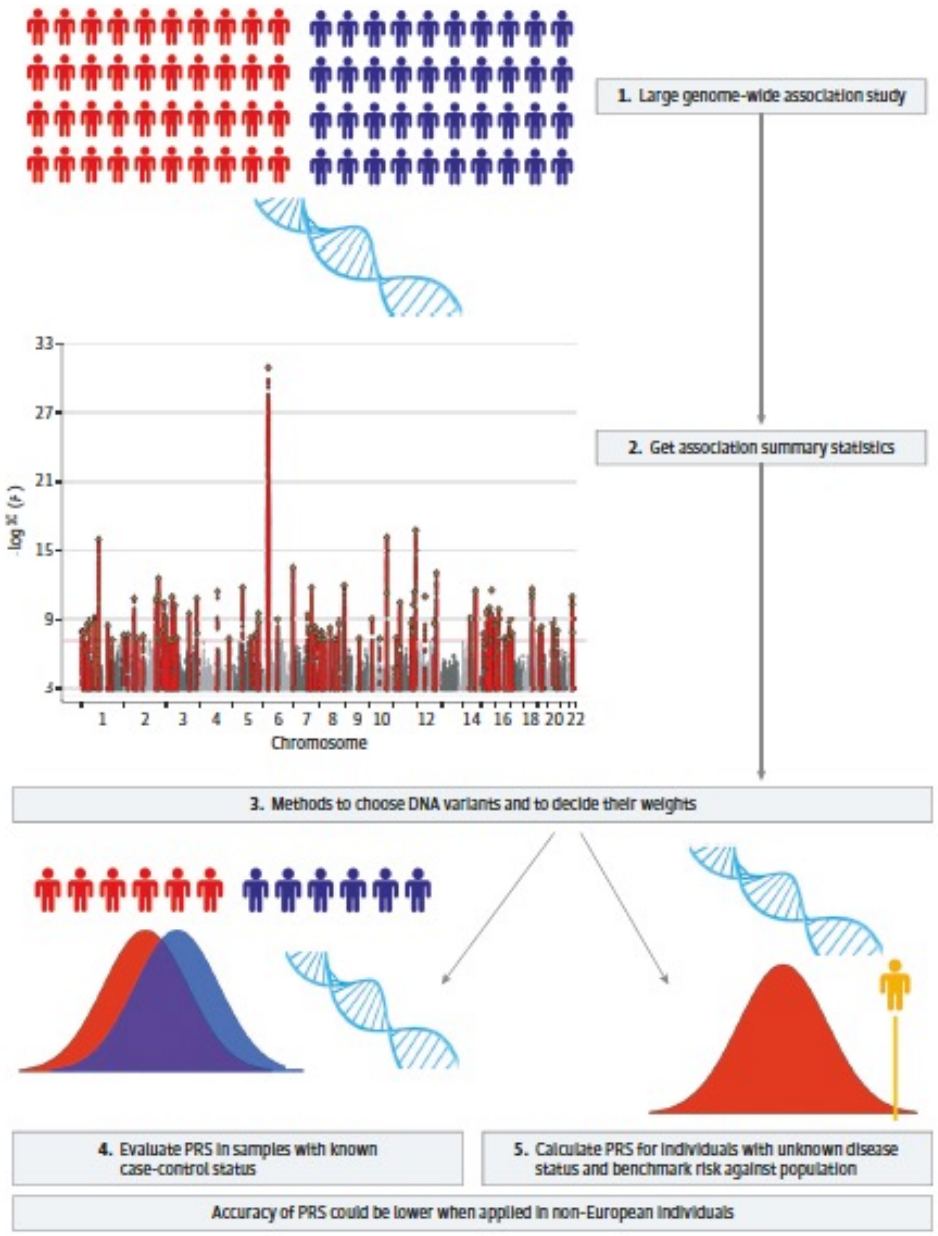
$h^2 = \frac{V(A)}{V(P)}$ heritability

Genetic variance between people attributed to all genetic factors associated with SNPs on genotyping arrays

$h^2_{SNP} = h^2_g = \frac{V(A:SNP)}{V(P)}$

$\mathrm{SNP-based\ heritability}$

Polygenic scores



- A weighted count of risk alleles

$$PGS = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

0, 1 or 2
Risk alleles

Which SNPs?

What weights?

- Don't need to know causal variants for prediction!
- Prediction can be based on correlated variants.

4. Evaluate
Y= b*PGS + e
R^2 = var(b*PGS)/Var(Y)

AUC statistic:
Probability that a case ranks higher than a control

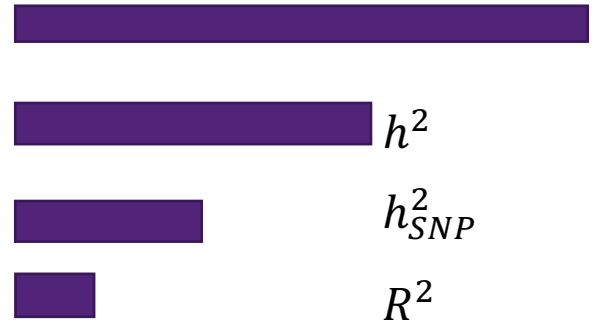
Limitations in prediction accuracy

- PGS have a **theoretical** upper limit dependent on the **heritability of the trait** (how much of the variance of trait values between people is attributed to genetic factors).

- PGS have a **technical** upper limit associated with the proportion of **variance tagged** by the DNA variants measured.

- PGS have a **practical** upper limit dependent on the **sample size of the discovery sample** used to estimate effect sizes of risk alleles, and the **quality** of the discovery sample.

- PGS can be pushed closer to the technical upper limit by the **statistical methodology** used to generate the optimal weighting given to the risk alleles, and new methods integrate new biological data.



