THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Bayesian Methods & Advanced Models for Polygenic Scores (PGS)

Jian Zeng

j.zeng@uq.edu.au

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | Institute for Molecular Bioscience

Program in Complex Trait Genomics

Outline

- Understanding to polygenic scores (PGS)
- Basic method to construct PGS
- Bayesian methods [15min Lecture | 15min Practical | 5min break]
- Using GWAS summary statistics [15min Lecture | 15min Practical | 5min break]
- Incorporating functional genomic annotations [15min Lecture | 15min Practical]
- Wrap-up/discussion/questions

Practical exercises will use a toy example and R scripts

All materials are available at:

https://gctbhub.cloud.edu.au/data/teaching/

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.

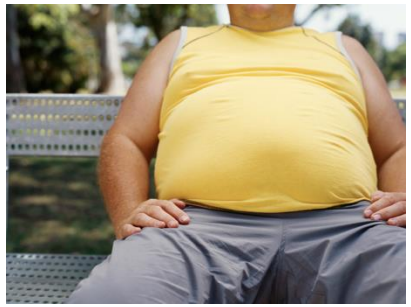
Head in the clouds Head in the sand

Source: Strachan & Read Human Molecular Genetics 3.

Theory and methodology of polygenic scores (PGS) are built on our understanding of "polygenicity" in complex traits.

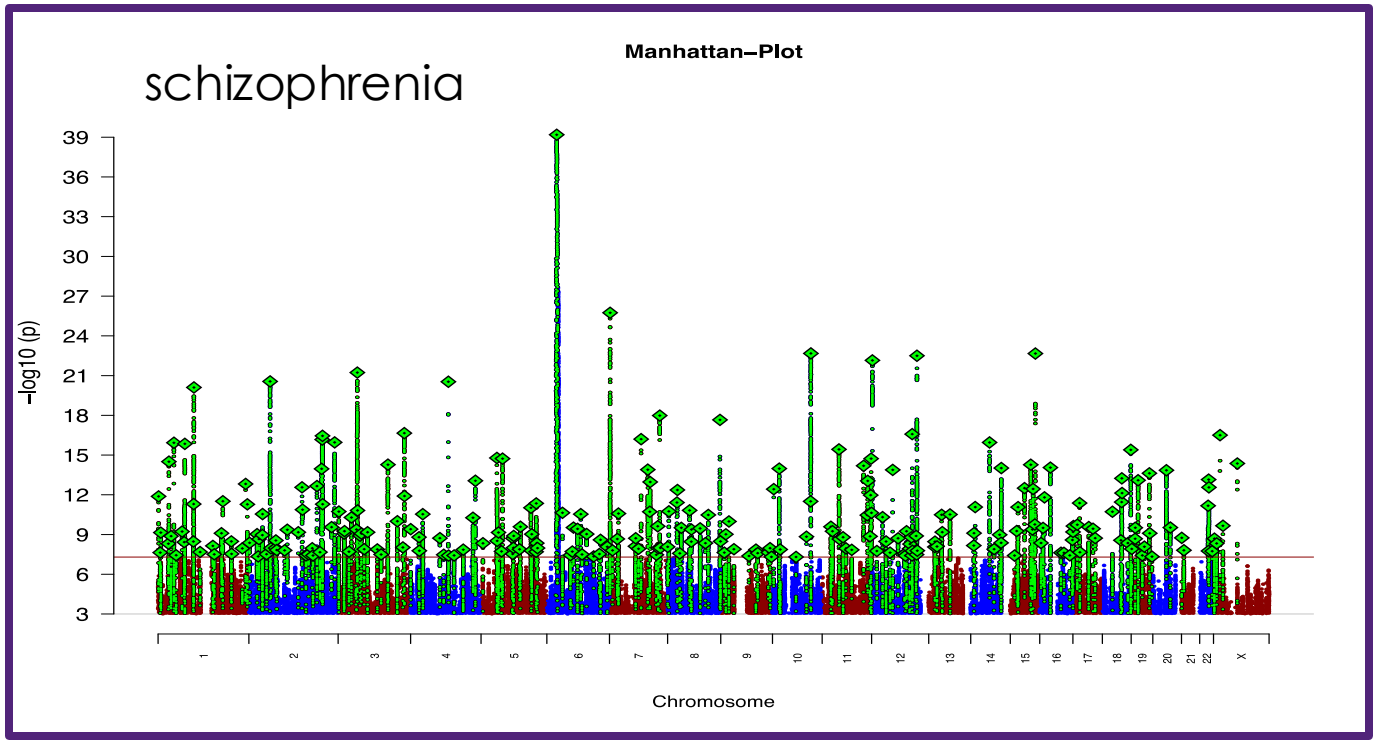
Height

Schizophrenia



Obesity

Common diseases are polygenic



nature

Explore content · About the journal · Publish with us

nature › articles › article

Article | Published: 08 April 2022

Mapping genomic loci implicates genes and synaptic biology in schizophrenia

Vassily Trubetskoy, Antonio F. Pardiñas, Ting Qi, Georgia Panagiotaropoulou, Swapnil Awasthi, Tim B. Bigdeli, Julien Bryois, Chia-Yen Chen, Charlotte A. Dennison, Lynsey S. Hall, Max Lam, Kyoko Watanabe, Oleksandr Frei, Tian Ge, Janet C. Harwood, Frank Koopmans, Sigurdur Magnusson, Alexander L. Richards, Julia Sidorenko, Yang Wu, Jian Zeng, Jakob Grove, Minsoo Kim, Zhiqiang Li, Indonesia Schizophrenia Consortium, PsychENCODE, Psychosis Endophenotypes International Consortium, The SynGO Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium

+ Show authors

Nature **604**, 502–508 (2022) | Cite this article

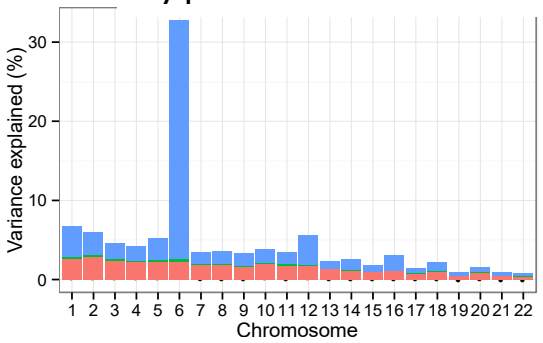
57k Accesses | **321** Citations | **463** Altmetric | Metrics

248 risk loci identified at genome-wide significance level.

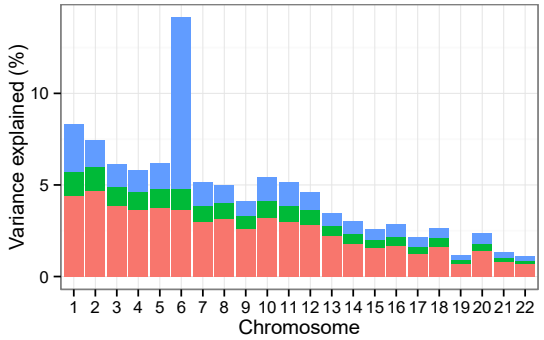
We predict thousands are associated with schizophrenia.

Many polygenic genetic architectures

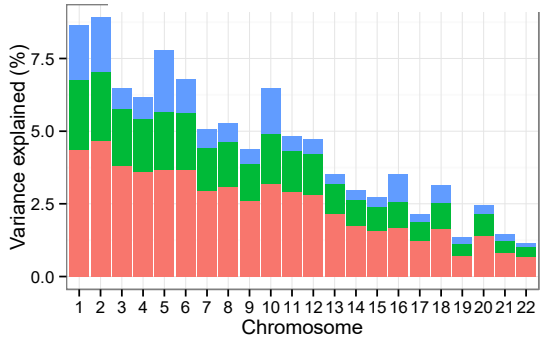
Type 1 Diabetes



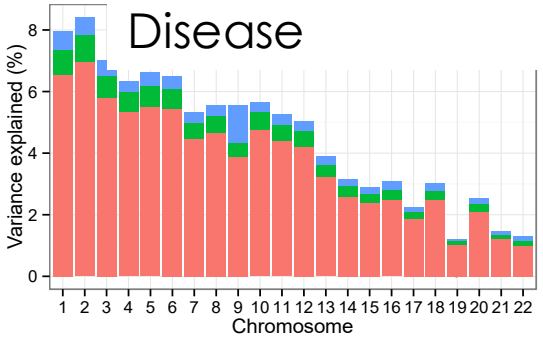
Rheumatoid Arthritis



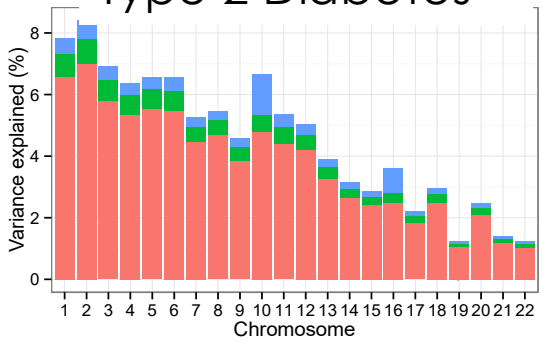
Crohn's Diseasse



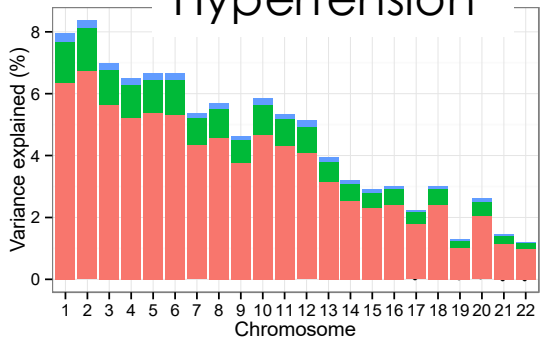
Coronary Artery Disease



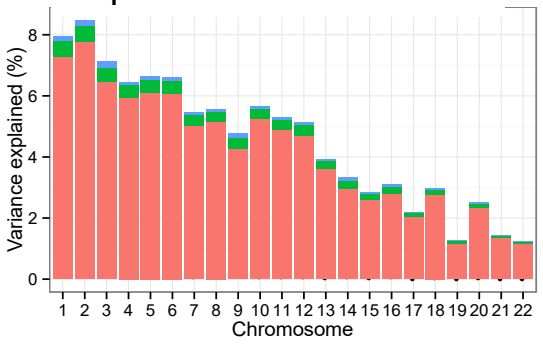
Type 2 Diabetes



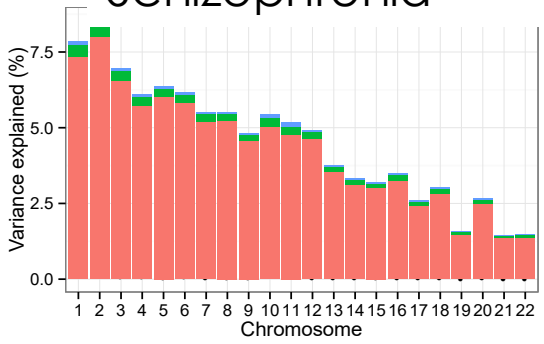
Hypertension



Bipolar Disorder



Schizophrenia

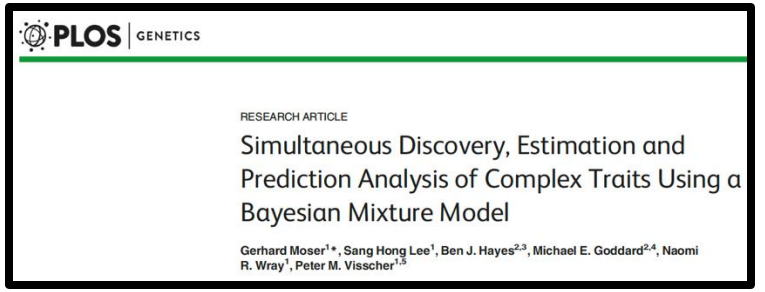


Mixture component

$10^{-4} \times \sigma^2_g$ small

$10^{-3} \times \sigma^2_g$ middle

$10^{-2} \times \sigma^2_g$ big



PLOS | GENETICS

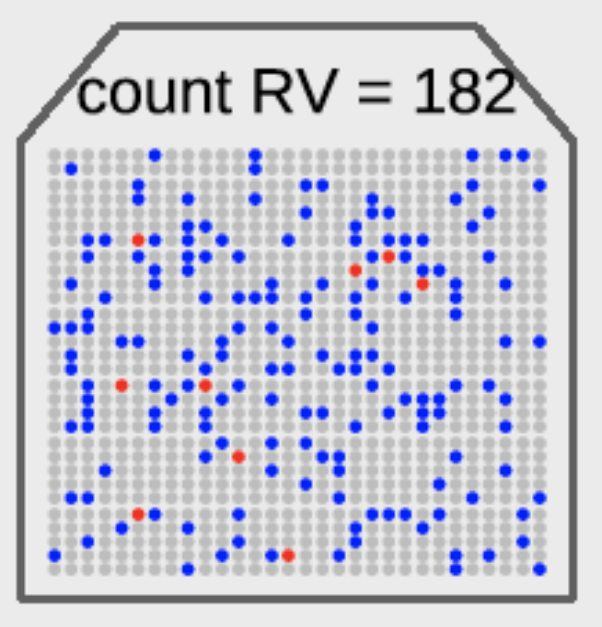
RESEARCH ARTICLE

Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model

Gerhard Moser[1]*, Sang Hong Lee[1], Ben J. Hayes[2,3], Michael E. Goddard[2,4], Naomi R. Wray[1], Peter M. Visscher[1,5]

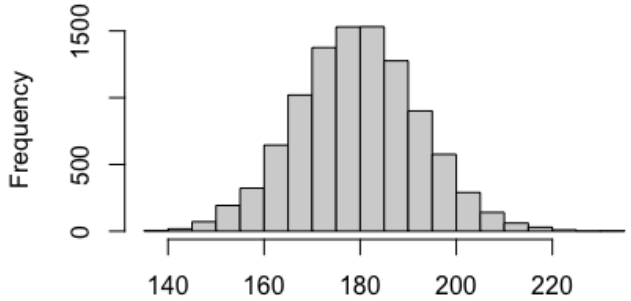
Many DNA variants contribute to genetic risk, and most have very small effects.

Polygenic disease for an individual



900 DNA polymorphic sites

RV =risk variant



Count of RV in population

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

Average person 900*2*0.1 = 180 risk variant

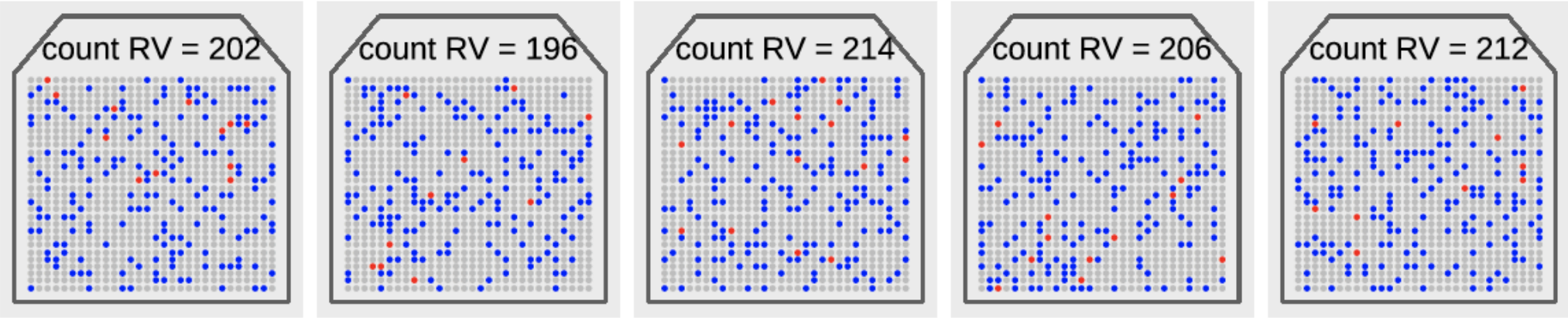
Mean +/- 3SD: 142 to 218

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

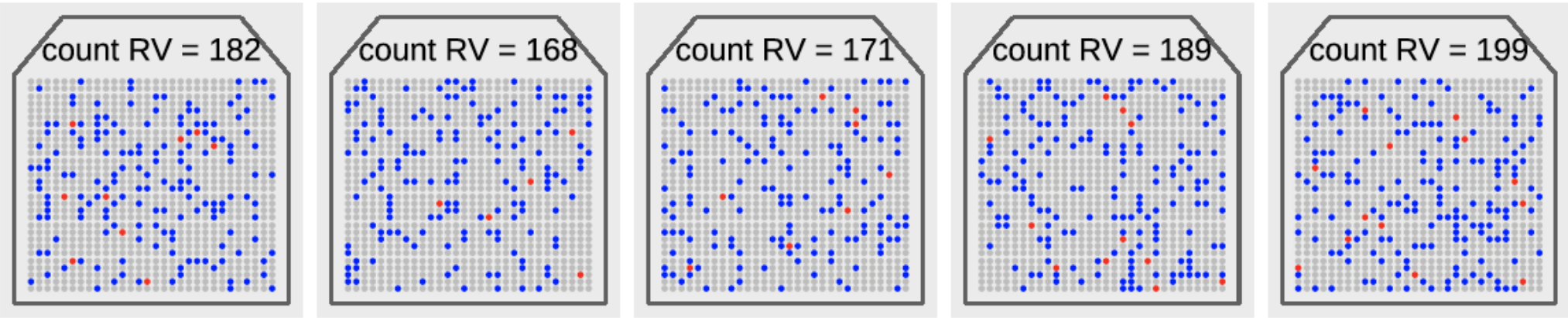
Toy example

Polygenic disease for an individual

Affected over lifetime

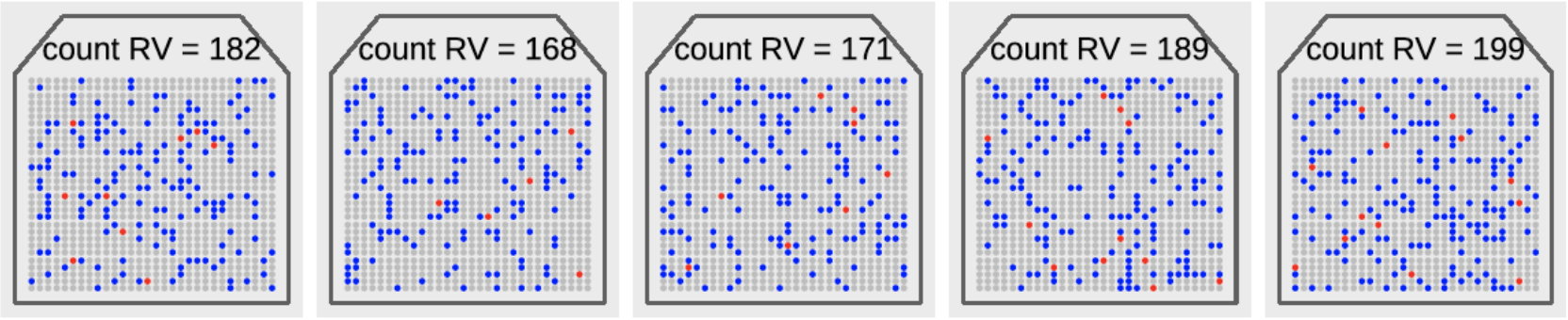


Not affected over lifetime



- 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

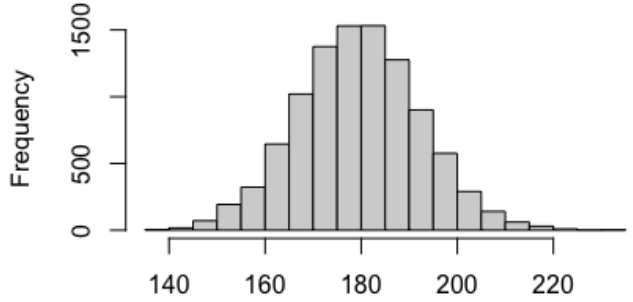
Polygenic score



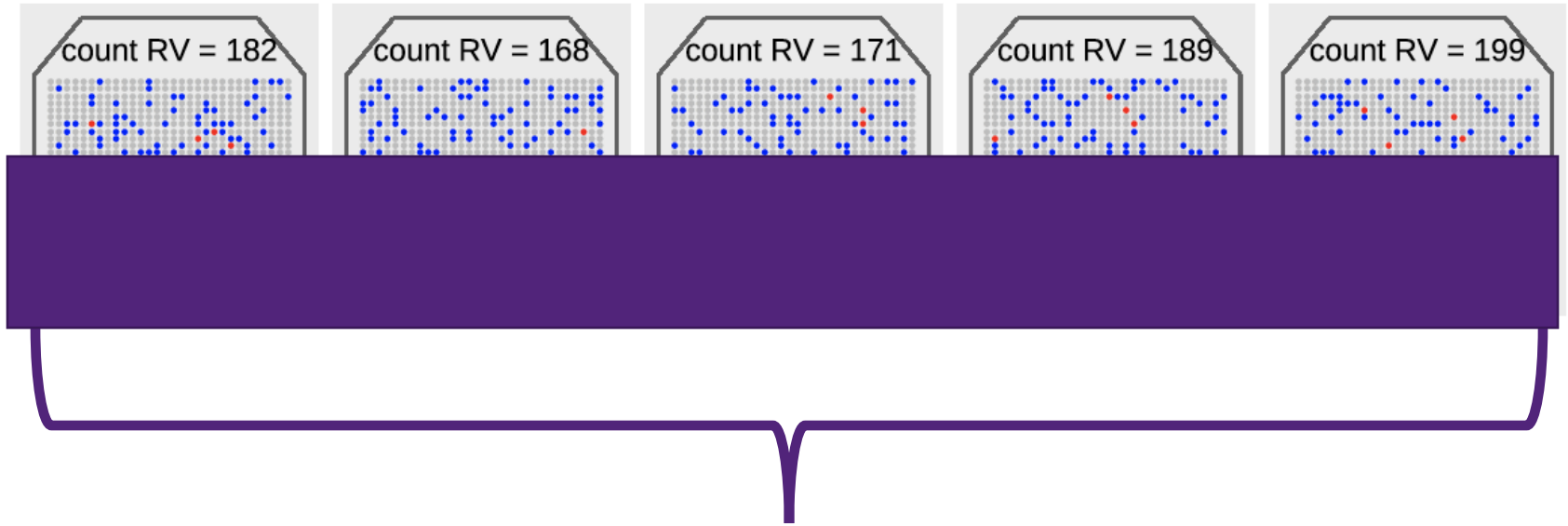
→ "True" polygenic score

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

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

$\mathrm{SNP-based\ heritability}$

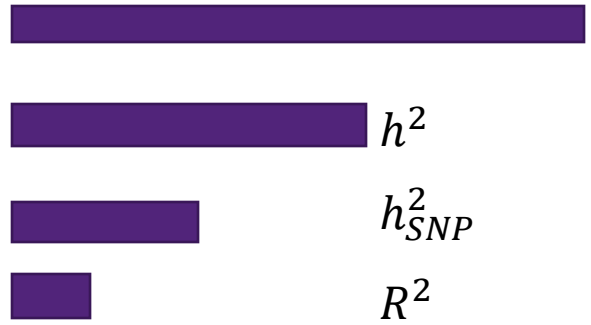
Limitations in prediction accuracy

- 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

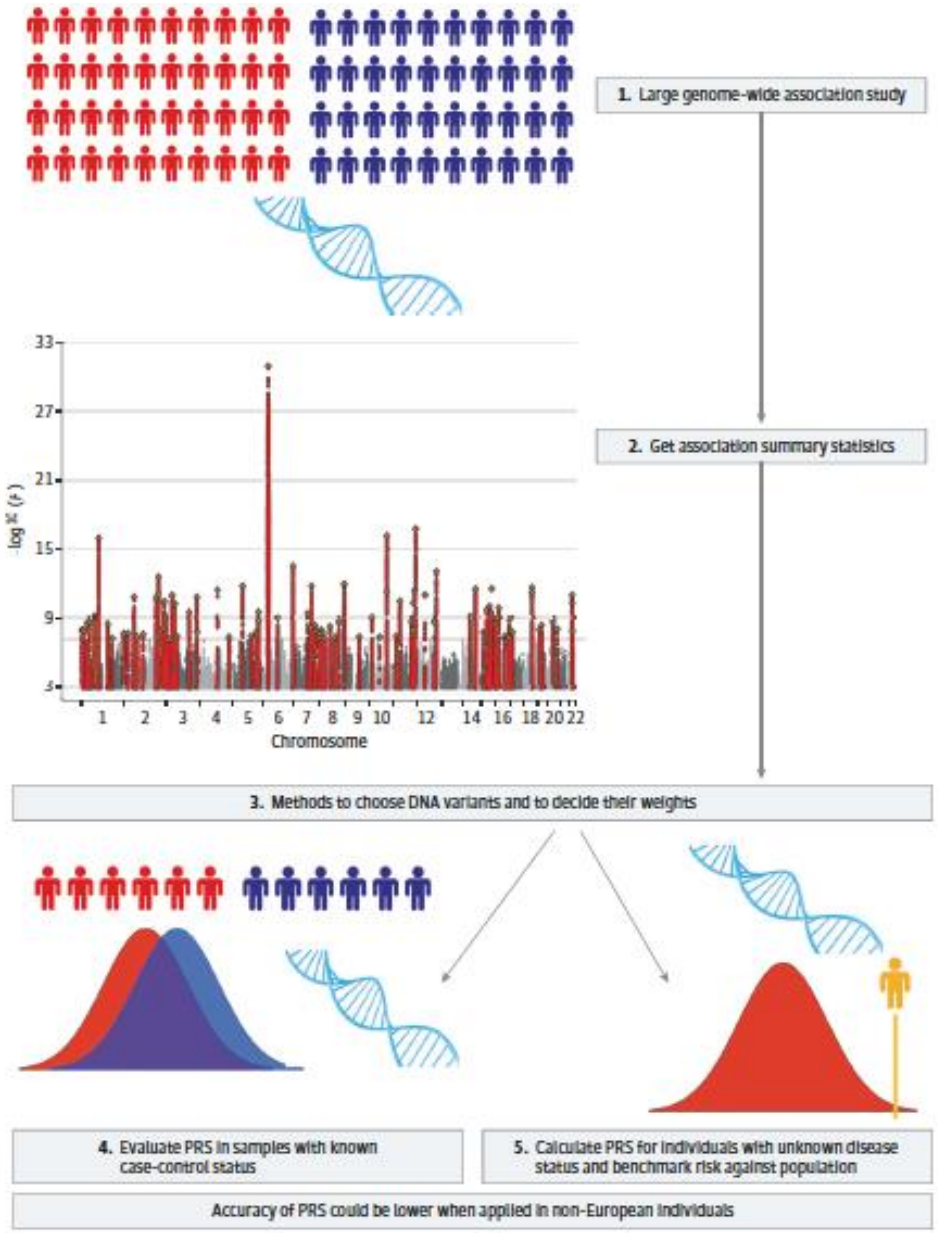
Max:
25% Liability
AUC 0.84

Current:
11% Liability
AUC 0.74

Polygenic scores cannot be highly accurate predictors of phenotypes

Basic method for PGS

Polygenic scores (PGS)