Schizophrenia

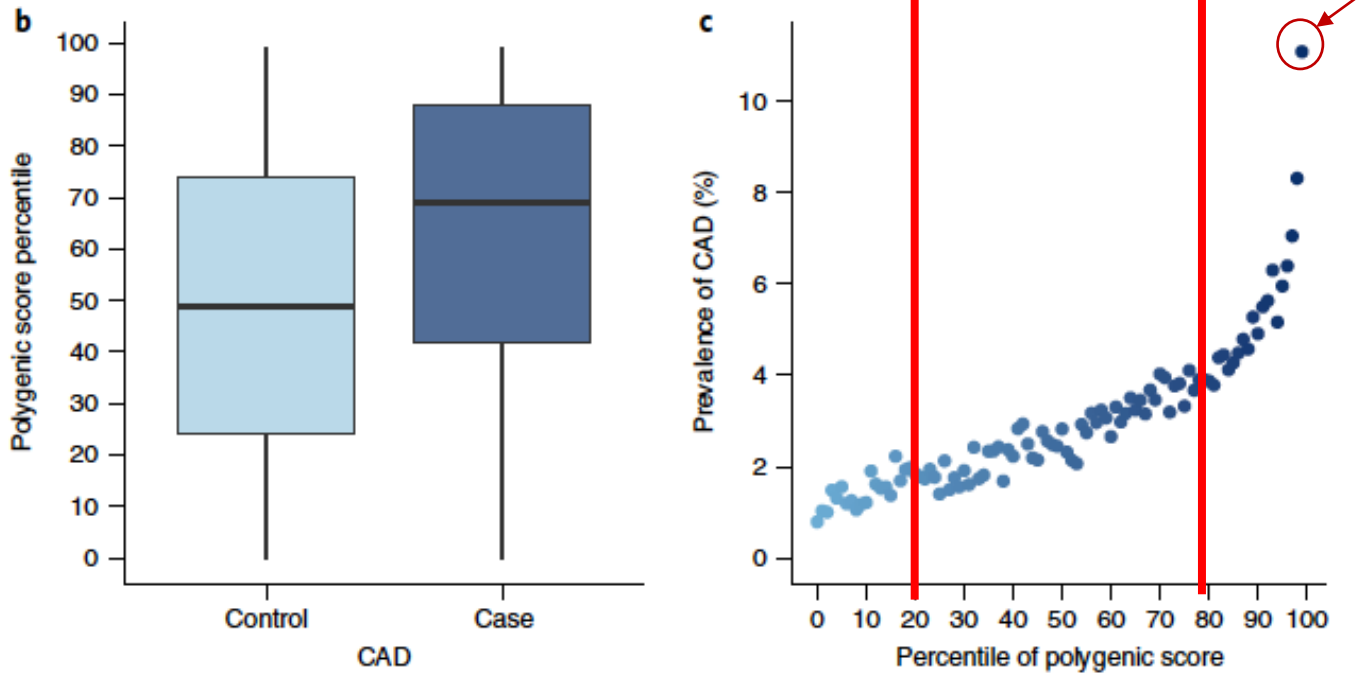
Max:
25% Liability
AUC 0.84

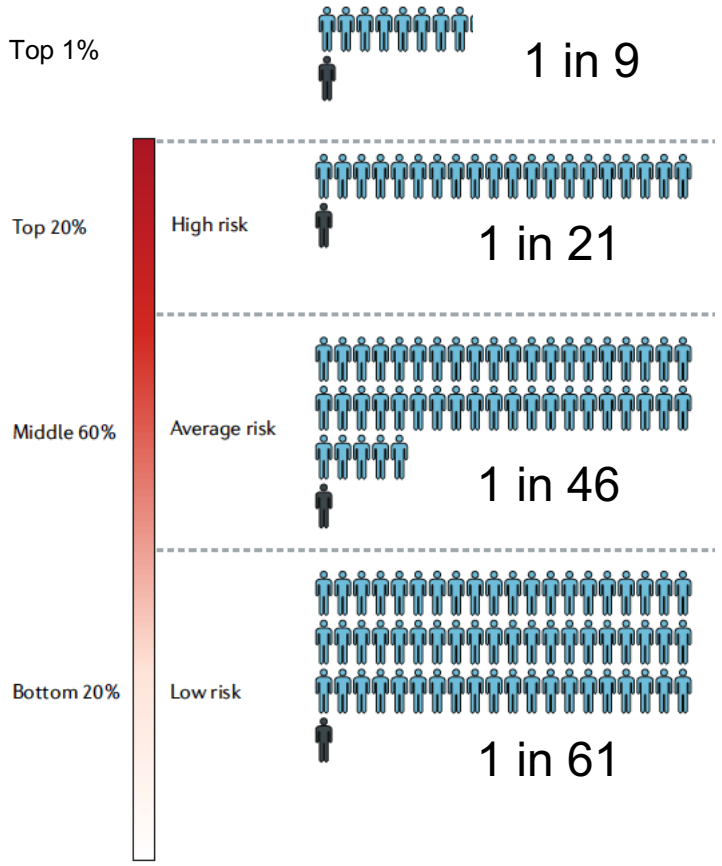
Current:
11% Liability
AUC 0.74

Polygenic scores cannot be highly accurate predictors of phenotypes

Evaluating PRS

Different views of the same data



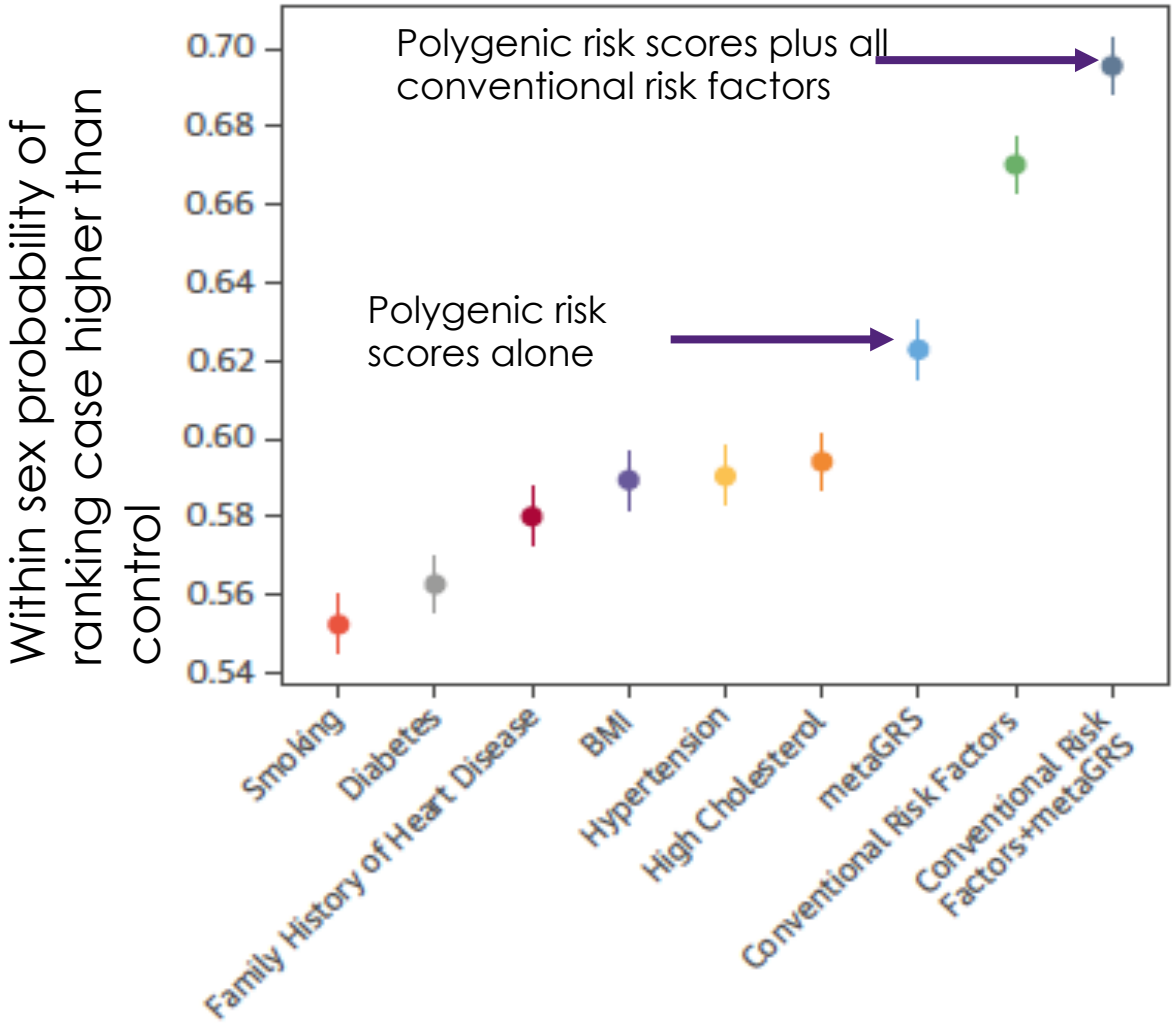


Khera et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics

Torkamani et al, Nat Rev Genetics, 2018

Increase total risk prediction accuracy

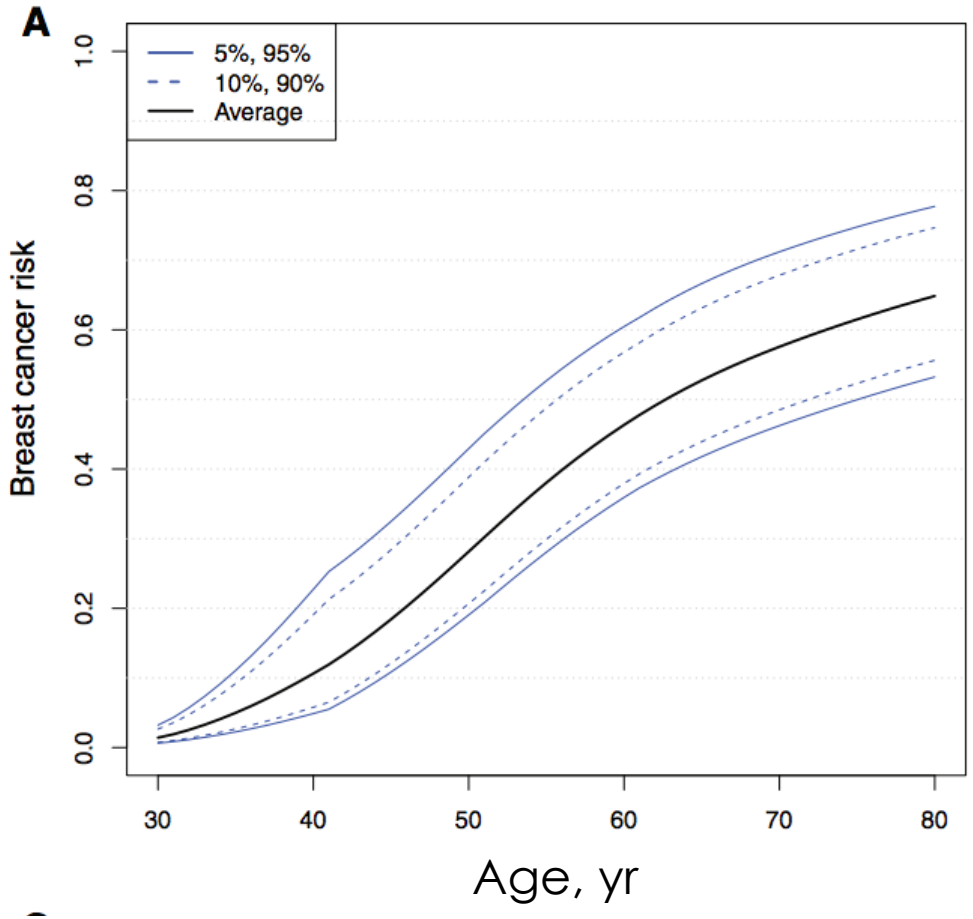
Combine PRS with conventional risk predictors
Coronary Artery Disease



Inouye et al (2018) Genomic risk prediction of CAD in 480K adults. JACC

Disease heterogeneity within patients

Combine PRS with known risk mutations
Breast cancer



BRCA1 carriers

Kuchenbaecker et al: Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. J Natl Cancer Inst (2017)

Polygenic risk score applications

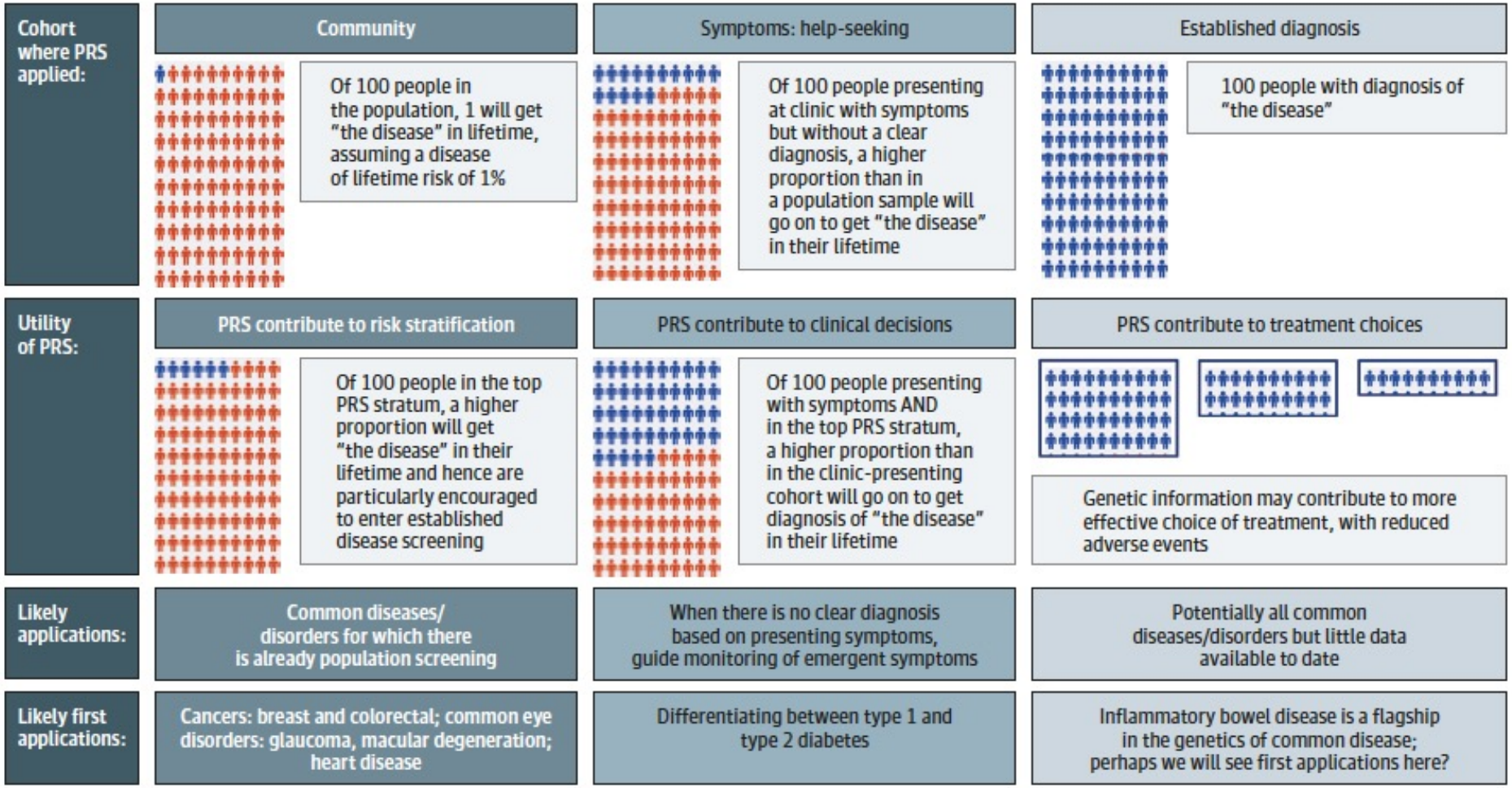
JAMA Psychiatry | Review

From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD

Goal:

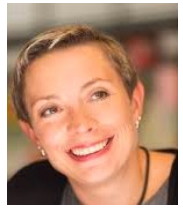
- Understandable by interested clinician
- Technically accurate – backed up in Supplement & Rscript



Naomi Wray, UQ & UoOxford

Graham Murray, UoCambridge

Ian Hickie, UoSydney



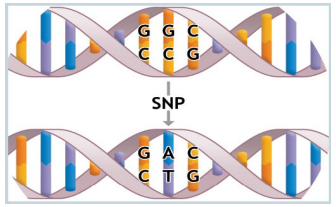
Jehannine Austin, UoBritish Columbia

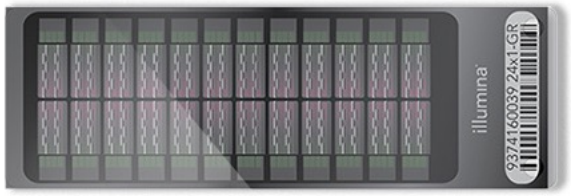
John McGrath, UQ

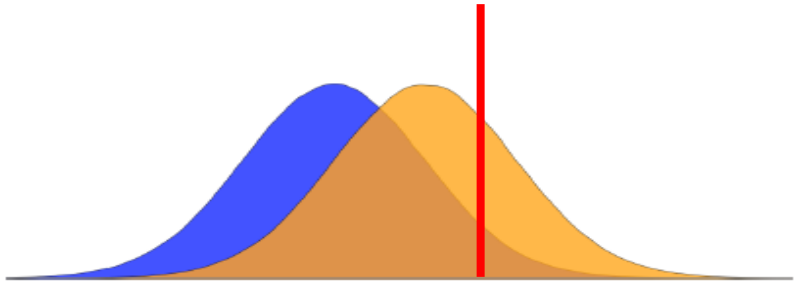
Tian Lin, UQ

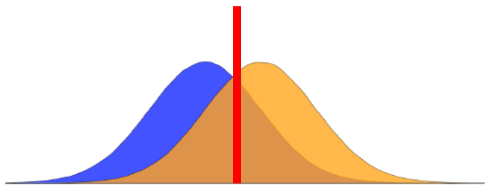
Justify for one disease and the rest come for free!