- A weighted count of risk alleles

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

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

PGS methods

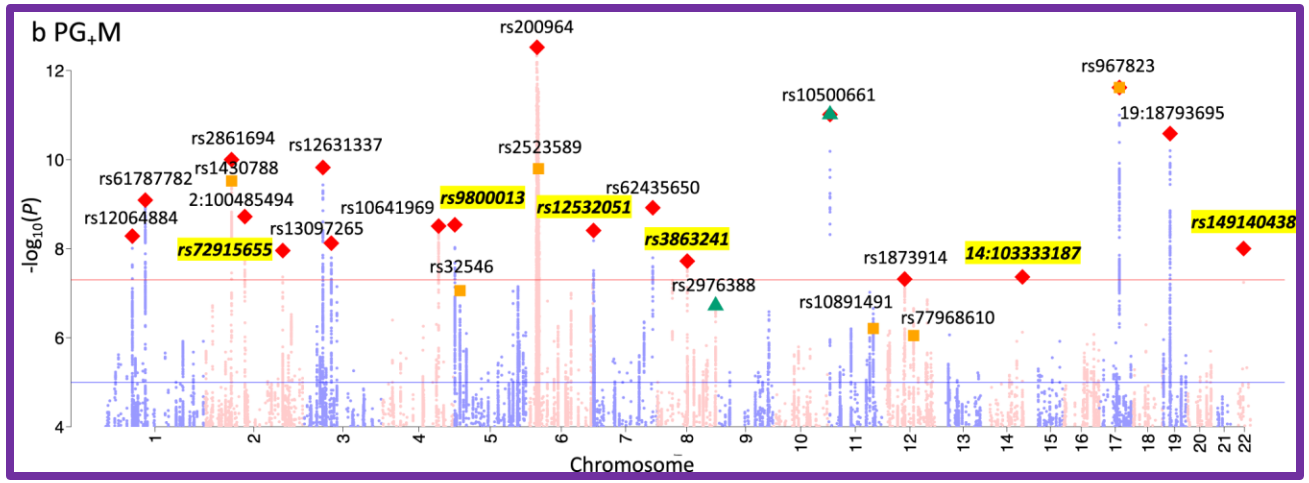
A weighted sum of the count of risk alleles

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

How many SNPs?
Which SNPs?
What weights?

Basic method:

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



PGS methods

A weighted sum of the count of risk alleles

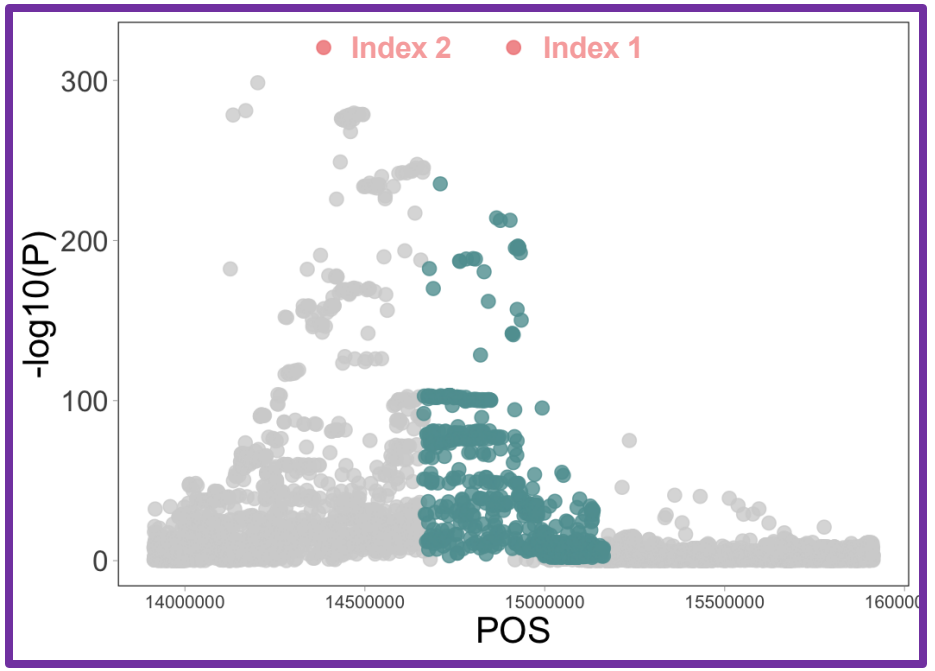
$$\text{PGS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

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

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



PGS methods

A weighted sum of the count of risk alleles

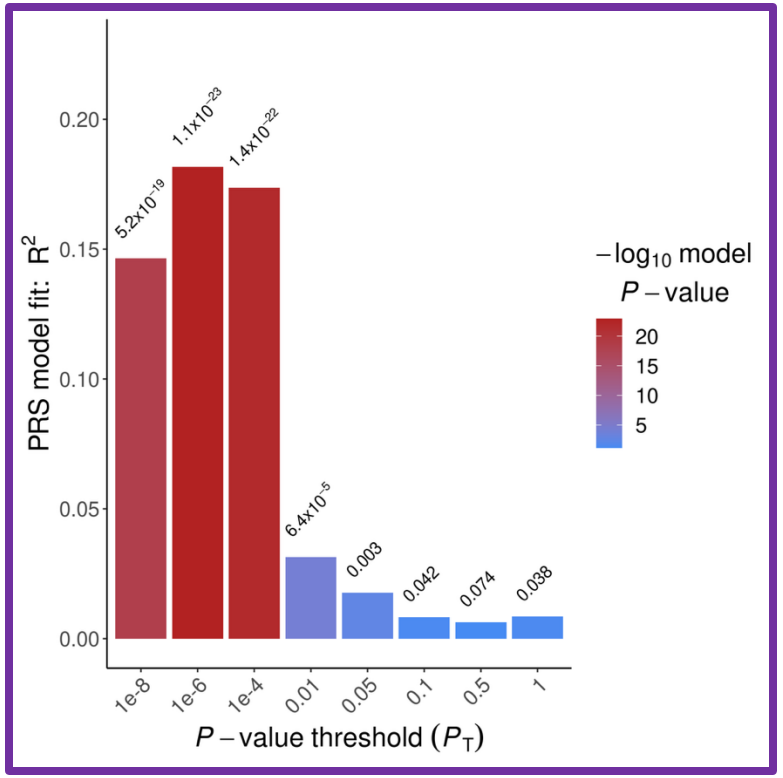
$$\text{PGS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

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

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



PGS methods

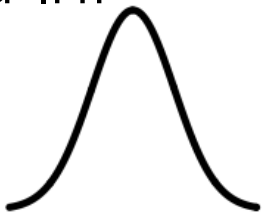
A weighted sum of the count of risk alleles

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

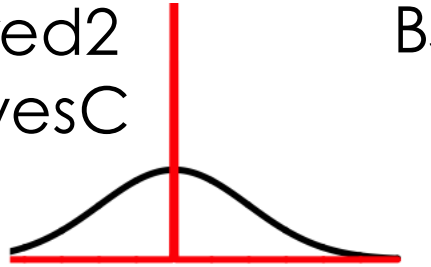
How many SNPs?
Which SNPs?
What weights?

New methods model genetic architecture

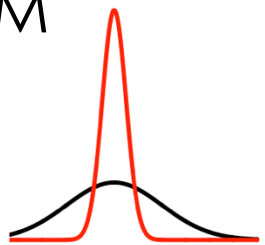
LDpred-Inf
SBLUP



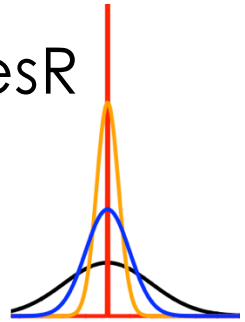
LDPred2
SBayesC



BSLMM



SBayesR



17

PGS methods

Table 1. Summary of Methods Used to Generate Polygenic Scores

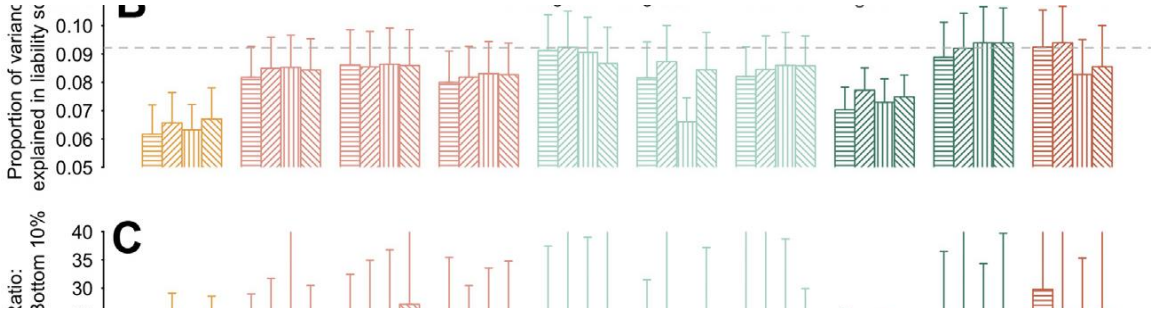
Method	Distribution of SNP Effects (β)	Tuning Sample	Predefined Parameters	Parameters Estimated in Tuning Sample
PC+T	None	Yes	–	*p*-value threshold
SBLUP	$\beta \sim N\left(0, \frac{h_g^2}{m}\right)$ h_g^2: SNP-based heritability, m: number of SNPs; $\lambda = m(1 - h_g^2)/h_g^2$	No	λ LD radius in kb	–
Ldpred2-Inf	Same as SBLUP	No	h_g^2 LD radius in cM or kb	–
LDpred-funct	$\beta_j \sim N(0, c\sigma_j^2)$ $\sum_{j=1}^{M} 1_{\sigma_j^2>0} c\sigma_j^2 = h_g^2$, c is a normalizing constant, σ_j^2 is the expected per SNP heritability under the baseline-LD annotation model estimated by stratified LDSC from the discovery GWAS within LDpred-funct software	No	h_g^2 LD radius in number of SNPs	–
LDpred2	$\beta_j \sim \begin{cases} N\left(0, \frac{h_g^2}{\pi m}\right), & \text{with probability of } \pi \\ 0, & \text{with probability of } 1-\pi \end{cases}$ When sparsity is "true," the β_j for SNPs in the $(1 - \pi)$ partition are all set to zero	Yes	h_g^2 π software default values, LD radius in cM or kb	π, sparsity
Lassosum	$f(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + (1-s)\boldsymbol{\beta}^T\mathbf{X}_r^T\mathbf{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + s\boldsymbol{\beta}^T\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1^1$ $\mathbf{X}_r$: $n \times m$ matrix of genotypes of LD reference sample, where n is sample size	Yes	LD blocks	λ, s
PRS-CS	$\beta_j \sim N\left(0, \frac{\sigma^2}{n}\psi_j\right)$ $\psi_j \sim G(a, \delta_j)$ $\delta_j \sim G(b, \phi)$, ϕ is a global scaling parameter	Yes	$a = 1$, $b = 0.5$ n LD blocks	ϕ
PRS-CS-auto	Same as PRS-CS, but estimates ϕ from the discovery GWAS	No	$a = 1$, $b = 0.5$ n LD blocks	–
SBayesR	$\beta_j \mid \boldsymbol{\pi}, \sigma_\beta^2 \sim \begin{cases} 0, & \text{with probability of } \pi_1 \\ N(0, \gamma_2\sigma_\beta^2), & \text{with probability of } \pi_2 \\ \vdots \\ N(0, \gamma_C\sigma_\beta^2), & \text{with probability of } 1 - \sum_{c=1}^{C-1}\pi_c \end{cases}$ $\sigma_\beta^2 \sim Inv - \chi^2(d.f. = 4)$ $\pi_i \sim Dir(\mathbf{1})$, estimated from discovery GWAS in SBayesR software γ_i are scaling parameters	No	LD radius in cM or kb C = 4 γ software default values	–
MegaPRS	Lasso: $\beta_j \sim DE(\lambda/\sigma_j)$ Ridge regression: $\beta_j \sim N(0, v\sigma_j^2)$ BOLT-LMM: $\beta_j \sim \begin{cases} N\left(0, \frac{(1-f_2)\sigma_j^2}{\pi}\right), & \text{with probability of } \pi \\ N\left(0, \frac{f_2\sigma_j^2}{1-\pi}\right), & \text{with probability of } 1-\pi \end{cases}$ f_2 is the proportion of the total mixture variance in the second normal distribution BayesR: similar to SBayesR with C = 4, and π_i and γ_i estimated in the tuning sample σ_j^2 is the expected per SNP-heritability under BLD-LDAK model using SumHer	Yes	LD radius in cM or kb Parameters used in BLD-LDAK Grid search parameter values for each method	The tuning cohort is used to estimate the parameters that maximize prediction for each model, and from these the model that maximizes prediction is selected