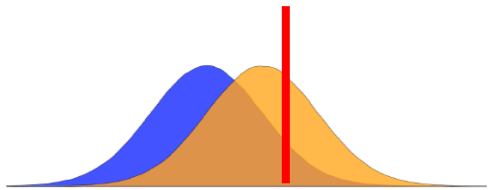
One disease

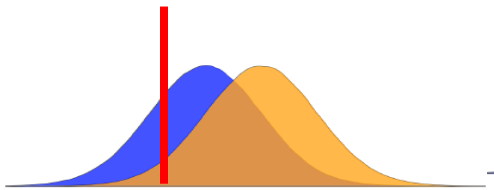


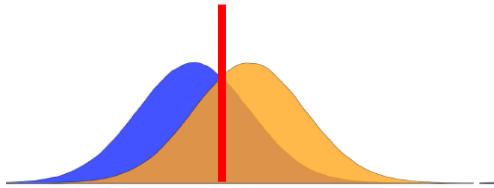


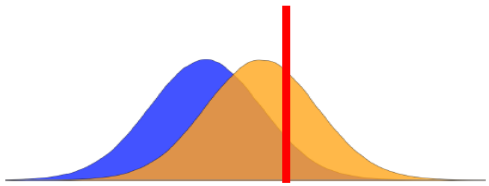


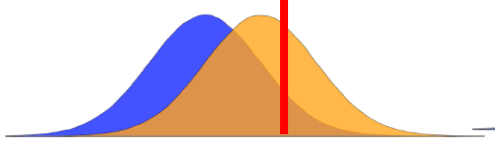

Disease 2

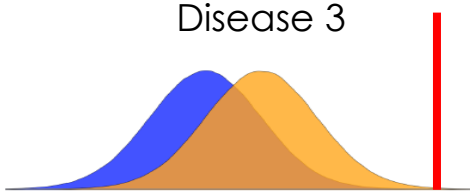

Disease 3

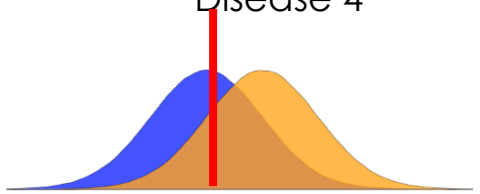

Disease 4

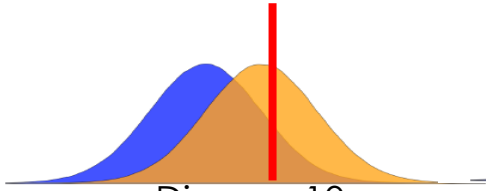

Disease 5

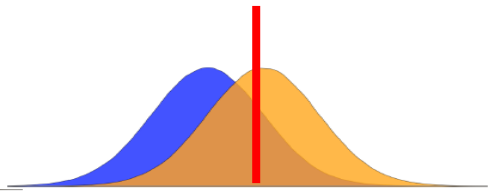

Disease 6

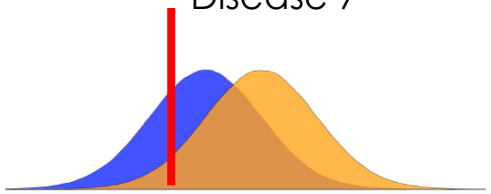

Disease 7

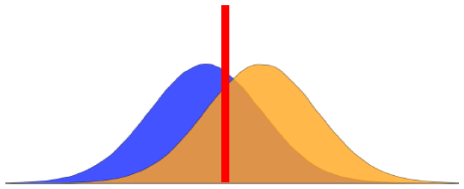

Disease 8

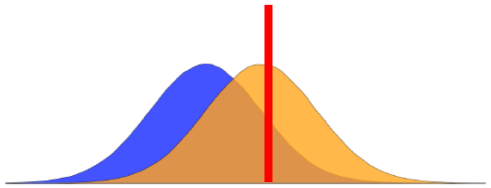

Disease 9

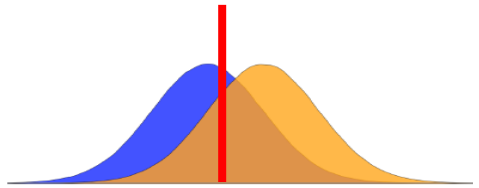

Disease 10

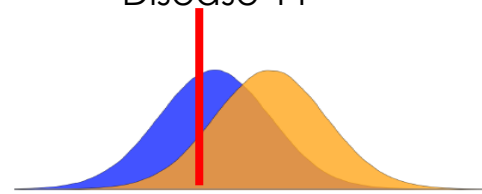

Disease 11


Disease 12


Disease 13


Disease 14


Disease 15


Disease 16

Polygenic risk score methods

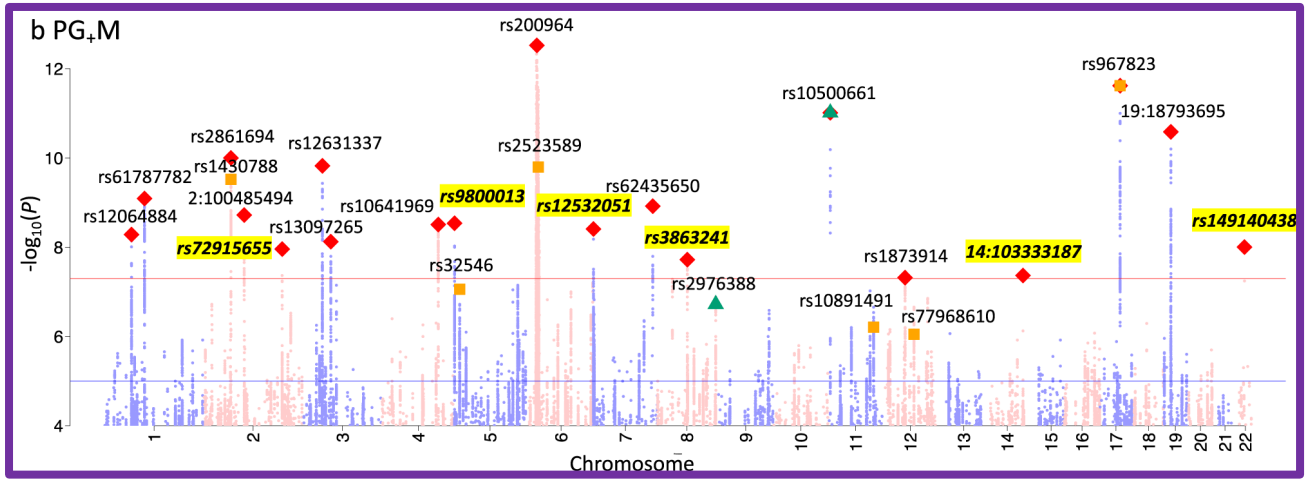
A weighted sum of the count of risk alleles

$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

Clumping & P-value thresholding (C+PT, aka P+T):



Polygenic risk score methods

A weighted sum of the count of risk alleles

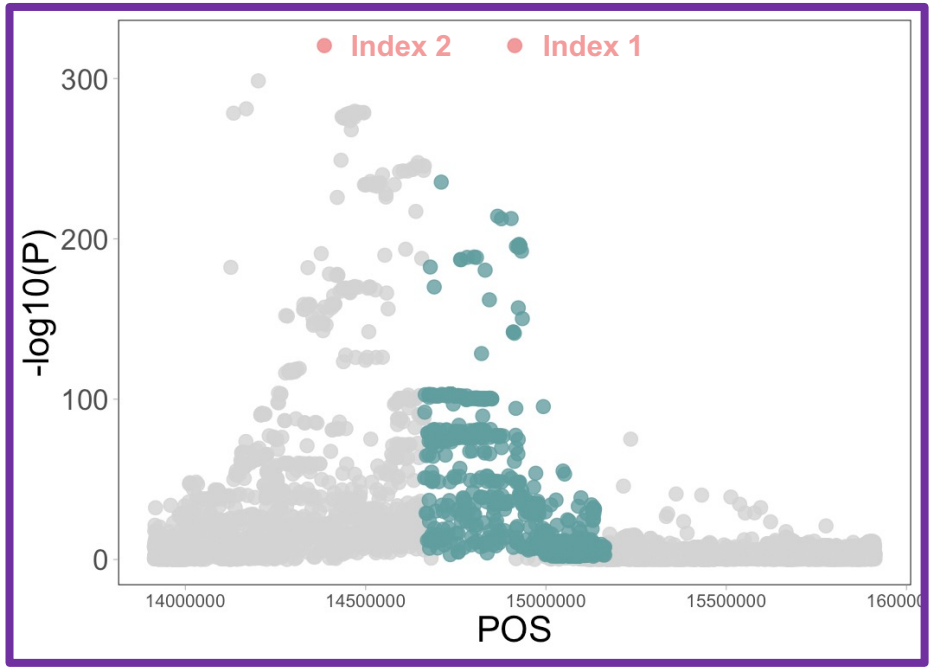
$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

Clumping & P-value thresholding (C+PT, aka P+T):

- Select most associated SNP in tower – LD-based clumping



Polygenic risk score methods

A weighted sum of the count of risk alleles

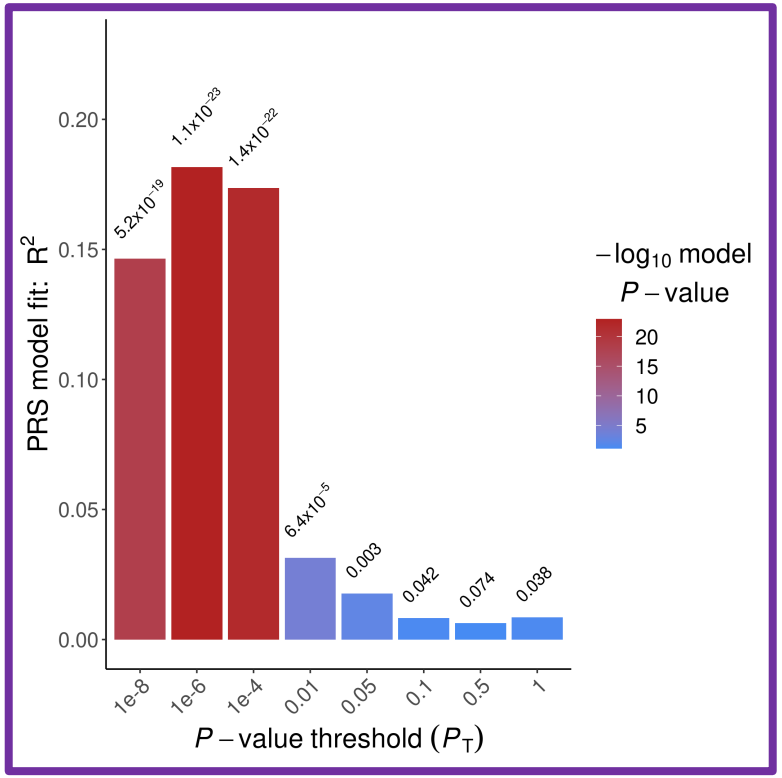
$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

Clumping & P-value thresholding (C+PT, aka P+T):

- Select most associated SNP in tower – LD-based clumping
- Select on a p-value threshold



Polygenic risk score methods

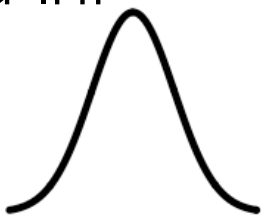
A weighted sum of the count of risk alleles

$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

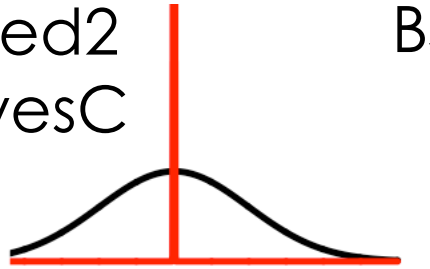
How many SNPs?
Which SNPs?
What weights?

New methods model genetic architecture

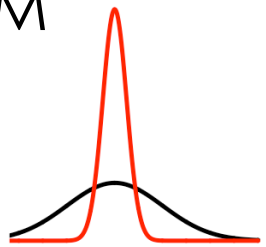
LDpred-Inf
SBLUP



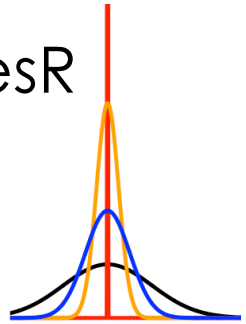
LDPred2
SBayesC



BSLMM



SBayesR

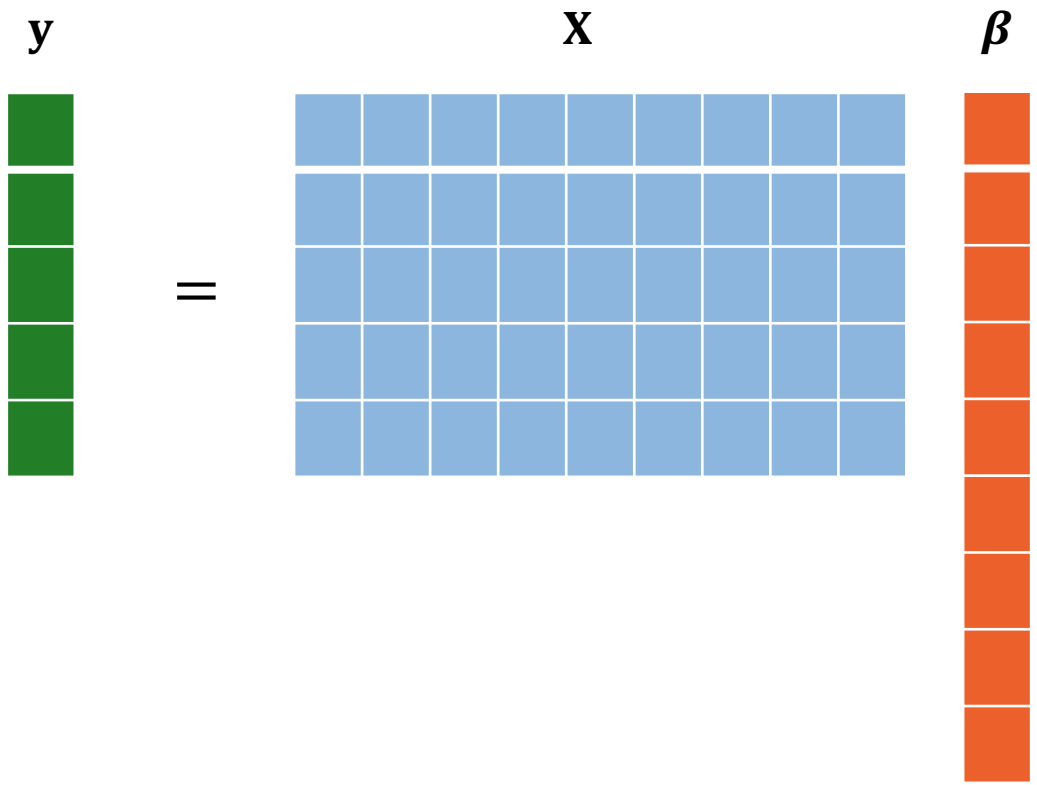


19

Fitting all SNPs

Multiple regression

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$



parameters >> # observations

Bayesian methods

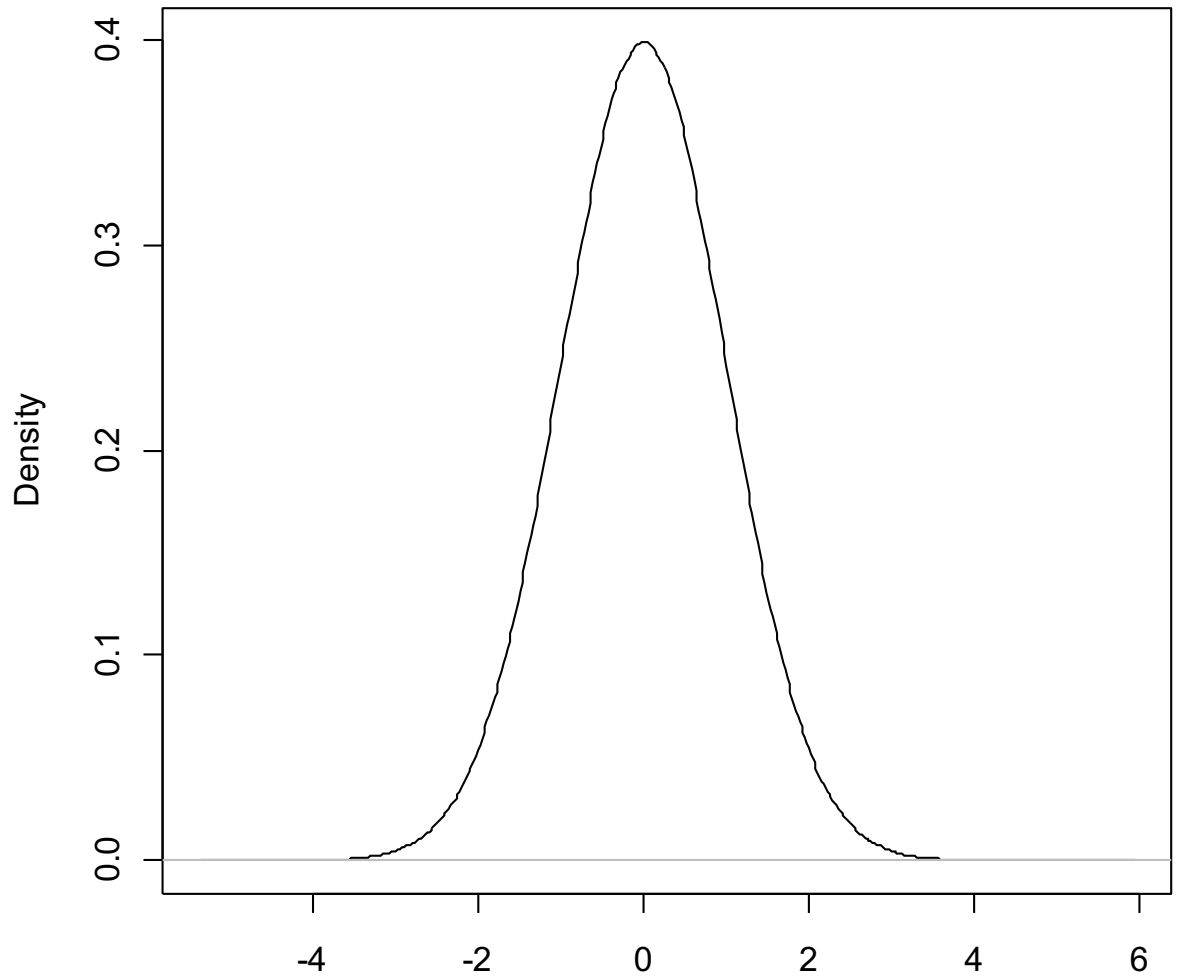
- Bayesian methods can estimate all parameters including SNP effects simultaneously by "borrowing" information across SNPs

- Allow assumptions regarding the distribution of SNP effects

What are SNP effect distributions that make sense?

Distribution of SNP effects

$$\beta_j \sim N(0, \sigma^2_\beta)$$



Assumes SNPs effects are:

- all non-zero
- very small
- normally distributed

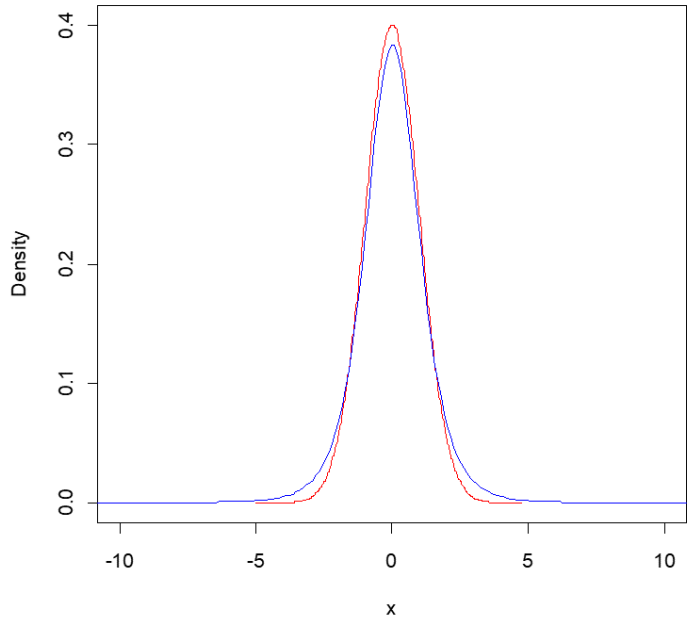
This is BLUP (Best Linear Unbiased Prediction)

How realistic is it?

Assumptions for SNP effect distribution

Alternative distributions

Assumption	Distribution of SNP effects	Method
Small number of moderate to large effects, many small effects	Students t	BayesA



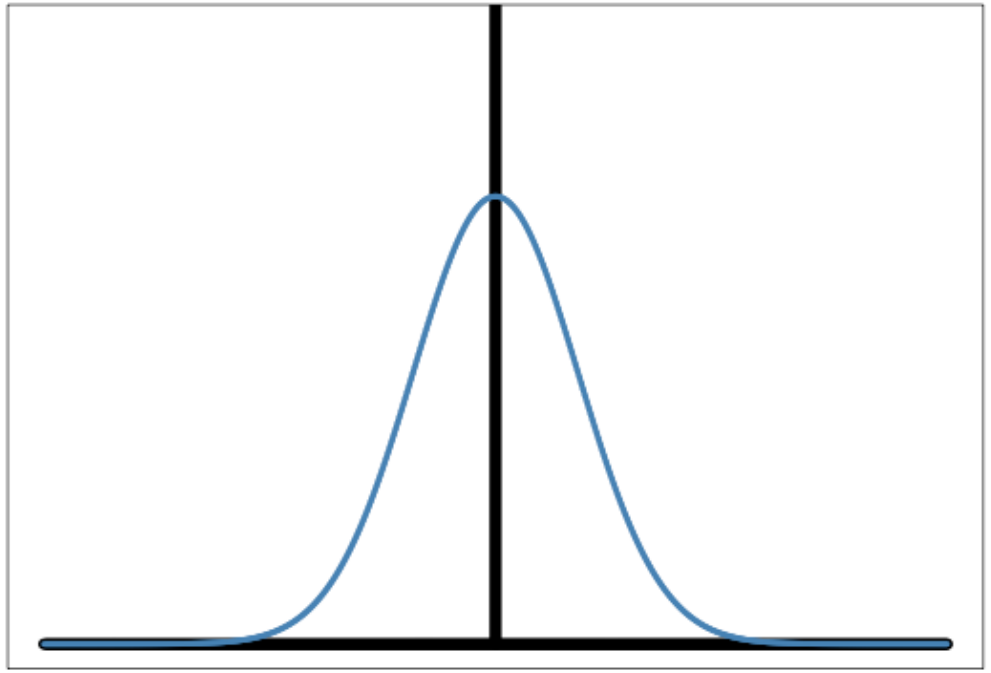
Assumptions for SNP effect distribution

Alternative distributions

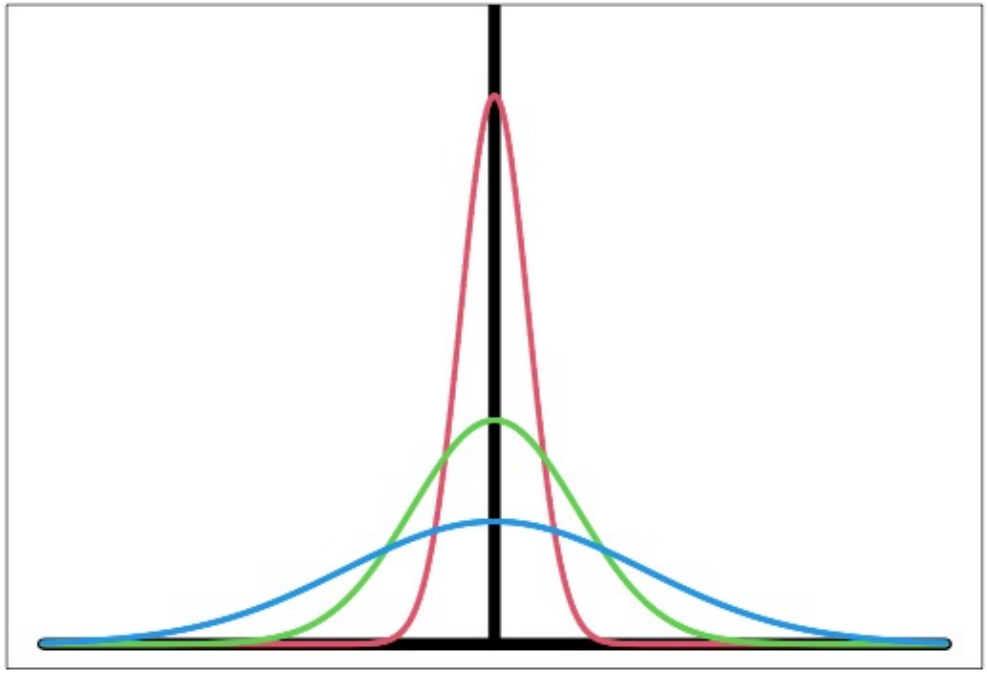
Assumption	Distribution of SNP effects	Method
Small number of moderate to large effects, many small effects	Students t	BayesA
Small number of moderate to large effects, many zero effects	Mixture, spike at zero, Students t	BayesB
Small number of small effects, many zero effects	Mixture, spike at zero, normal distribution	BayesC
Many zero effects, proportion of small effects, some moderate to large effects	Mixture, multiple normals	BayesR

Assumptions for SNP effect distribution

BayesC



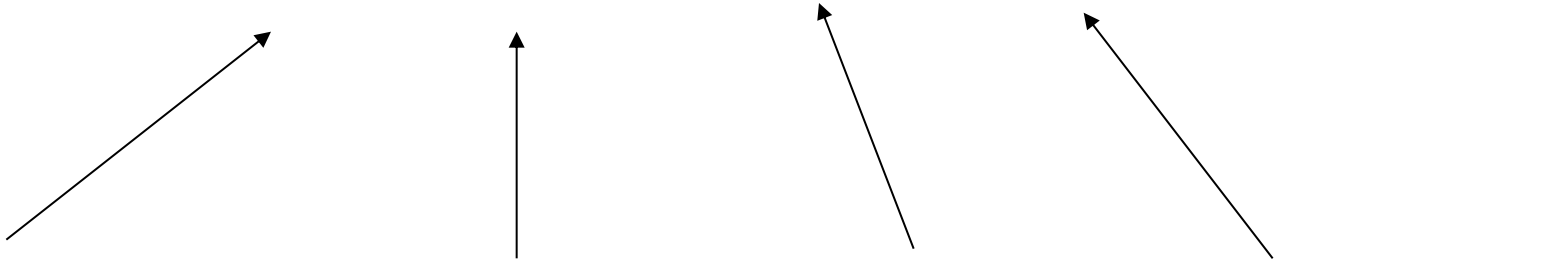
BayesR



How to incorporate this prior knowledge in the estimation of SNP effects?

Bayes theorem

$$P(x \mid y) \propto P(y \mid x)P(x)$$



Probability of parameters x given the data y (posterior)

Is proportional to

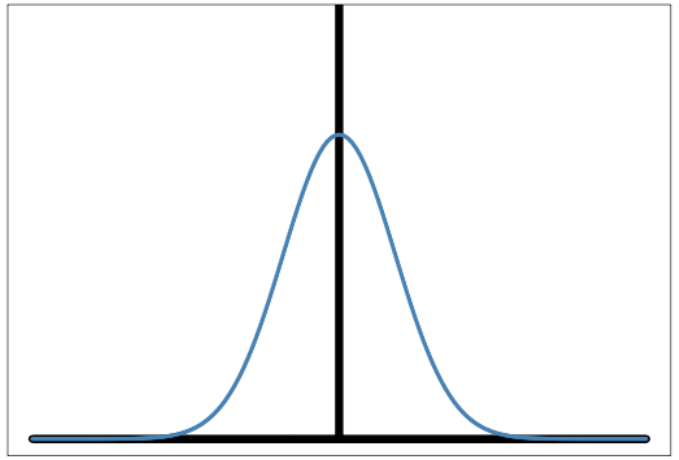
Probability of data y given the x (likelihood of data)

Prior probability of x

Model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\beta_j \begin{cases} \sim N(0, \sigma^2_\beta) & \text{with probability } \pi \\ = 0 & \text{with probability } 1 - \pi \end{cases}$$



Posterior distribution of SNP effects

$$f(\boldsymbol{\beta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta})$$

Posterior mean of SNP effects

$$\widehat{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|\boldsymbol{y}) = \int_{\boldsymbol{\beta}} \boldsymbol{\beta} f(\boldsymbol{\beta}|\mathbf{y}) \mathrm{d}\boldsymbol{\beta}$$

$$= \int_{\beta_1} \dots \int_{\beta_m} \beta_j (\sigma_e^2)^{-\frac{n}{2}} \exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma_e^2}\right\} \prod_{j=1}^{m} \left[(\sigma_\beta^2)^{-\frac{1}{2}} \exp\left\{-\frac{\beta_j^2}{2\sigma_\beta^2}\right\} \pi + \varphi_0(1-\pi)\right] \mathrm{d}\beta_1 \dots \mathrm{d}\beta_m$$

- Cannot solve directly → no closed form solution
- Use Markov chain Monte Carlo (MCMC) algorithm!

Gibbs Sampling

A special case of MCMC to sample from posterior distribution of each parameter **conditional** on all other parameters.