Archival Report

Biological Psychiatry

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

Guiyan Ni, Jian Zeng, Joana A. Revez, Ying Wang, Zhili Zheng, Tian Ge, Restuadi Restuadi, Jacqueline Kiewa, Dale R. Nyholt, Jonathan R.I. Coleman, Jordan W. Smoller, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Jian Yang, Peter M. Visscher, and Naomi R. Wray



- Random effects models > fixed effects models
- Mixture models > non-mixture (infinitesimal) models

Bayesian methods for PGS

Bayesian methods

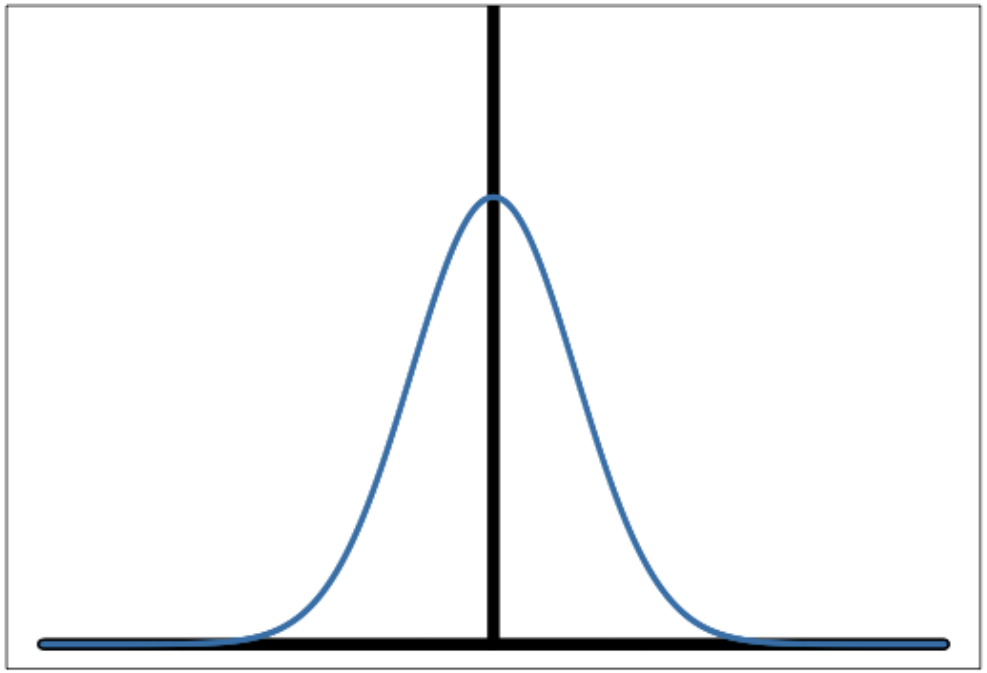
- Bayesian methods can estimate all parameters including SNP effects simultaneously
- Allow alternative assumptions regarding the distribution of SNP effects
- Different Bayesian methods mainly differ in
  - Assumption of the SNP effect distribution
  - Implementation

Assumptions for SNP effect distribution

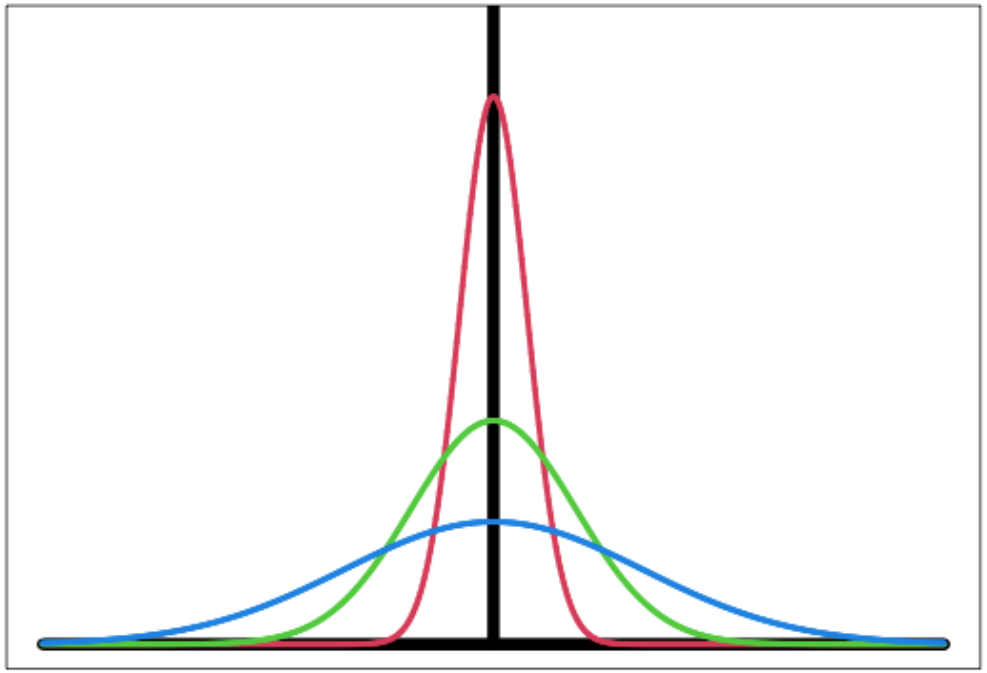
Assumption	Distribution of SNP effects	Method
Infinite number of small effects	Normal distribution	BLUP/LDpred2-inf
Small number of moderate to large effects, many small effects	Student's t distribution, hierarchical distribution	BayesA/PRS-CS
Small number of small effects, many zero effects	Mixture, spike at zero, normal distribution	LDpred2/BayesC
Many zero effects, proportion of small effects, some moderate to large effects	Mixture, multiple normals	BayesR

Assumptions for SNP effect distribution

BayesC



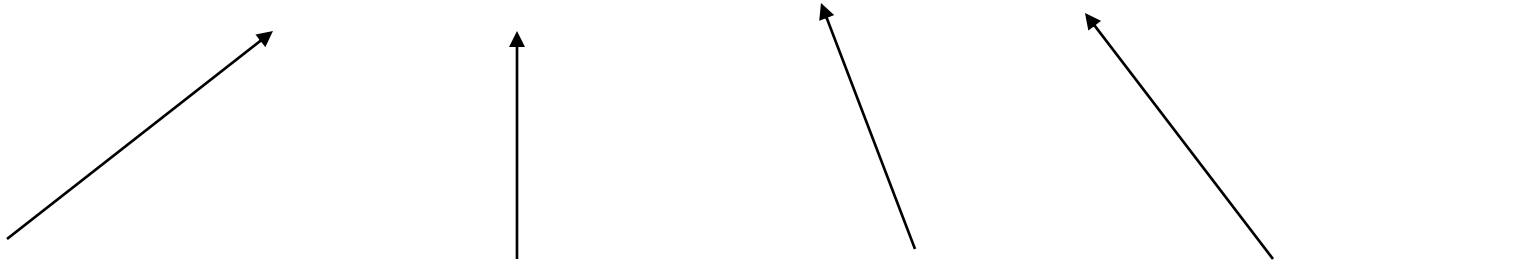
BayesR



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

Introduction to Bayesian methods

Consider an experiment where we measure height of 10 people **to estimate average height**

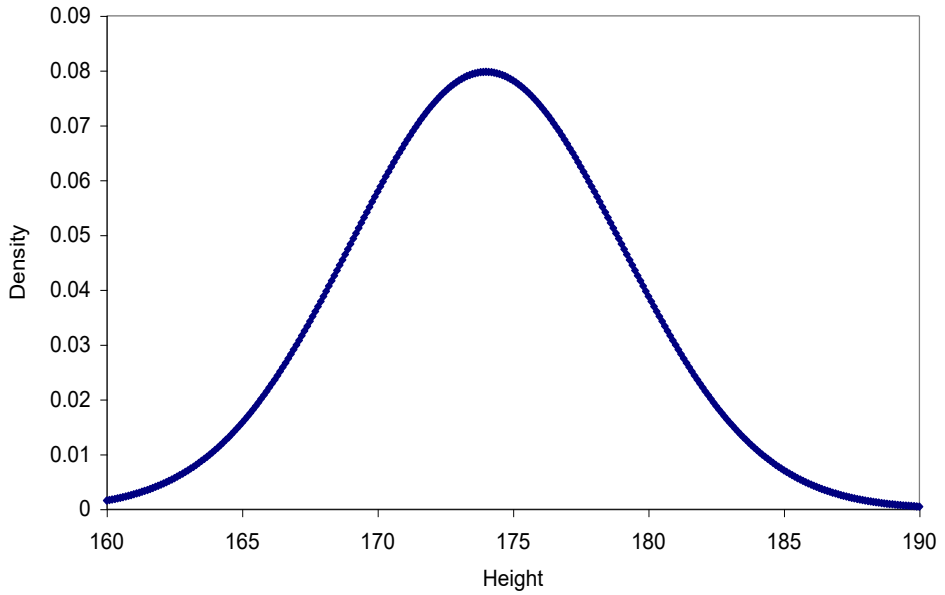
We want to use prior knowledge from many previous studies that average height is 174cm with standard error 5cm

$$y = \text{average height} + e$$

Bayes theorem

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

Prior probability of x (average height)



Bayes theorem

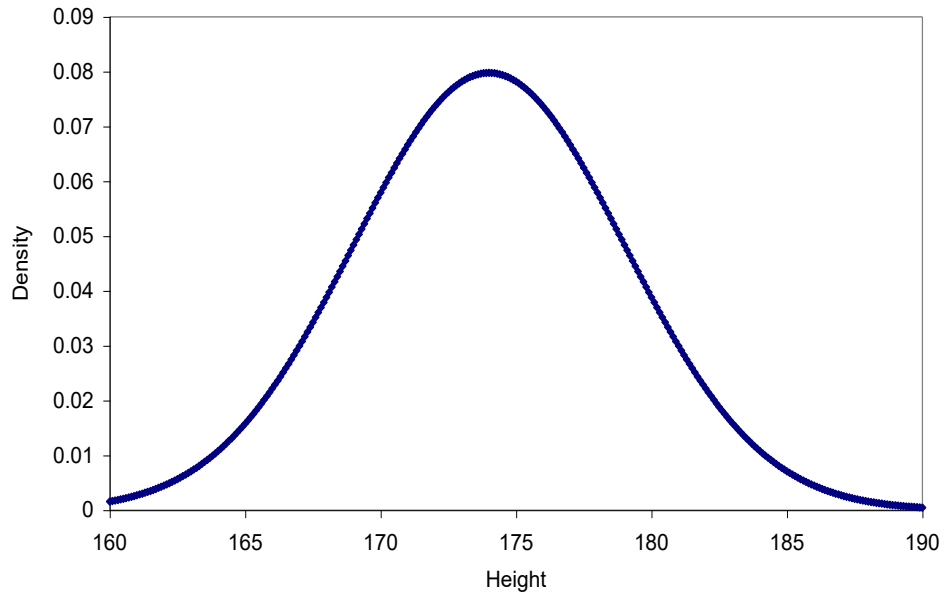
$$P(x \mid y) \propto P(y \mid x)P(x)$$

From the data……

$$\bar{x} = 178$$

$$s.e = 5$$

Prior probability of x (average height)