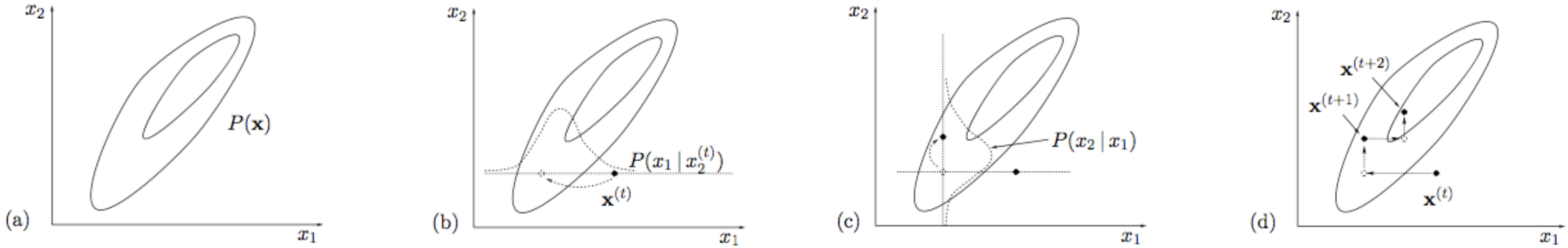


Figure source

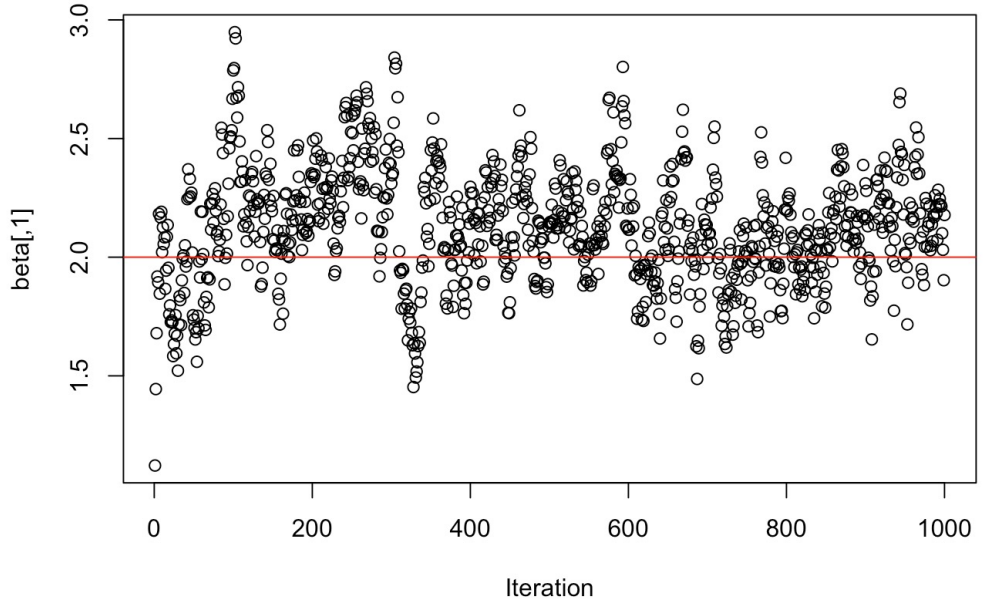
Gibbs sampling

- Set starting values for $(\mu, \boldsymbol{\beta}, \sigma_\beta^2, \pi, \sigma_e^2)$
- Then (for many iterations)
  - For each SNP, sample β_j conditional on other parameters
  - Sample $\mu, \sigma_\beta^2, \pi, \sigma_e^2$ with updated $\boldsymbol{\beta}$

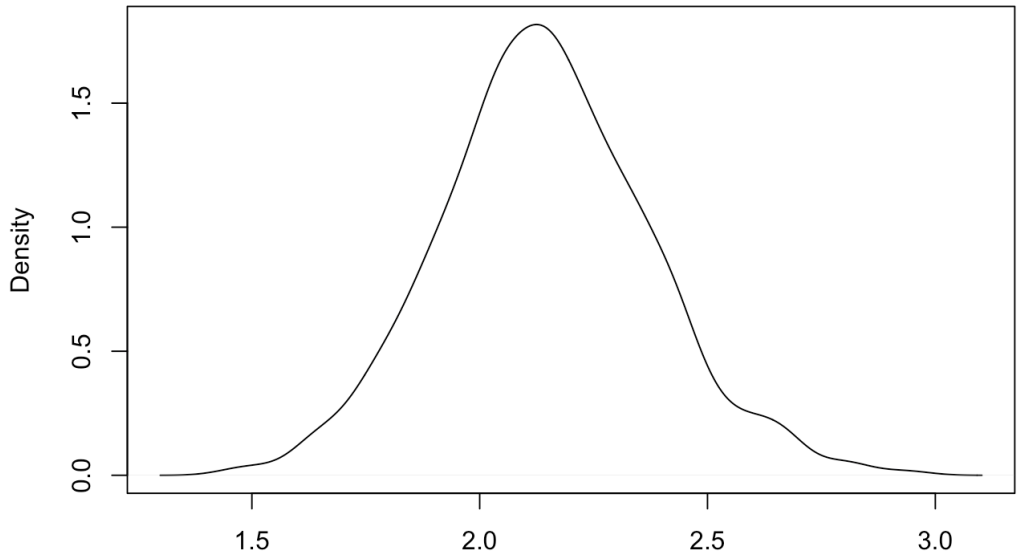
Samples reconstruct posterior distributions of parameters

Gibbs sampling

Trace plot



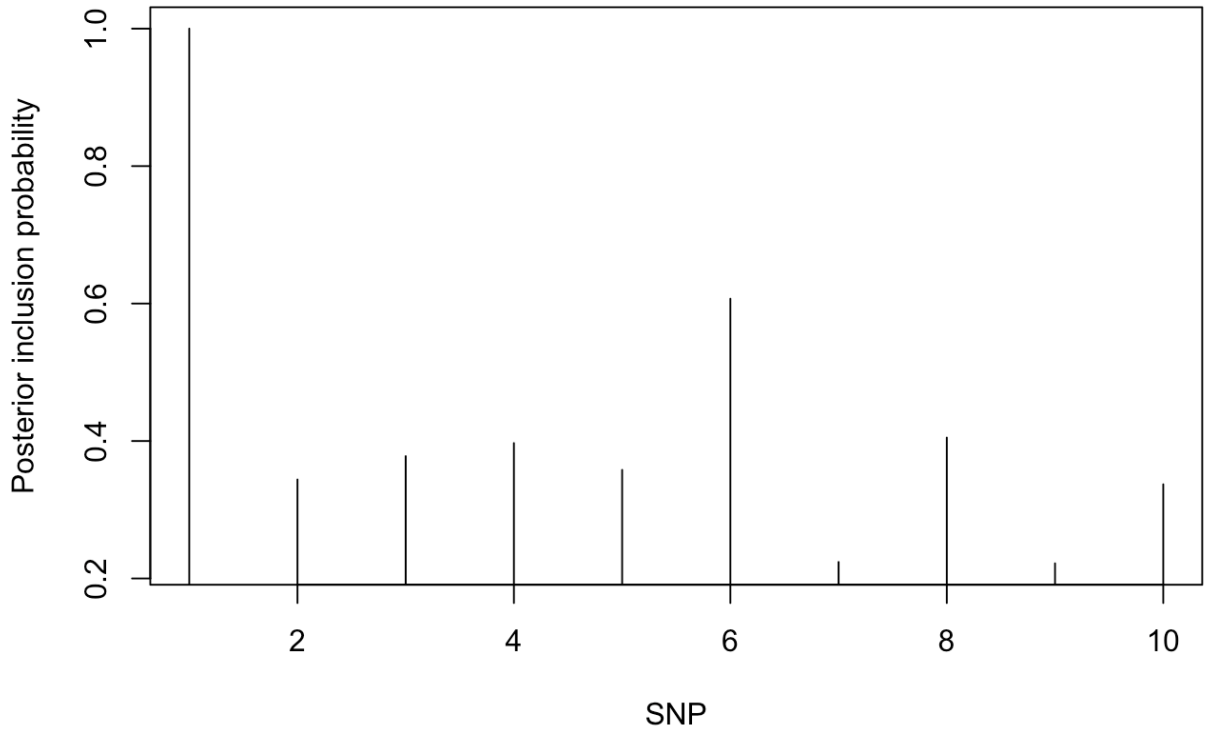
Posterior distribution



Posterior mean is used as the point estimate of the SNP effect

As a method of fine-mapping

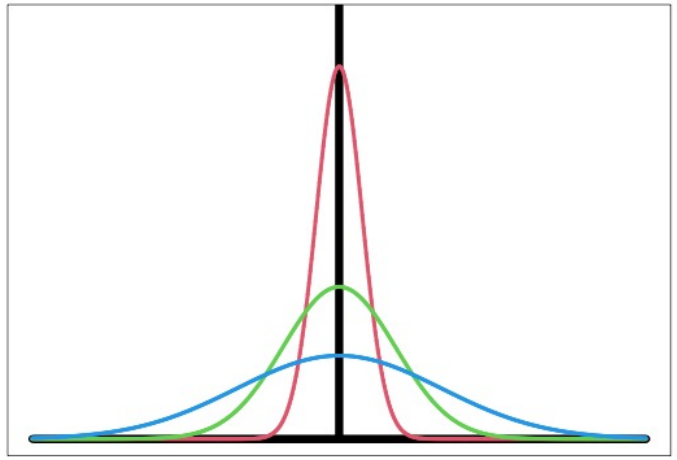
Posterior inclusion probability (PIP):
probability that the SNP is included in the model with a nonzero effect.



Model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

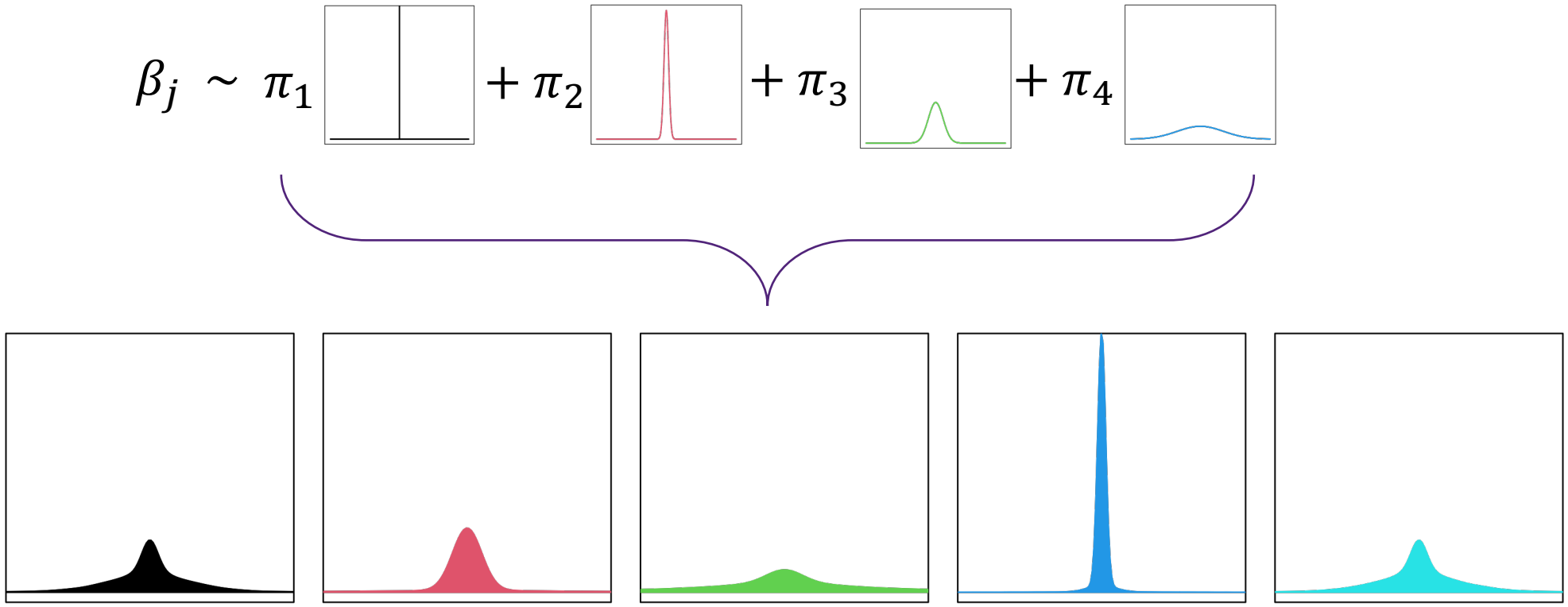
$$\beta_j|\pi, \sigma^2_\beta = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0, \gamma_2\sigma^2_\beta) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C\sigma^2_\beta) & \text{with probability } 1 - \sum_{c=1}^{C-1} \pi_c, \end{cases}$$



$$\boldsymbol{\gamma} = (0, 0.01, 0.1, 1.0)'$$

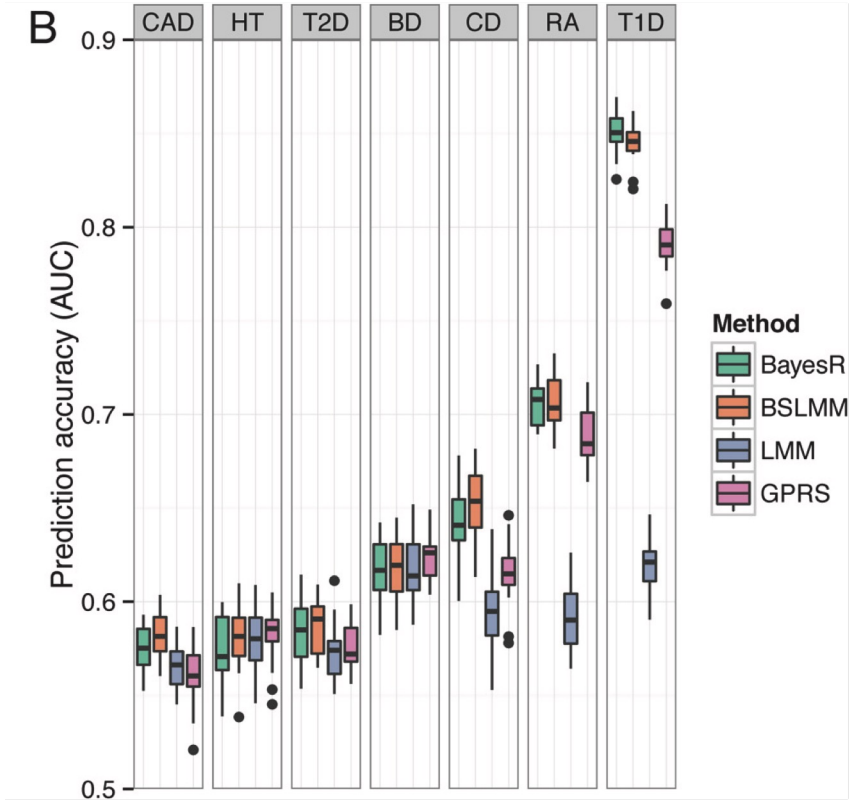
BayesC is a special case of BayesR with two components