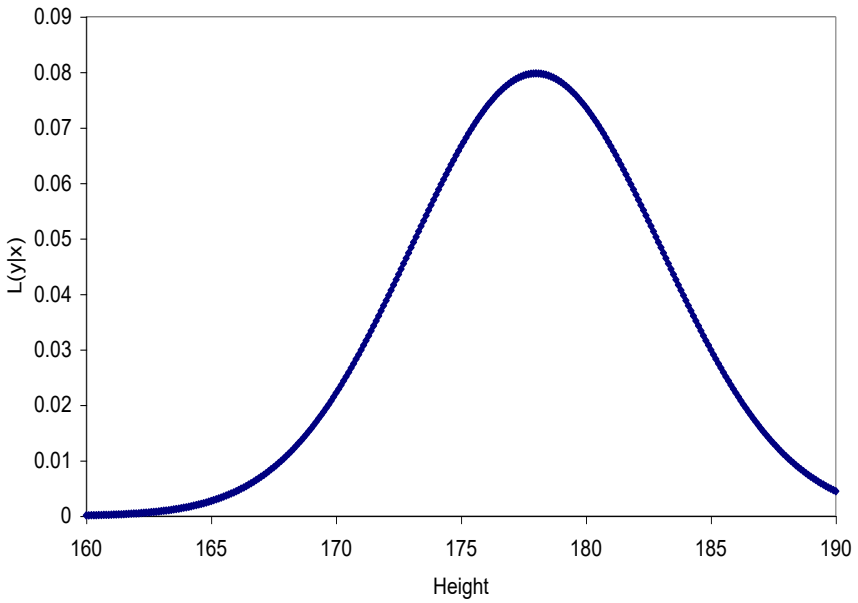


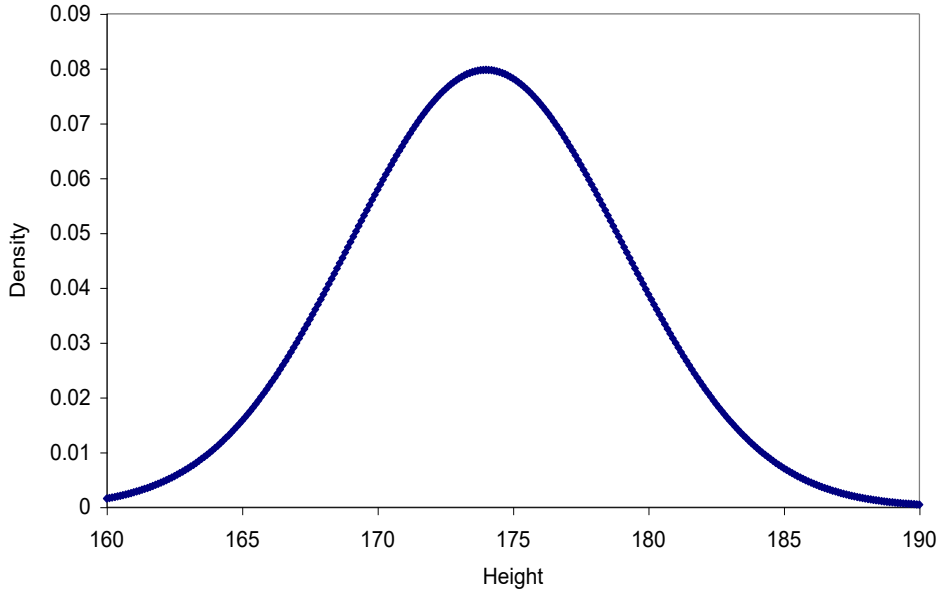
Bayes theorem

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

Likelihood of data (y) given height x, most likely x = 178cm

Prior probability of x (average height)





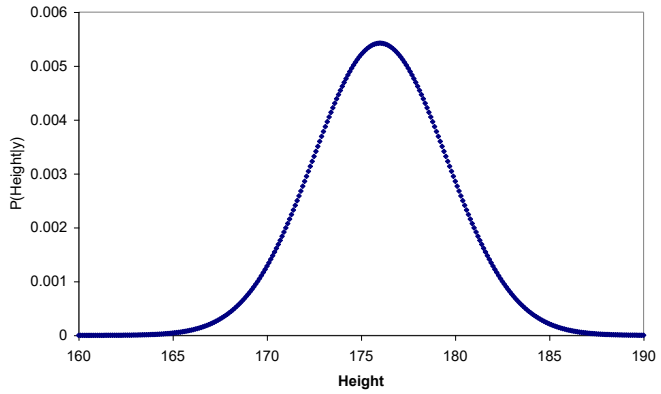
Bayes theorem

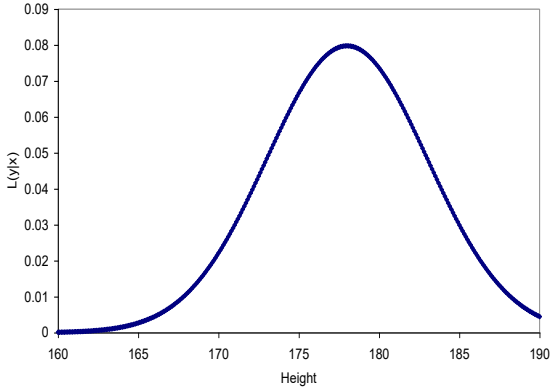
$$P(x \mid y) \propto P(y \mid x)P(x)$$

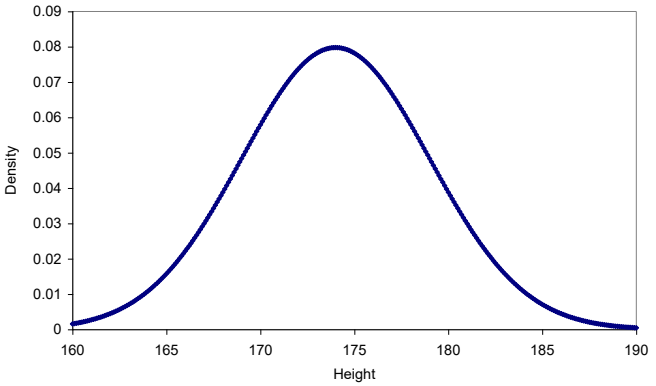
P(x|y) mean = 176cm

L(y|x)

P(x)







Introduction to Bayesian methods

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

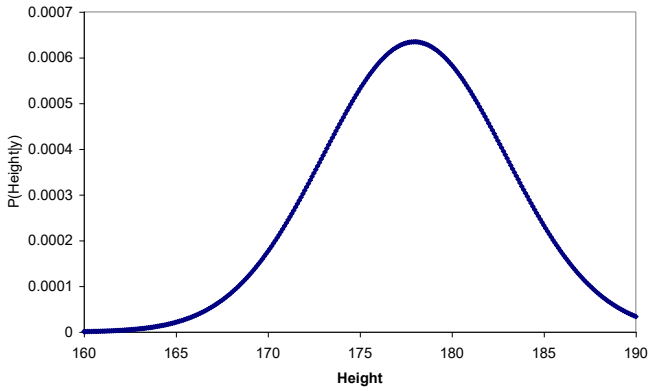
Less certainty about prior information? Use *less* informative (flat) prior

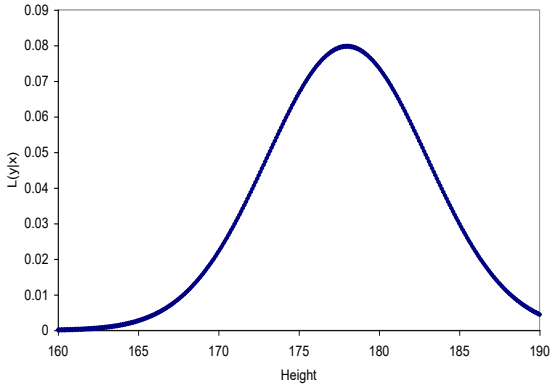
$$P(x \mid y) \propto P(y \mid x)P(x)$$

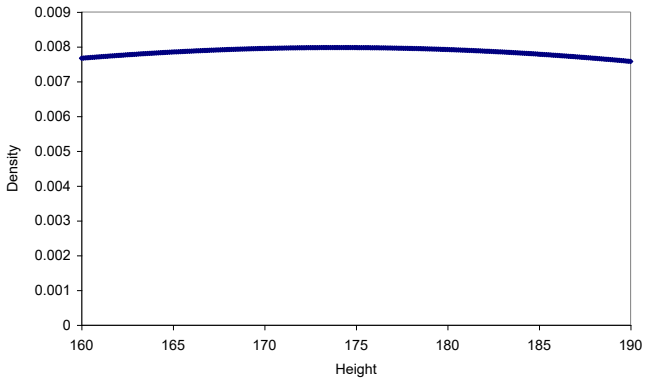
P(x|y) mean = 178cm

L(y|x)

P(x)







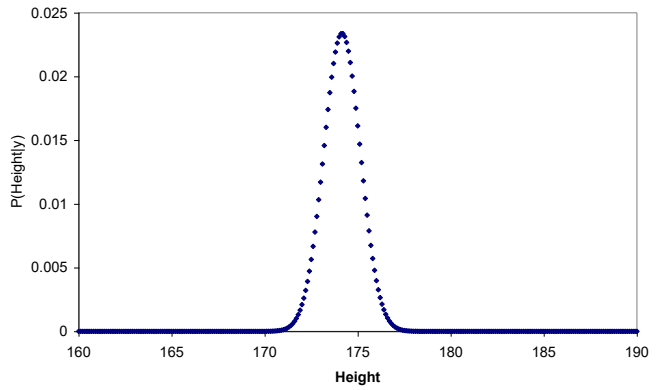
More certainty about prior information? Use *more* informative prior

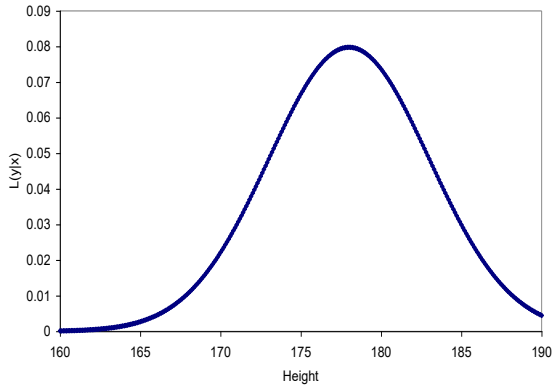
$$P(x \mid y) \propto P(y \mid x)P(x)$$

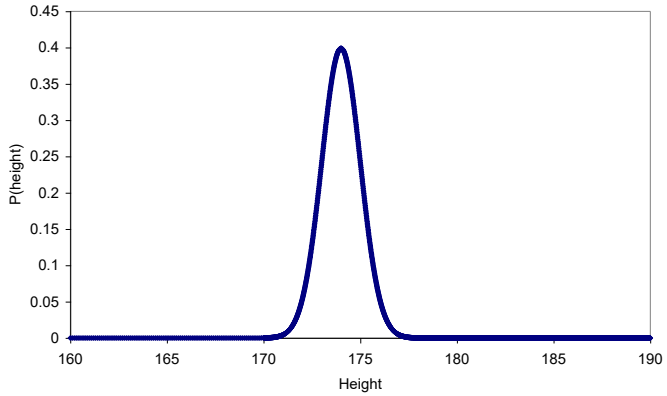
P(x|y) mean = 174.5cm

L(y|x)

P(x)



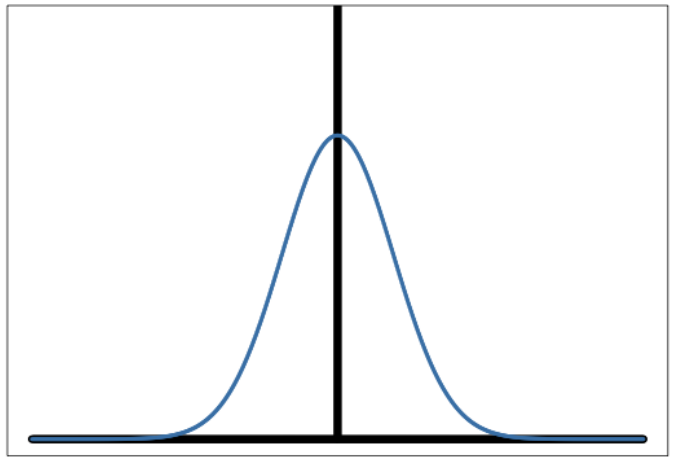




Model

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

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



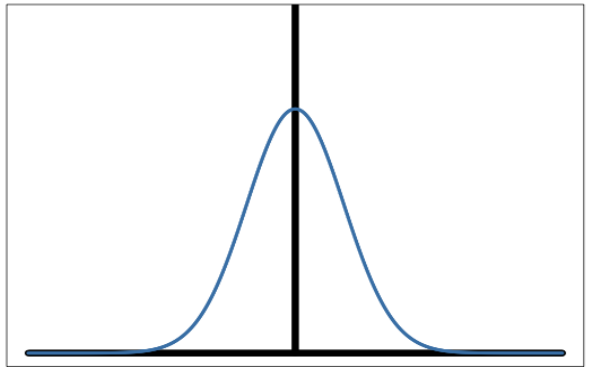
Posterior inference on SNP effects

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

$$P(\boldsymbol{\beta}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{\beta})P(\boldsymbol{\beta})$$

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

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



SNP effect estimates:

$$\widehat{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|\boldsymbol{y}) = \int \boldsymbol{\beta} P(\boldsymbol{\beta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\beta}$$

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

Posterior inference on SNP effects

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

- Cannot solve directly → no closed form solution
- Estimates of parameters depend on other parameters
- Use Markov chain Monte Carlo (MCMC) algorithm!

MCMC algorithm

Markov chain

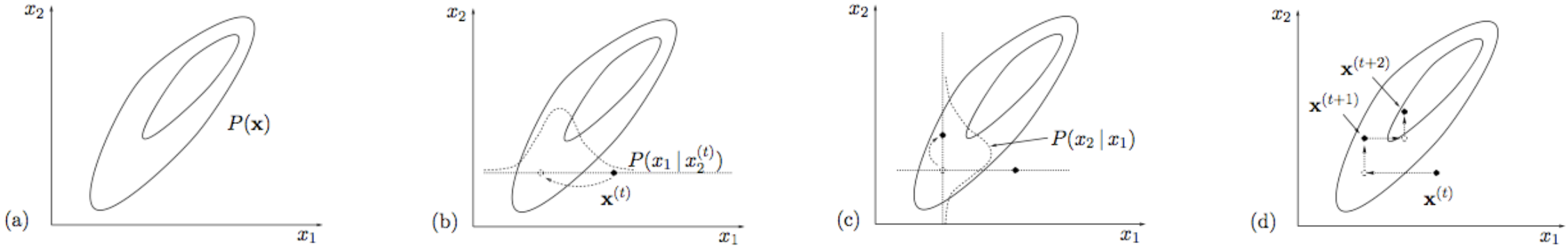
A sequence of samples where each sample depends only on the previous one (memoryless). This property allows the algorithm to gradually explore the distribution.