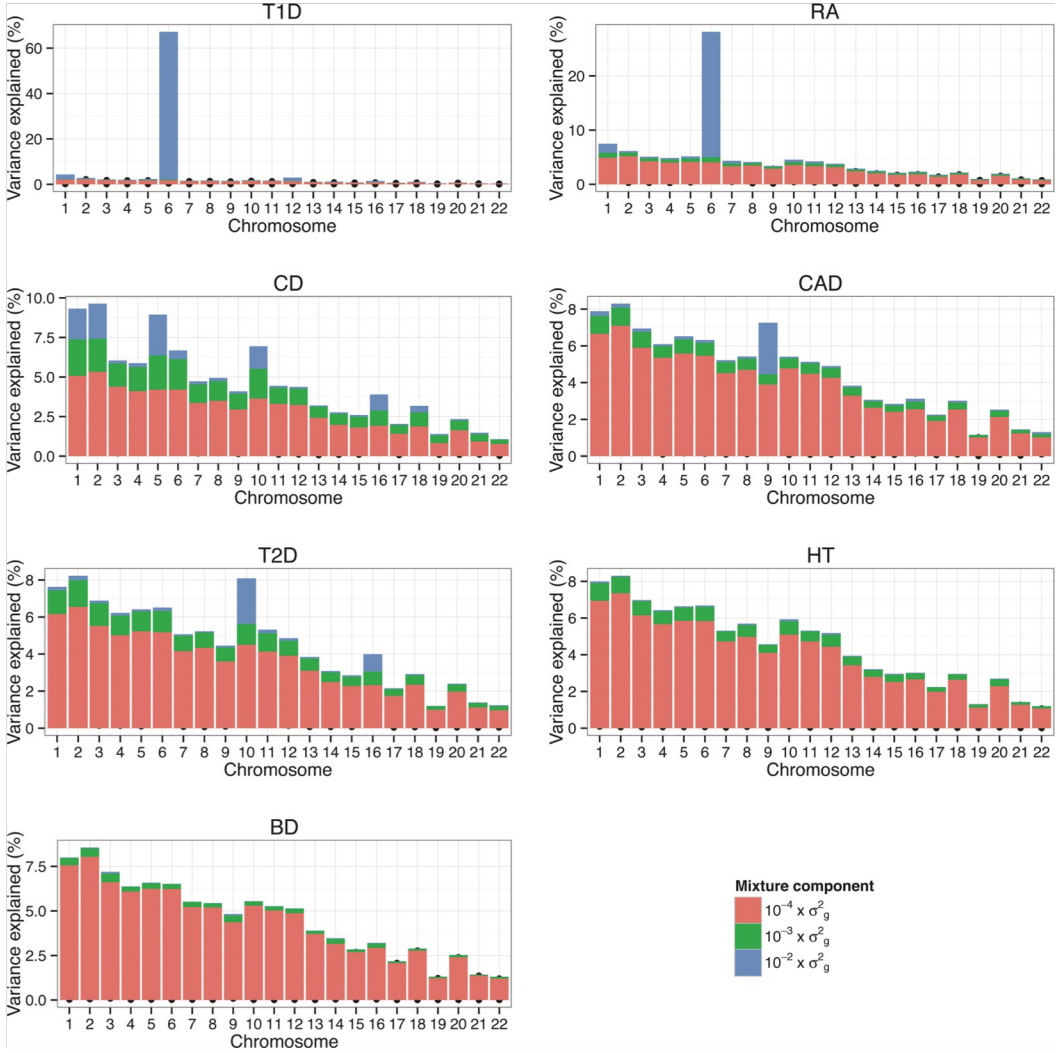
Why use multi-normal mixture?



Account for almost any distribution!

Prediction of disease risk in humans





Moser et al PLoS Genetics 2015

Advancing the methodology

Summary-data-based model

Consider an individual-data model with a standardised genotype matrix **X:**

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Multiply both sides by $\frac{1}{N}\mathbf{X}'$ gives

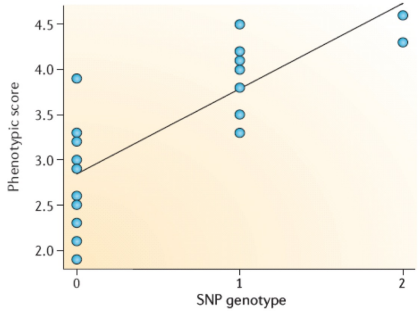
$$\frac{1}{N}\mathbf{X}'\mathbf{y} = \frac{1}{N}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \frac{1}{N}\mathbf{X}'\mathbf{e}$$

$$\mathbf{b} = \mathbf{R}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

GWAS marginal SNP effects

$$Var(\boldsymbol{\epsilon}) = \frac{1}{N}\mathbf{R}\sigma_e^2$$

LD correlation matrix



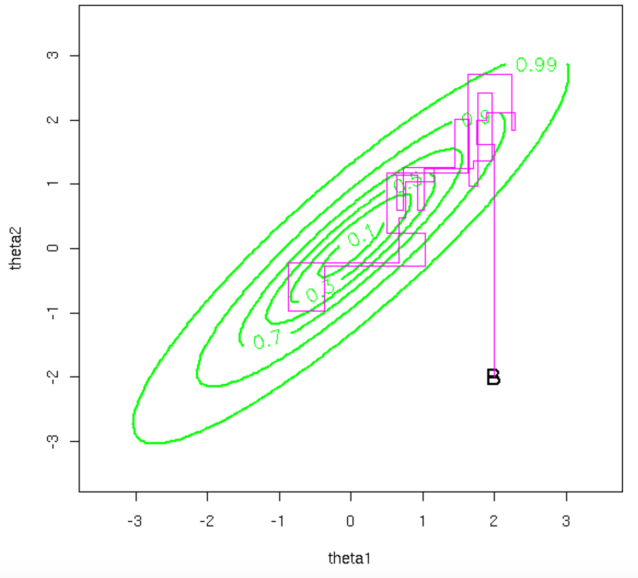
Compare BayesR and SBayesR algorithms

Gibbs sampling

Full conditional distribution for β_j, if in a nonzero dist'n,

$$f\left(\beta_j \mid \mathbf{b}, else\right) = N\left(\frac{r_j}{C_j}, \frac{\sigma_e^2}{C_j}\right)$$

where



Individual-level data

$$r_j = \mathbf{X}_j'\left(\mathbf{y} - \sum_{k \neq j} \mathbf{X}_k \beta_k\right)$$

$$C_j = \mathbf{X}_j'\mathbf{X}_j + \frac{\sigma_e^2}{\gamma_j \sigma_\beta^2}$$

Summary-level data

$$r_j = nb_j - \sum_{k \neq j} nR_{jk}\beta_k$$

$$C_j = n + \frac{\sigma_e^2}{\gamma_j \sigma_\beta^2}$$

Potential issue

- In principle, SBayes and Bayes are equivalent methods when **same data** are used ($\mathbf{X}'\mathbf{y}$ and $\mathbf{X}'\mathbf{X}$ are sufficient statistics).

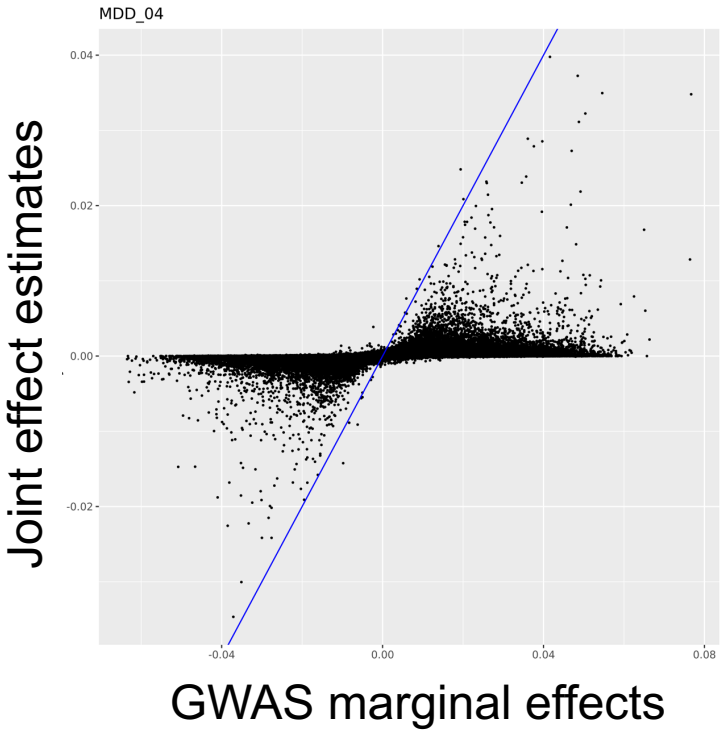
- However, when LD is estimated from a reference sample, SBayes is only an approximation to Bayes.

- Whether the difference is negligible depends on the heterogeneity in LD between the GWAS and LD reference samples.

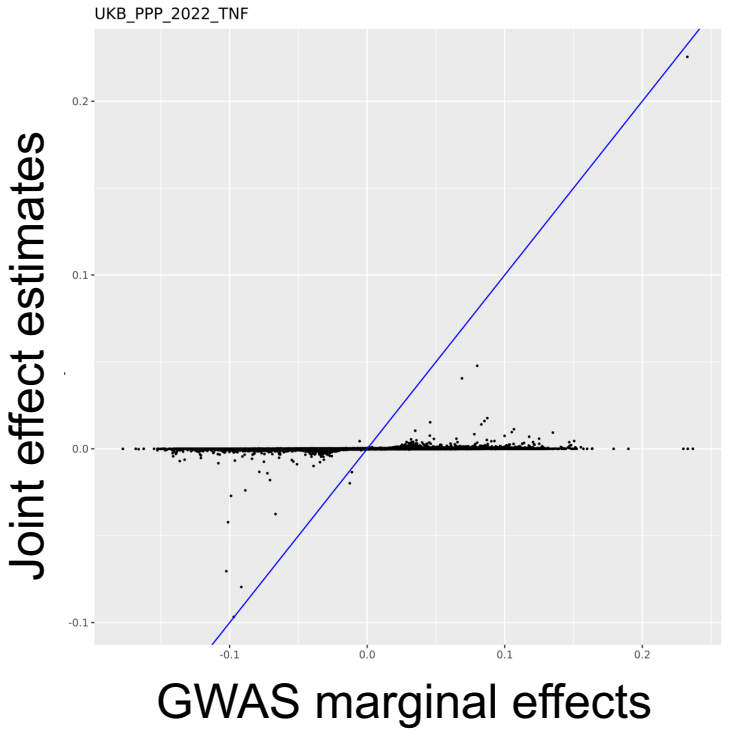
Always good to check SNP effect estimates

GWAS marginal effect size vs. Estimated joint effect size

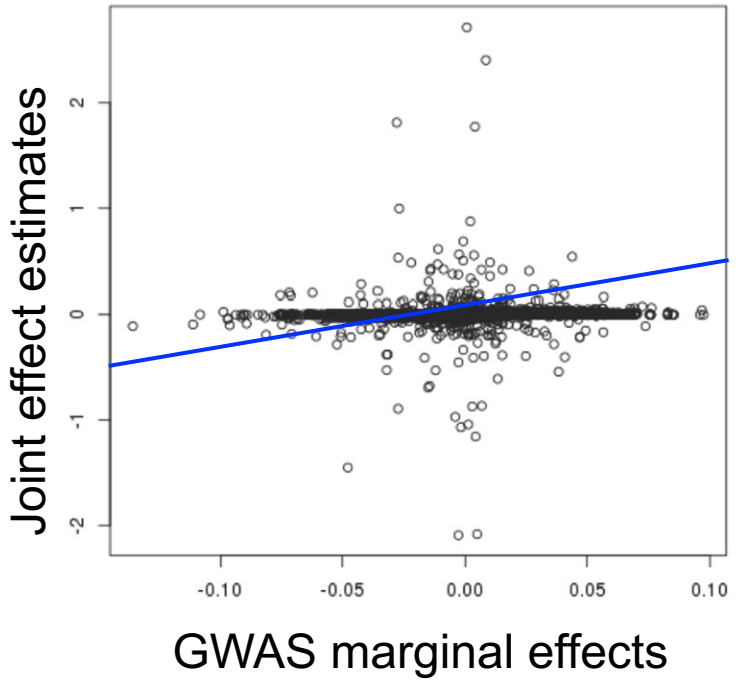
Most common



Presence of large effects



Bad convergence!

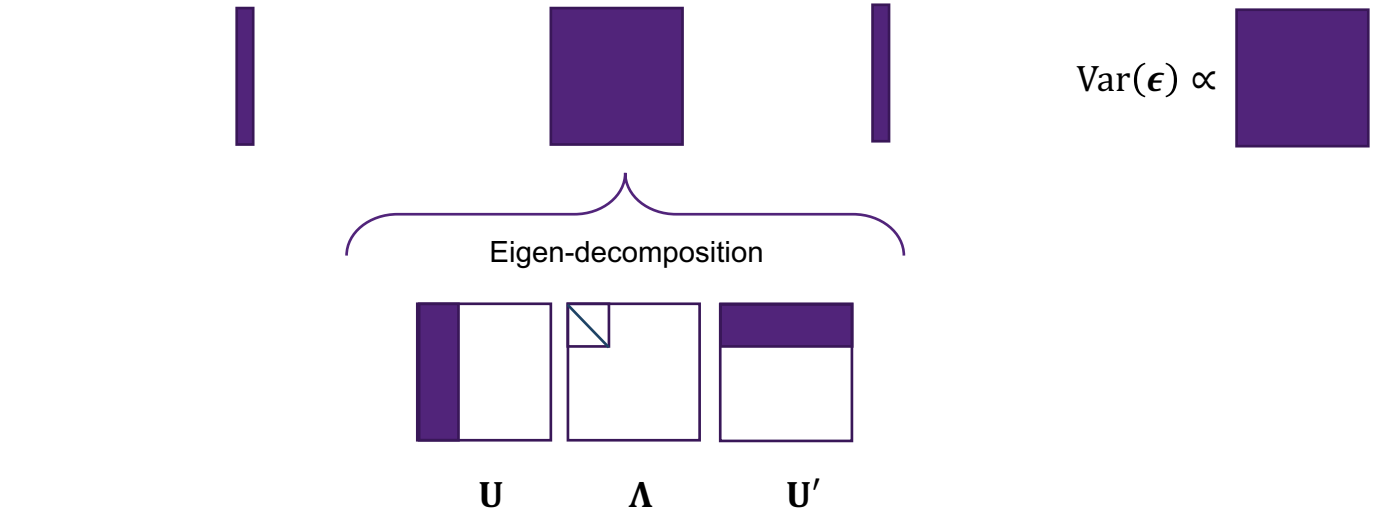


Low-rank model (fits 10M SNPs or more)

In each quasi-independent LD block:

$$\mathbf{b} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

GWAS SNP marginal effects | LD correlation matrix | SNP joint effects | Residuals



$$\mathbf{b} \sim \mathrm{N}(\mathbf{R}\boldsymbol{\beta}, \frac{1}{n}\mathbf{R}\sigma_e^2)$$

It only requires the top 20% PCs to explain 99.5% of the variance in LD!

$$\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{b} = \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}'\boldsymbol{\beta} + \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\boldsymbol{\epsilon}$$

$$\mathbf{w} = \mathbf{Q}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Improved computational efficiency and robustness

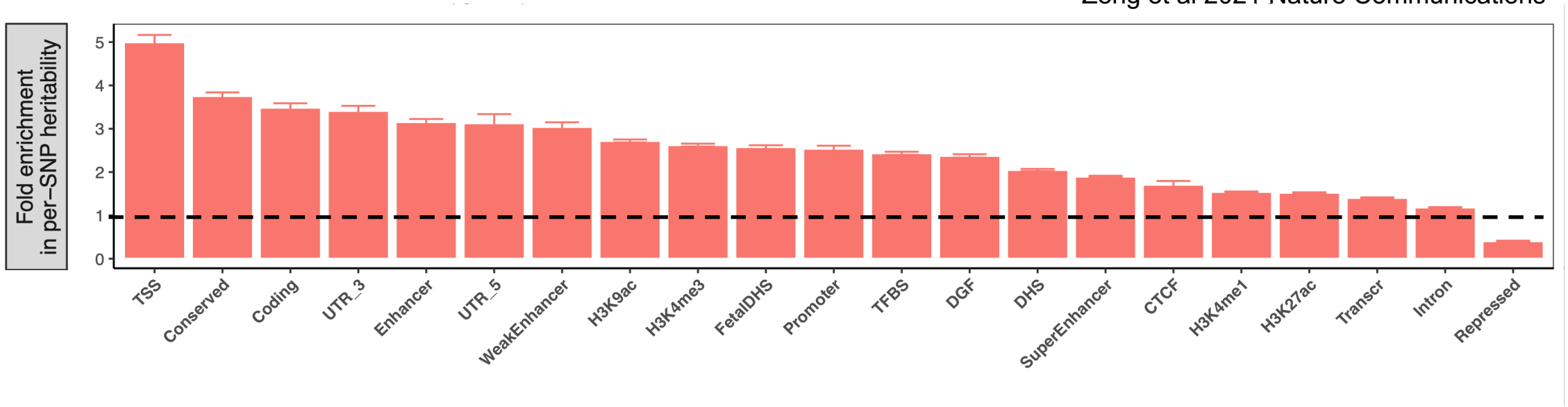
$$\mathbf{w} \sim \mathrm{N}(\mathbf{Q}\boldsymbol{\beta}, \frac{1}{n}\mathbf{I}\sigma_e^2)$$

Functional genomic annotations

Functional genomic annotations provide orthogonal information useful for polygenic prediction.

- Chromatin states
- Biological functions
- Molecular quantitative trait loci (xQTL)
- ……

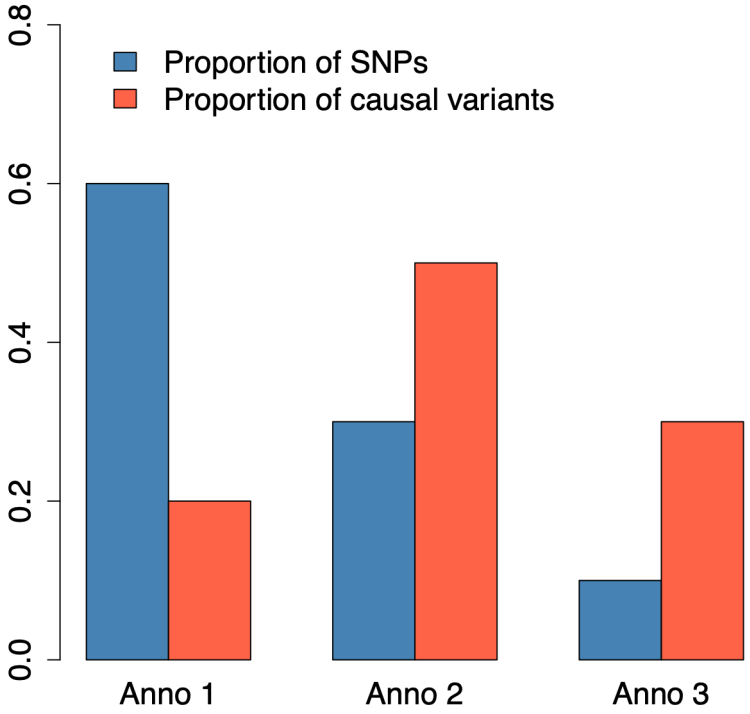
Zeng et al 2021 Nature Communications



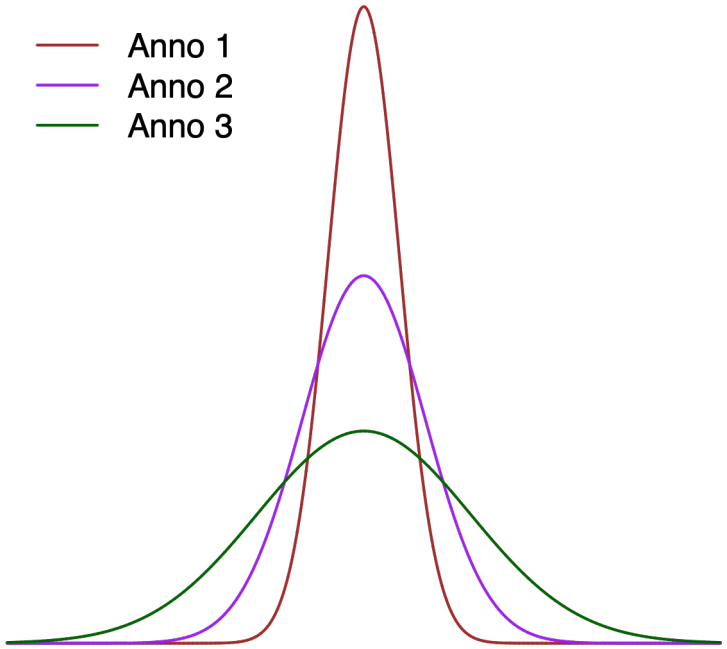
Opportunities/challenges

Functional annotations are informative on both the presence of causal variants and the distribution of causal effect sizes.

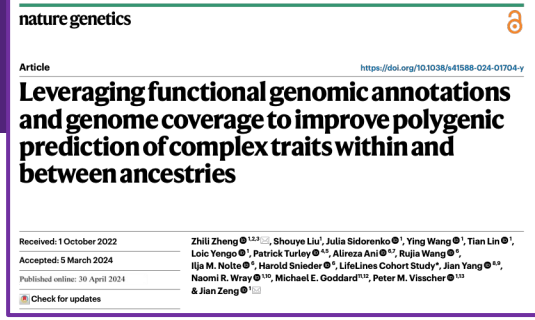
Differences in proportion of causal variants



Differences in distribution of causal effects



SBayesRC


nature genetics

Article

https://doi.org/10.1038/s41588-024-01704-y

Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries

Received: 1 October 2022

Accepted: 5 March 2024

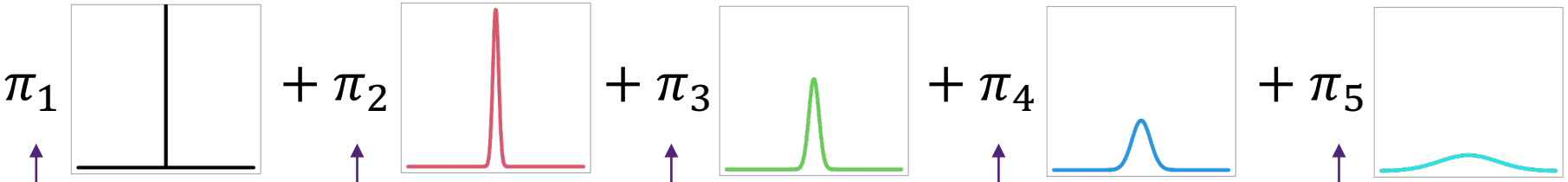
Published online: 30 April 2024

Check for updates

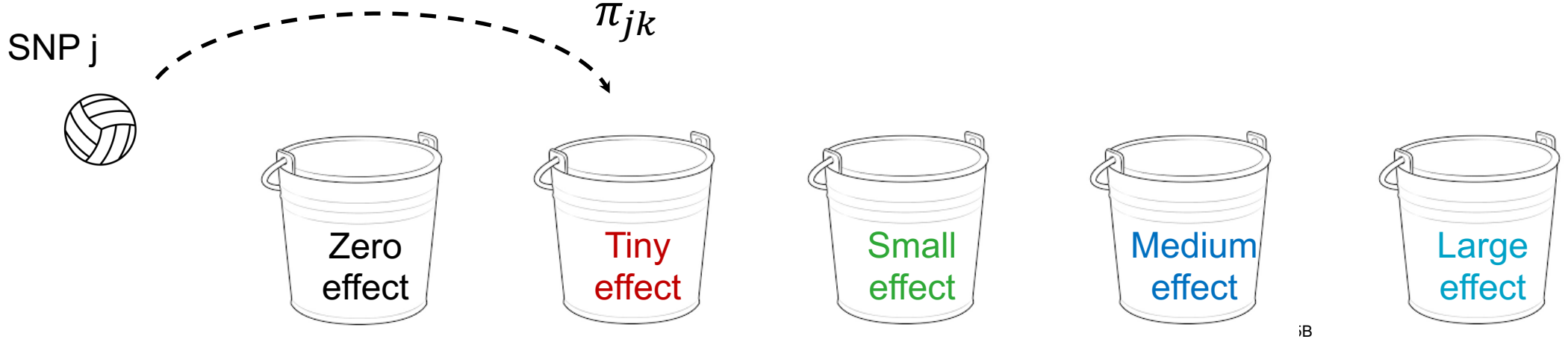
Zhili Zheng, Shouye Liu, Julia Sidorenko, Ying Wang, Tian Lin, Loic Yengo, Patrick Turley, Alireza Ani, Rujia Wang, Ilja M. Nolte, Harold Snieder, LifeLines Cohort Study, Jian Yang, Naomi R. Wray, Michael E. Goddard, Peter M. Visscher & Jian Zeng

Incorporate functional annotations through a hierarchical prior:

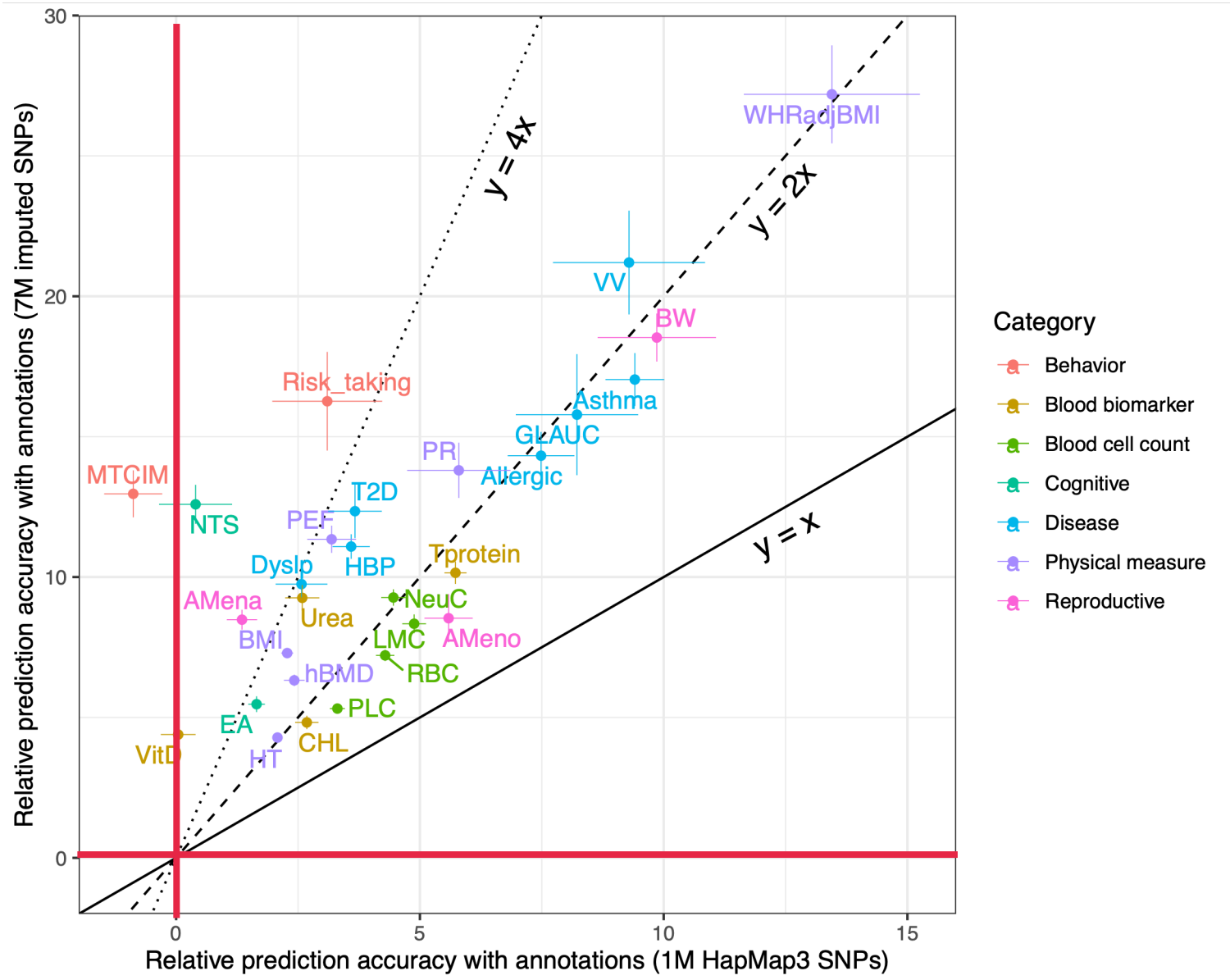
$$\beta_j \sim \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5$$



$$probit(\pi_{jk}) = \text{SNP annotations} \times \text{annotation effects}$$



Improved prediction within European ancestry



Improvement (%) in prediction accuracy with vs. without annotations:

$$\frac{R^2_{\text{annot}} - R^2_{\text{wo}}}{R^2_{\text{wo}}}$$

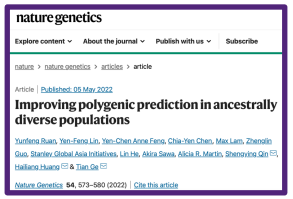
using 7M imputed SNPs (y-axis) or 1M HapMap3 SNPs (x-axis).

Annotations matter more with more SNPs - why?

SNP markers can tag the causal variant by LD but may not tag by annotation.

Trans-ancestry prediction

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


nature genetics

Explore content · About the journal · Publish with us · Subscribe

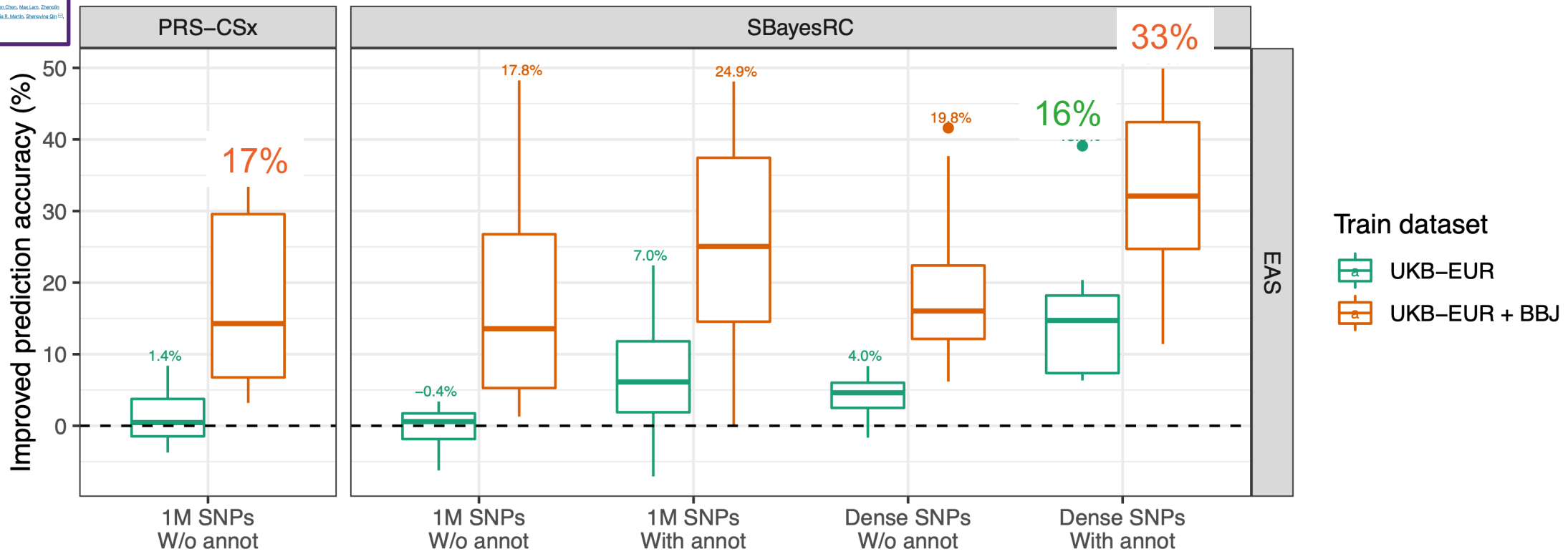
nature › nature genetics › articles › article

Article | Published: 05 May 2022

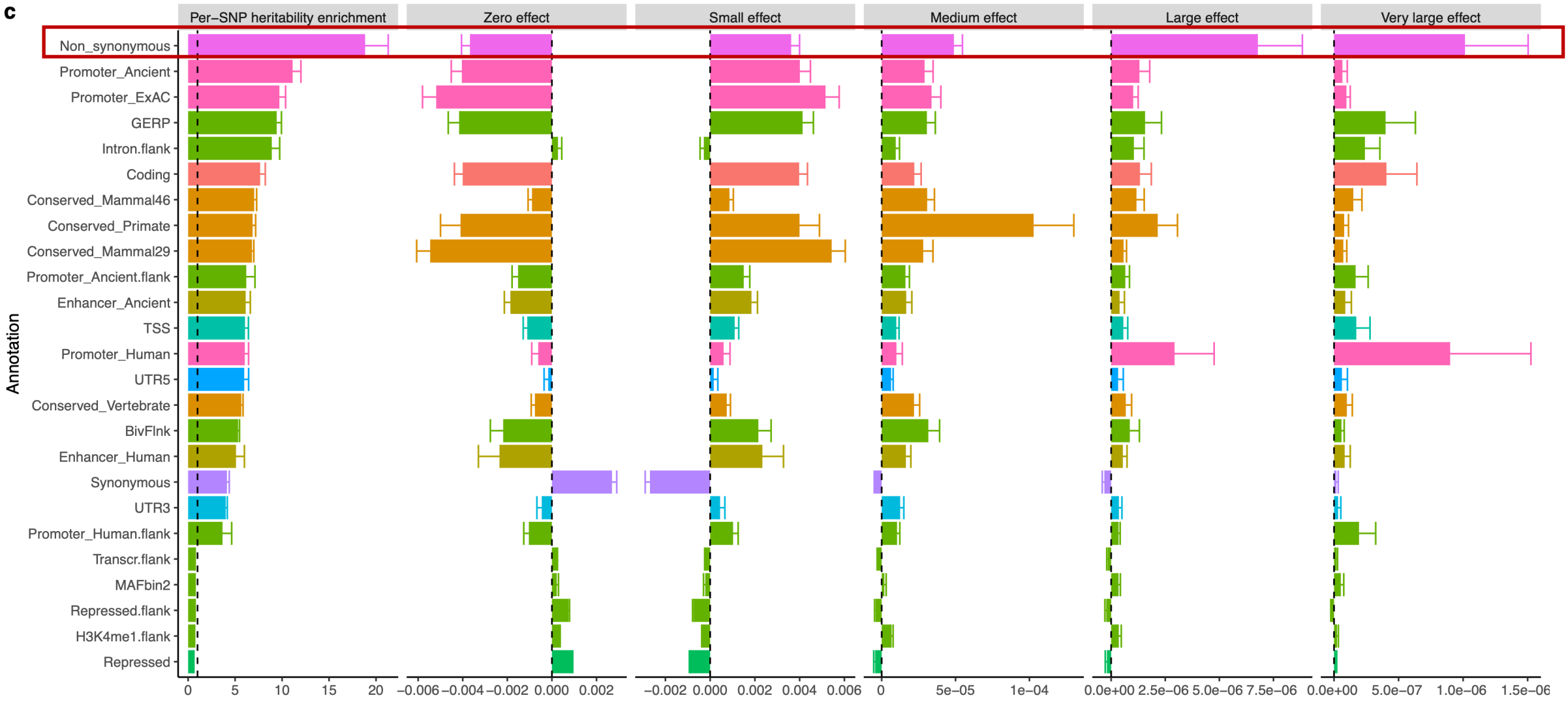
Improving polygenic prediction in ancestrally diverse populations

Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Stanley Global Asia Initiatives, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin, Hailiang Huang & Tian Ge

Nature Genetics 54, 573–580 (2022) | Cite this article



Functional genetic architecture



Summary

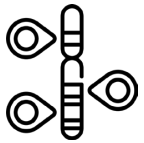
- Polygenic scores (PGS) are imperfect genetic predictors with inherently limited accuracy.
- Basic method (C+PT, aka P+T) for PGS prediction requires SNP selection and tuning.
- Bayesian methods simultaneously estimate all SNP effects and incorporate prior knowledge in estimation of SNP effects.
- State-of-the-art Bayesian methods utilize GWAS summary statistics, which unleash the power of large GWAS sample size, and functional annotations, which provide orthogonal information to GWAS data to better estimate SNP effects.
- PGS methods can also be used for understanding functional architecture and for genetic fine-mapping.

Genetics & Genomics Winter School

July 6 - 10, 2026 | Brisbane, Australia

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Statistical and Computational Methods



Statistical Genomics 1 Genetic Mapping

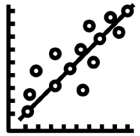
Dr Kathryn Kemper

- Genome-wide association study (GWAS)
- Data processing & quality control
- Resources & meta-analysis

Statistical Genomics 2 Heritability Estimation

Prof Loic Yengo

- Concepts, methods & implications
- Estimation using GWAS data
- Genomic REML
- Genomic partitioning analysis



Statistical Genomics 3 Polygenic Prediction

Dr Jian Zeng

- Polygenic risk score
- Utilities, opportunities & limitations
- Methodology & analytical pipeline
- Bayesian methods

Cellular Transcriptomics

A/Prof Quan Nguyen

- Single-cell & spatial transcriptomics
- Cell type analysis
- Machine learning for imaging and sequencing data

Genetic Epidemiology

Dr Daniel Hwang

- Causal inference using genetic data
- Mendelian randomization (MR)
- Structural equation modelling (SEM)

Systems Genomics & Pharmacogenomics

A/Prof Sonia Shah

- Transcriptome-wide QTL analysis
- Integrating GWAS with omics data
- Prediction of drug effects
- Connectivity Map for therapeutics

On-site lectures + hands-on practical exercises

Construct your own week-long course

6 modules offered, each 1.5 days, from 9am to 4pm

Each class size limited to 60 participants

Special Seminar, Social Events, HPC & Lab tour

Scholarships available for undergraduate students

Registration opens on 7th April