Monte Carlo

Using random sampling to perform numerical estimation, e.g., integrating over a probability distribution by averaging over samples.

Gibbs Sampling

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



The key is to derive $P(x_1|x_2)$ and $P(x_2|x_1)$

Figure source

To run Gibbs sampling, we need to derive the full conditional distribution for each parameter

- $P\left(\mu \middle| \boldsymbol{y}, \boldsymbol{\beta}, \sigma_\beta^2, \pi, \sigma_e^2\right)$
- $P\left(\beta_j \middle| \boldsymbol{y}, \boldsymbol{\beta}_{-j}, \sigma_\beta^2, \pi, \sigma_e^2\right)$
- $P\left(\sigma_\beta^2 \middle| \boldsymbol{y}, \boldsymbol{\beta}, \pi, \sigma_e^2\right)$
- $P\left(\pi \middle| \boldsymbol{y}, \boldsymbol{\beta}, \sigma_\beta^2, \sigma_e^2\right)$
- $P\left(\sigma_e^2 \middle| \boldsymbol{y}, \boldsymbol{\beta}, \sigma_\beta^2, \pi\right)$

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

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

Gibbs sampling

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

Samples reconstruct posterior distributions of parameters

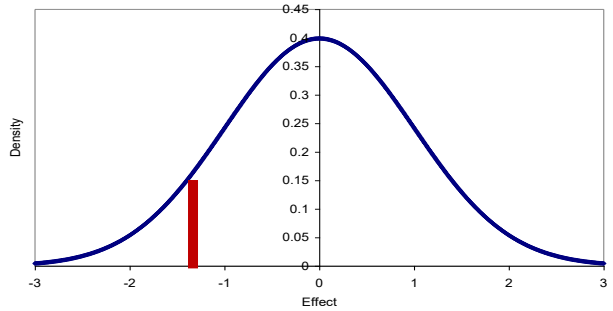
Gibbs Sampling

For example, for SNP effect β_j

- First sample if in zero or nonzero effect component of distribution (δ_j)
- Then if in nonzero component of the distribution, sample from

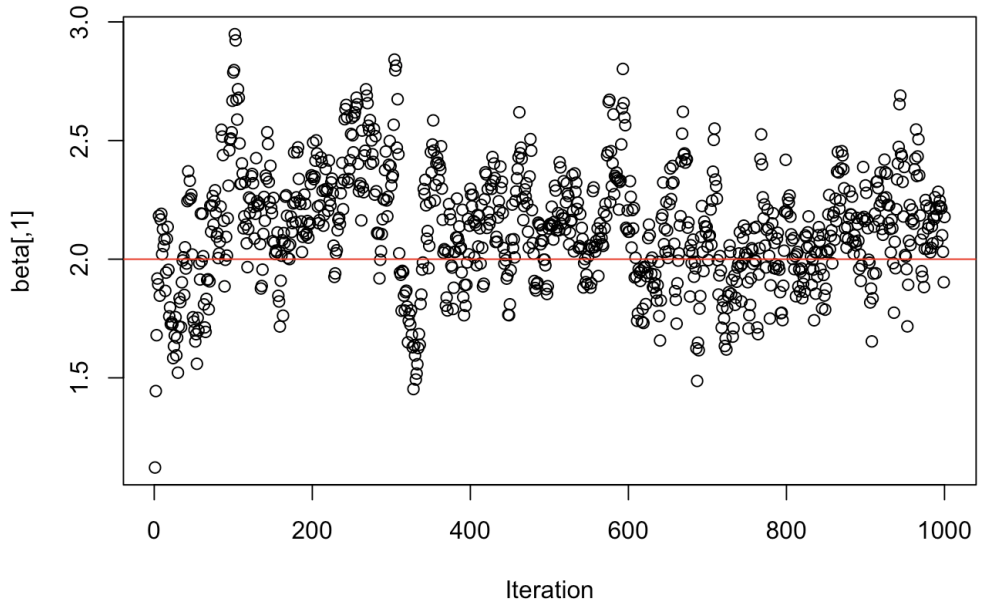
$$N\left(\frac{\mathbf{X}_j'(\mathbf{y}-\mathbf{1_n}\mu-\sum_{k\neq j}\mathbf{X}_k'\beta_k)}{\mathbf{X}_j'\mathbf{X}_j+\sigma_e^2/\sigma_\beta^2},\sigma_e^2/\left(\mathbf{X}_j'\mathbf{X}_j+\sigma_e^2/\sigma_\beta^2\right)\right)$$

Introduce shrinkage

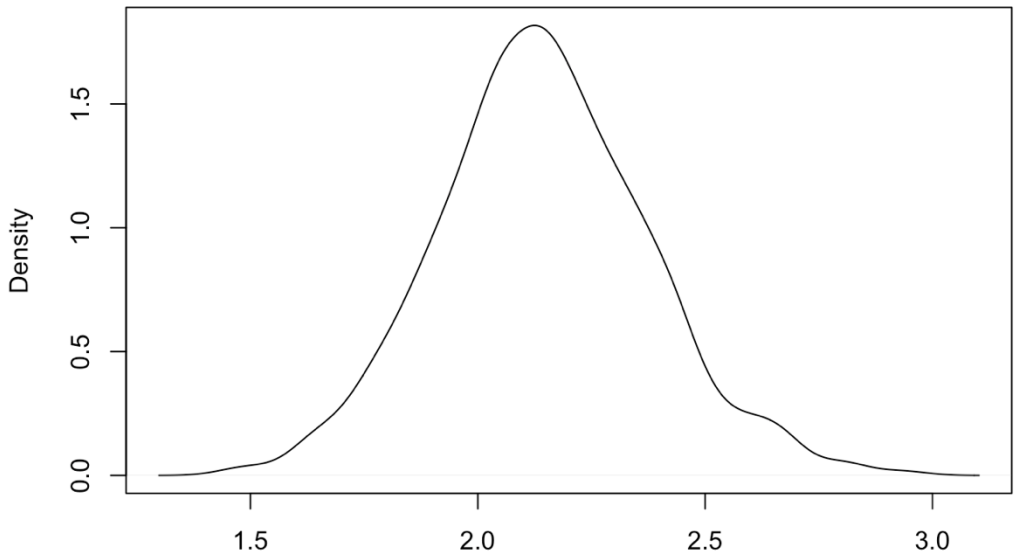


Gibbs sampling

Trace plot



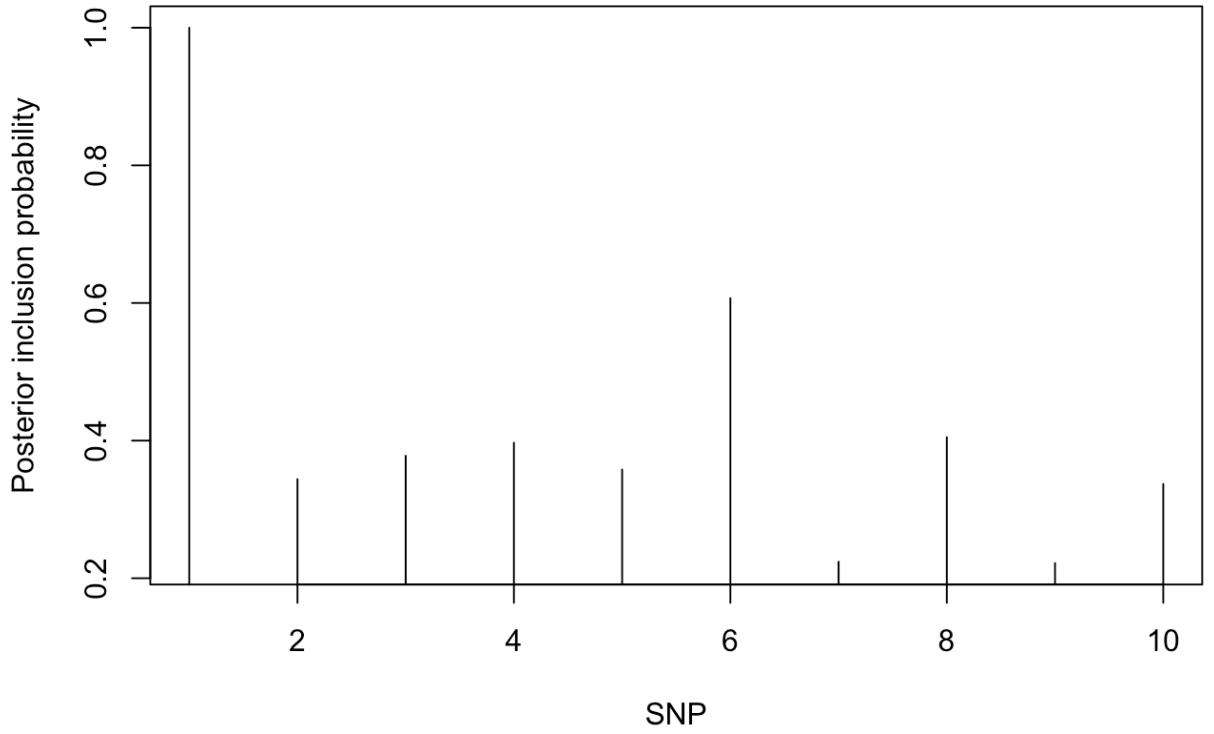
Posterior distribution



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

As a method of fine-mapping

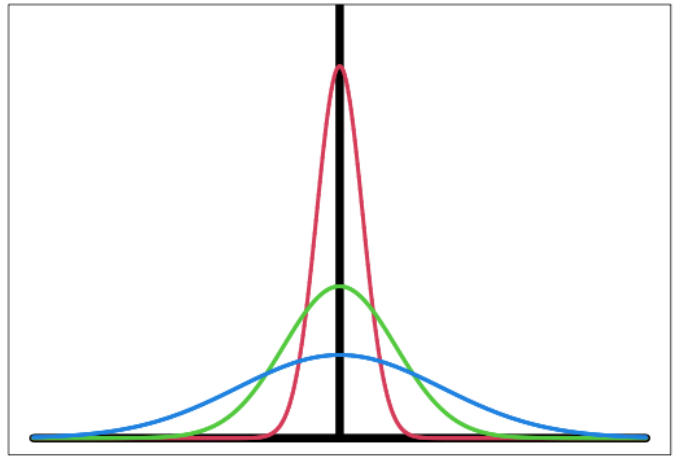
Posterior inclusion probability (PIP):
probability that the SNP is included in the model with a nonzero effect.



Model

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

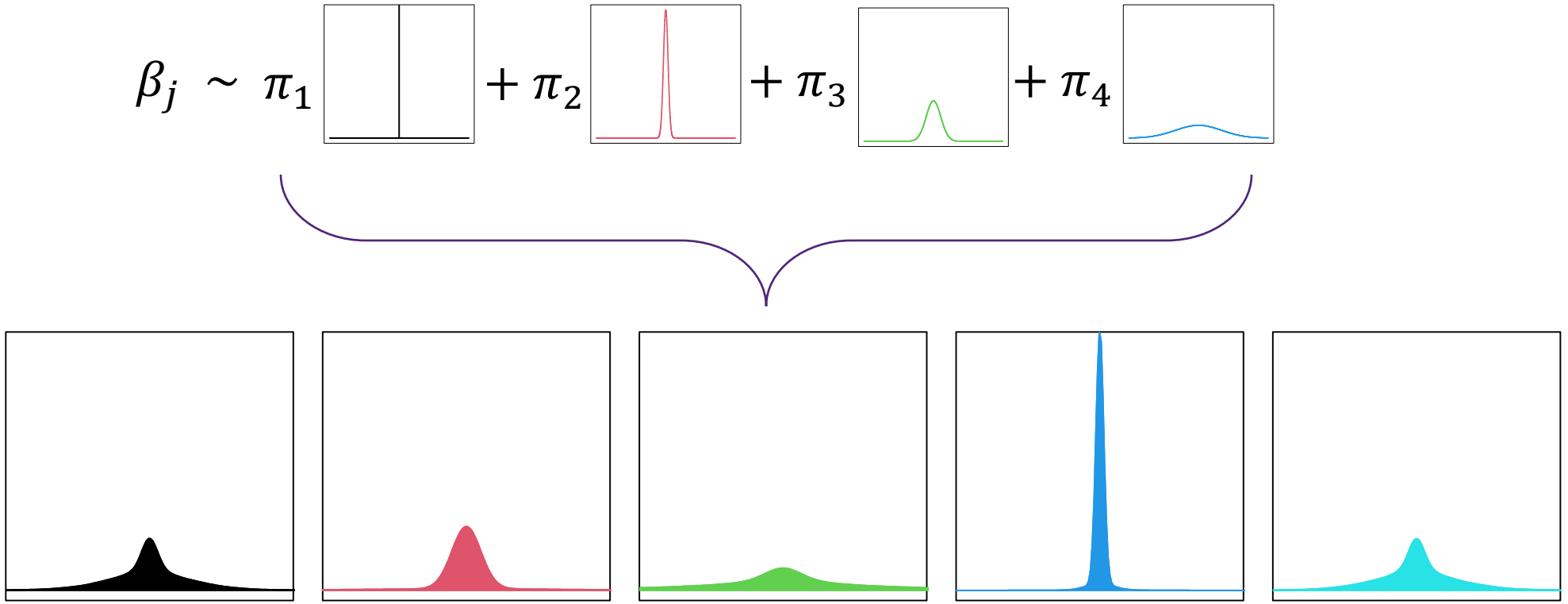
$$\beta_j|\pi, \sigma^2_\beta = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0, \gamma_2\sigma^2_\beta) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C\sigma^2_\beta) & \text{with probability } 1 - \sum_{c=1}^{C-1} \pi_c, \end{cases}$$



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

BayesC is a special case of BayesR with two components

Why use multi-normal mixture?



Account for almost any distribution!

Practical Exercise Part 1: BayesR

https://gctbhub.cloud.edu.au/data/teaching/

Bayesian methods using GWAS summary statistics (sumstats)

Motivation

- Best prediction methods take genetic values as random effect.
- These methods require individual genotypes and phenotypes.
- These data are often not publicly accessible.
- Computationally demanding with large # individuals/SNPs.
- Could be addressed by using GWAS summary statistics (sumstats).
- Methodology in human genetics has moved forward to use GWAS sumstats only.

Cell Genomics

CellPress
OPEN ACCESS

2021

Perspective

Workshop proceedings: GWAS summary statistics standards and sharing

Jacqueline A.L. MacArthur,[1,2,*] Annalisa Buniello,[1] Laura W. Harris,[1] James Hayhurst,[1] Aoife McMahon,[1] Elliot Sollis,[1] Maria Cerezo,[1] Peggy Hall,[3] Elizabeth Lewis,[1] Patricia L. Whetzel,[1] Orli G. Bahcall,[4] Inês Barroso,[5] Robert J. Carroll,[6] Michael Inouye,[7,8,9] Teri A. Manolio,[3] Stephen S. Rich,[10] Lucia A. Hindorff,[3] Ken Wiley,[3] and Helen Parkinson[1,*]

Table 1. Recommended standard reporting elements for GWAS SumStats

Data element	Column header	Mandatory/Optional
variant id	variant_id	One form of variant ID is mandatory, either rsID or chromosome, base pair location, and genome build[a]
chromosome	chromosome	
base pair location	base_pair_location	
p value	p_value	Mandatory
effect allele	effect_allele	Mandatory
other allele	other_allele	Mandatory
effect allele frequency	effect_allele_frequency	Mandatory
effect (odds ratio or beta)	odds_ratio or beta	Mandatory
standard error	standard_error	Mandatory
upper confidence interval	ci_upper	Optional
lower confidence interval	ci_lower	Optional

PRIMER

Check for updates

Genome-wide association studies

Emil Uffelmann[1], Qin Qin Huang[2], Nchangwi Syntia Munung[3], Jantina de Vries[3], Yukinori Okada[4,5], Alicia R. Martin[6,7,8], Hilary C. Martin[2], Tuuli Lappalainen[9,10,12] and Danielle Posthuma[1,11] ✉

Table 3 | **Databases of GWAS summary statistics**

Database	Content
GWAS Catalog[110]	GWAS summary statistics and GWAS lead SNPs reported in GWAS papers
GeneAtlas[8]	UK Biobank GWAS summary statistics
Pan UKBB	UK Biobank GWAS summary statistics
GWAS Atlas[273]	Collection of publicly available GWAS summary statistics with follow-up in silico analysis
FinnGen results	GWAS summary statistics released from FinnGen, a project that collected biological samples from many sources in Finland
dbGAP	Public depository of National Institutes of Health-funded genomics data including GWAS summary statistics
OpenGWAS database	GWAS summary data sets
Pheweb.jp	GWAS summary statistics of Biobank Japan and cross-population meta-analyses

For a comprehensive list of genetic data resources, see REF.[13]. GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism.

What are the minimum data required?

Given the standard GWAS with genotypes being allelic counts (0/1/2), the minimum data required for PGS prediction include:

- SNP marginal effect estimates
- Standard errors
- GWAS sample size

} GWAS sumstats

- LD correlations among SNPs → LD matrix

SNP marginal effect estimates

GWAS estimates effect of each SNP one at a time from single SNP regression, so the estimate is marginal to (unconditional on) other SNPs.

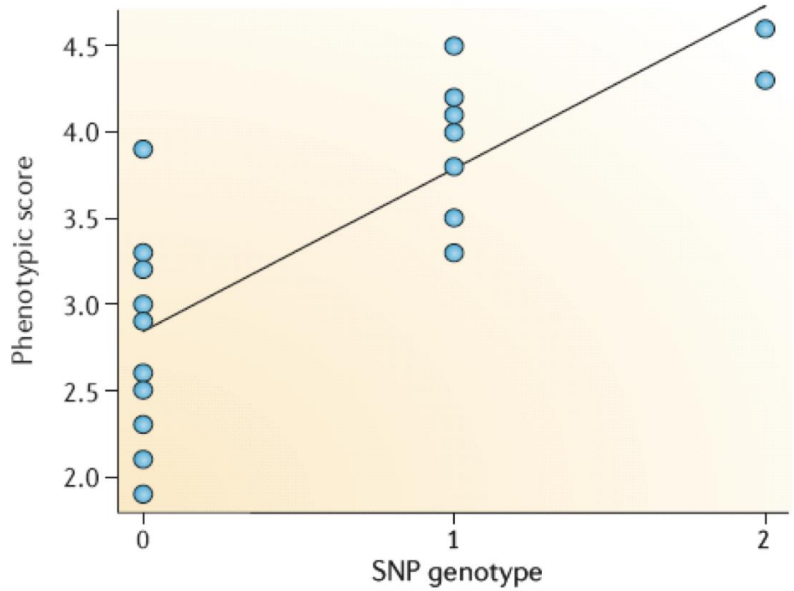
$$b_j = \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\mathbf{y}$$

Assuming $\mathbf{X}$ has been standardised with column mean zero and variance one, then

$$\mathbf{X}_j'\mathbf{X}_j = nVar\left(\mathbf{X}_j\right) = n$$

And

$$b_j = \frac{1}{n}\mathbf{X}_j'\mathbf{y}$$



Note that it has the inner product of the SNP genotypes and the phenotypes.

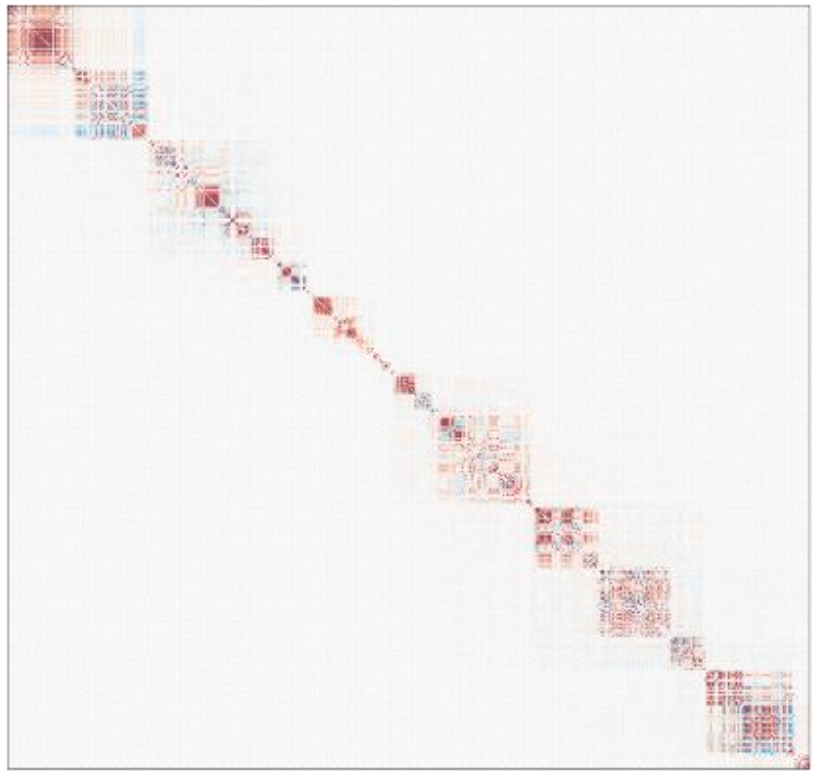
Linkage disequilibrium (LD) correlations

Usually obtained from a reference population

LD correlation matrix

$$\mathbf{R} = \frac{1}{n}\mathbf{X}'\mathbf{X}$$

assuming $\mathbf{X}$ is standardised with mean zero and variance one



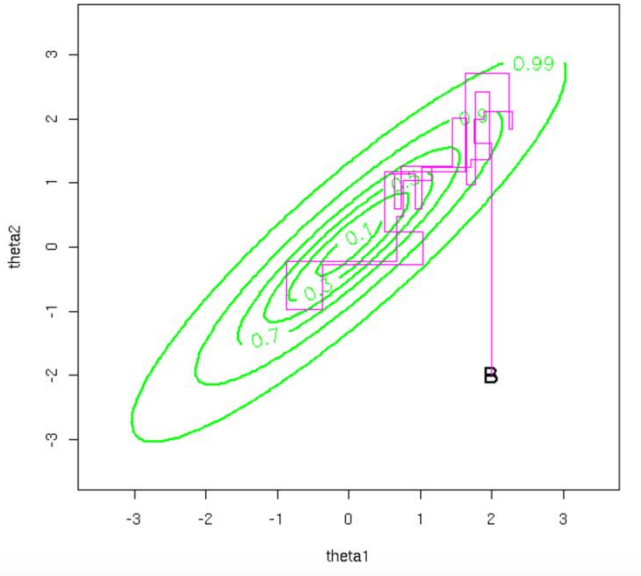
Compare BayesR and SBayesR algorithms

Gibbs sampling

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

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

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

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

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

Compare BayesR and SBayesR algorithms

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

*All **X'y** and **X'X** can be replaced by n**b** and n**R***

Algorithm 1 – Individual level data algorithm

Initialise parameters and read genotypes and phenotypes in PLINK binary format
Initialise $\mathbf{y}^* = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
for i :=1 **to** number of iterations **do**
  for i :=1 **to** p **do**
    Calculate $r_j^* = \mathbf{x}_j'\mathbf{y}^*$
    Calculate $r_j = r_j^* + \mathbf{x}_j'\mathbf{x}_j\beta_j^{(i-1)}$
    Calculate $\sigma_c^2 = \sigma_\beta^2\gamma_{\delta_j=c}$ for each of C classes (e.g., BayesR C=4 and $\boldsymbol{\gamma} = (0, 0.0001, 0.001, 0.01)$)
    Calculate the left hand side $l_{jc} = \mathbf{x}_j'\mathbf{x}_j + \frac{\sigma_\epsilon^2}{\sigma_c^2}$ for each of the C classes
    Calculate the log densities of given $\delta_j = c$ using $\log(\mathcal{L}_c) = -\frac{1}{2}\left[\log\left(\frac{\sigma_c^2 l_{jc}}{\sigma_\epsilon^2}\right) - \frac{r_j^2}{\sigma_\epsilon^2 l_{jc}}\right] + \log(\pi_c)$, where π_c is the current
    Calculate the full conditional posterior probability for $\delta_j = c$ for C classes with $\mathbb{P}(\delta_j = c|\boldsymbol{\theta}, \mathbf{y}) = \frac{1}{\sum_{l=1}^{C}\exp[\log(\mathcal{L}_l) - \log(\mathcal{L}_c)]}$
    Using full conditional posterior probabilities sample class membership for $\beta_j^{(i)}$ using categorical random variable sampler
    Given class sample SNP effect $\beta_j^{(i)}$ from $N\left(\frac{r_j}{l_{jc}}, \frac{\sigma_\epsilon^2}{l_{jc}}\right)$
    Given SNP effect adjust corrected phenotype side $(\mathbf{y}^*)^{(i)} = (\mathbf{y}^*)^{(i-1)} - \mathbf{x}_j\left(\beta_j^{(i)} - \beta_j^{(i-1)}\right)$
  od
  Sample update from full conditional for σ_β^2 from scaled inverse chi-squared distribution $\widetilde{\nu}_\beta = \nu_\beta + q$ and $\widetilde{S^2}_\beta = \frac{\nu_\beta S_\beta^2 + \sum_{j=1}^{q}\frac{\beta_j^2}{\gamma_c}}{\nu_\beta + q}$, where q is the number of non-zero variants
  Sample update from full conditional for σ_ϵ^2 from scaled inverse chi-squared distribution $\widetilde{\nu}_e = n + \nu_e$ and scale parameter $\widetilde{S}_e^2 = \frac{SSE + \nu_e S_e^2}{n + \nu_e}$ and $SSE = \mathbf{y}^{*\prime}\mathbf{y}^*$
  Sample update from full conditional for $\boldsymbol{\pi}$, which is Dirichlet$(C, \mathbf{c} + \boldsymbol{\alpha})$, where $\mathbf{c}$ is a vector of length C and contains the counts of the number of variants in each variance class and $\boldsymbol{\alpha} = (1, \ldots, 1)$
  Calculate genetic variance for h_{SNP}^2 calculation using $\sigma_g^2 = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta})$
  Calculate $h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}$
od

Algorithm 2 Summary data algorithm

Initialise parameters and read summary statistics
Reconstruct $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ from summary statistics and LD reference panel
Calculate $\mathbf{r}^* = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$
for i :=1 **to** number of iterations **do**
  for i :=1 **to** p **do**
    Calculate $\mathbf{r}_j = \mathbf{r}_j^* + \mathbf{x}_j'\mathbf{x}_j\beta_j$
    Calculate $\sigma_c^2 = \sigma_\alpha^2\gamma_{\delta_j=c}$ for each fo C classes (e.g., SBayesR C=4 and $\boldsymbol{\gamma} = (0, 0.01, 0.1, 1)'$)
    Calculate the left hand side $l_{jc} = \mathbf{x}_j'\mathbf{x}_j + \frac{\sigma_\epsilon^2}{\sigma_c^2}$ for each of the C classes
    Calculate the log densities of given $\delta_j = c$ using $\log(\mathcal{L}_c) = -\frac{1}{2}\left[\log\left(\frac{\sigma_c^2 l_{jc}}{\sigma_\epsilon^2}\right) - \frac{r_j^2}{\sigma_\epsilon^2 l_{jc}}\right] + \log(\pi_c)$, where π_c is the current
    Calculate the full conditional posterior probability for $\delta_j = c$ for C classes with $\mathbb{P}(\delta_j = c|\boldsymbol{\theta}, \mathbf{y}) = \frac{1}{\sum_{l=1}^{C}\exp[\log(\mathcal{L}_l) - \log(\mathcal{L}_c)]}$
    Using full conditional posterior probabilities sample class membership for $\beta_j^{(i)}$ using categorical random variable sampler
    Given class sample SNP effect $\beta_j^{(i)}$ from $N\left(\frac{\mathbf{r}_j}{l_{jc}}, \frac{\sigma_\epsilon^2}{l_{jc}}\right)$
    Given SNP effect adjust corrected right hand side $(\mathbf{r}^*)^{(i+1)} = (\mathbf{r}^*)^{(i)} - \mathbf{X}'\mathbf{x}_j\left(\beta_j^{(i+1)} - \beta_j^{(i)}\right)$. $\mathbf{X}'\mathbf{x}_j$ is the jth column of $\mathbf{X}'\mathbf{X}$.
  od
  Sample update from full conditional for σ_α^2 from scaled inverse chi-squared distribution $\widetilde{\nu}_\alpha = \nu_0 + q$ and $\widetilde{\tau^2}_\alpha = \frac{\nu_0\tau_0^2 + \sum_{j=1}^{q}\frac{\beta_j^2}{\gamma_{\delta_j}}}{\nu_0 + q}$, where q is the number of non-zero variants
  Sample update from full conditional for σ_ϵ^2 from scaled inverse chi-squared distribution $\widetilde{\nu}_e = n + \nu_e$ and scale parameter $\widetilde{\tau}_e^2 = \frac{SSE + \nu_e\tau_e^2}{n + \nu_e}$ and $SSE = \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{r}^* - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$
  Sample update from full conditional for $\boldsymbol{\pi}$, which is Dirichlet$(C, \mathbf{c} + \boldsymbol{\alpha})$, where $\mathbf{c}$ is a vector of length C and contains the counts of the number of variants in each variance class.
  Calculate genetic variance for h_{SNP}^2 calculation using $\sigma_g^2 = MSS/n$, where $MSS = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'r^*$
  Calculate $h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}$
od

Lloyd-Jones et al. 2019 Nat Comm, Supplementary Note

From individual- to summary-level 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

LD correlation matrix

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

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).

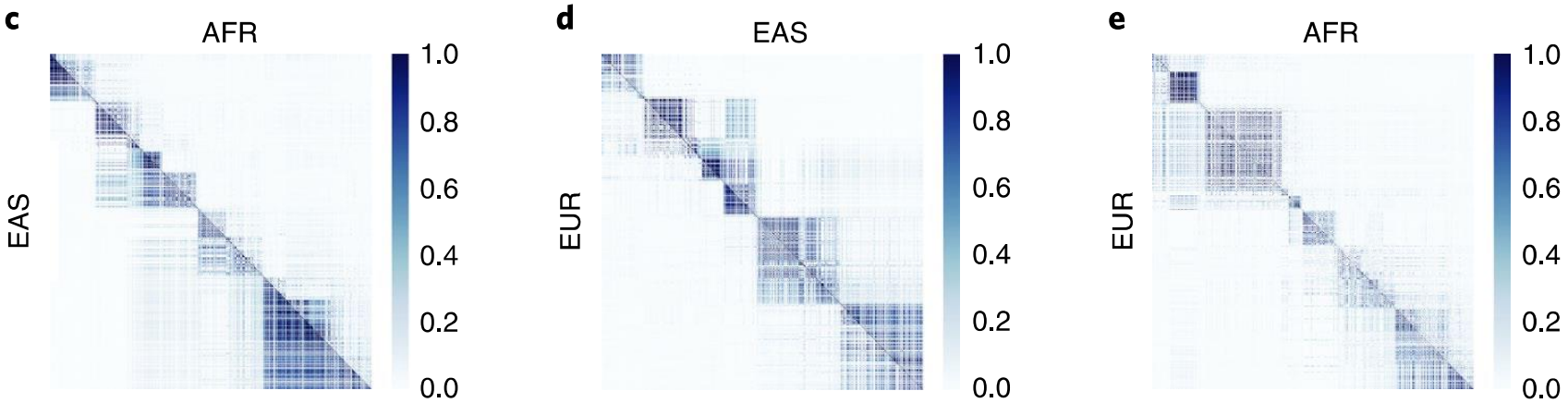
- However, when LD is estimated from a reference sample, SBayes is only an approximation to Bayes.

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

Assumptions regarding LD reference

LD reference population matches with GWAS population in genetics

- No systematic differences in LD → **same ancestry**
- Minimum sampling variance in LD → LD ref sample size cannot be too small

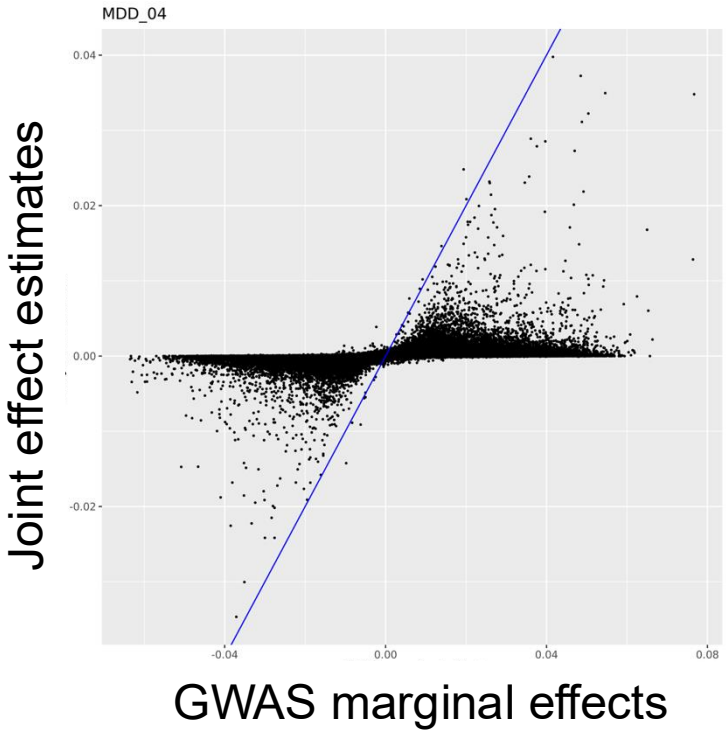


Failure to meet this assumption can result in a convergence issue!

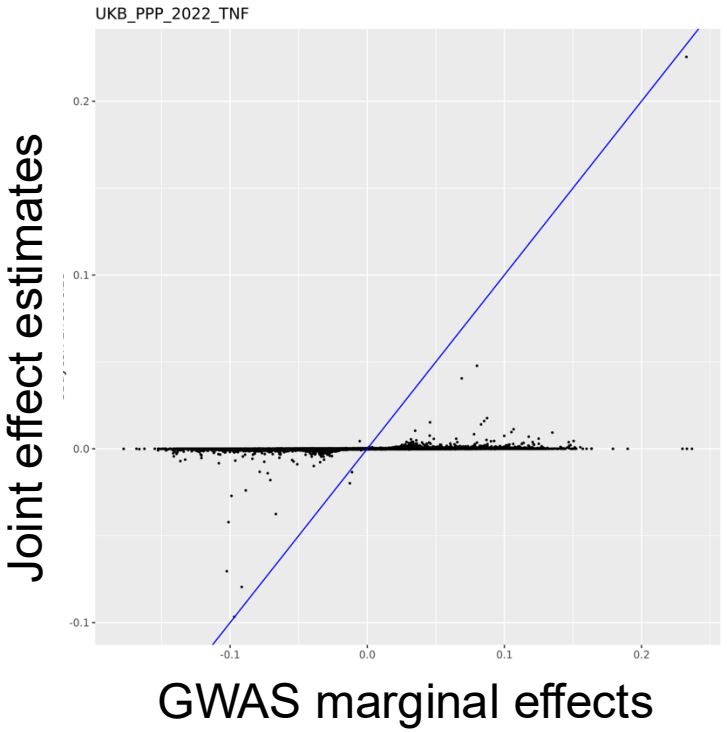
Always good to check SNP effect estimates

GWAS marginal effect size vs. Estimated joint effect size

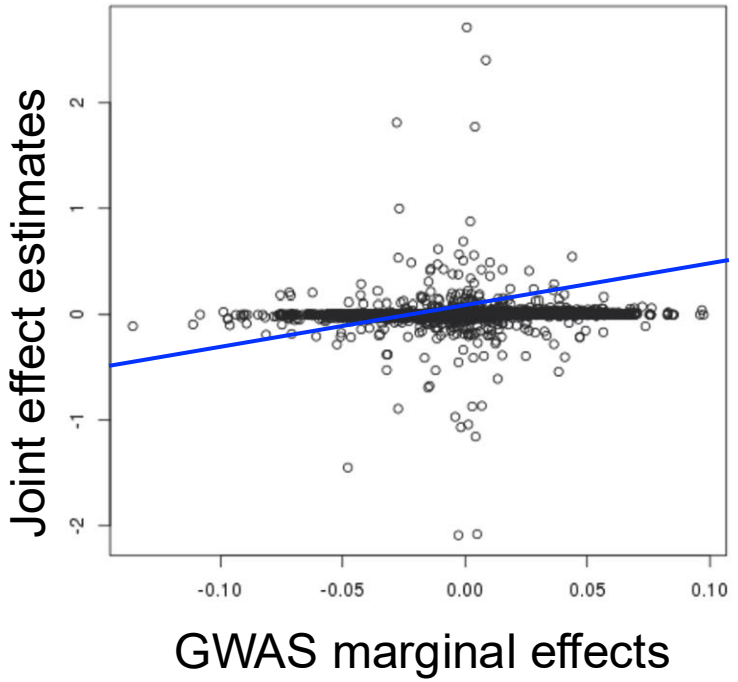
Most common



Presence of large effects



Bad convergence!



How do different methods handle this issue?

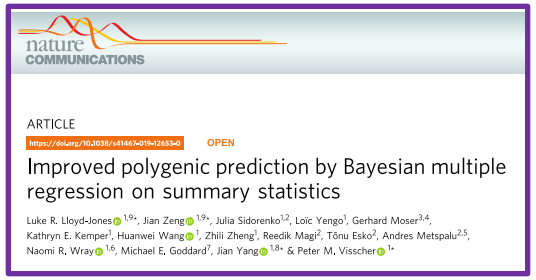
- Run multiple MCMC chains with different starting values
  - e.g., LDpred2

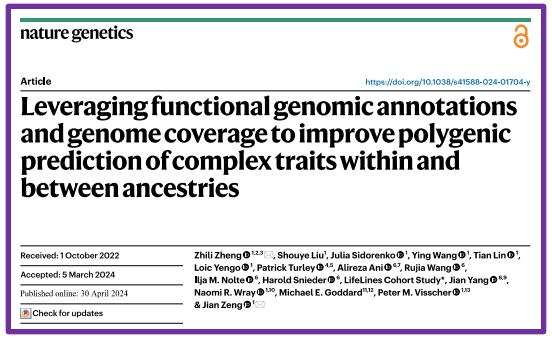
- Force an (ad hoc) minimum strong shrinkage to SNP effects
  - e.g., PRS-CS

$$N\left(\frac{\mathbf{X}_j'(\mathbf{y}-\mathbf{1_n}\mu-\sum_{k\neq j}\mathbf{X}_k'\beta_k)}{\mathbf{X}_j'\mathbf{X}_j+\sigma_e^2/\sigma_\beta^2},\sigma_e^2/(\mathbf{X}_j'\mathbf{X}_j+\sigma_e^2/\sigma_\beta^2)\right)$$

the shrinkage parameter is set to be greater than a threshold value

- Regulate LD matrices
  - e.g., SBayesR uses chromosome-wide shrunk LD matrices
  - e.g., SBayesRC uses eigen-decomposed matrices from LD blocks



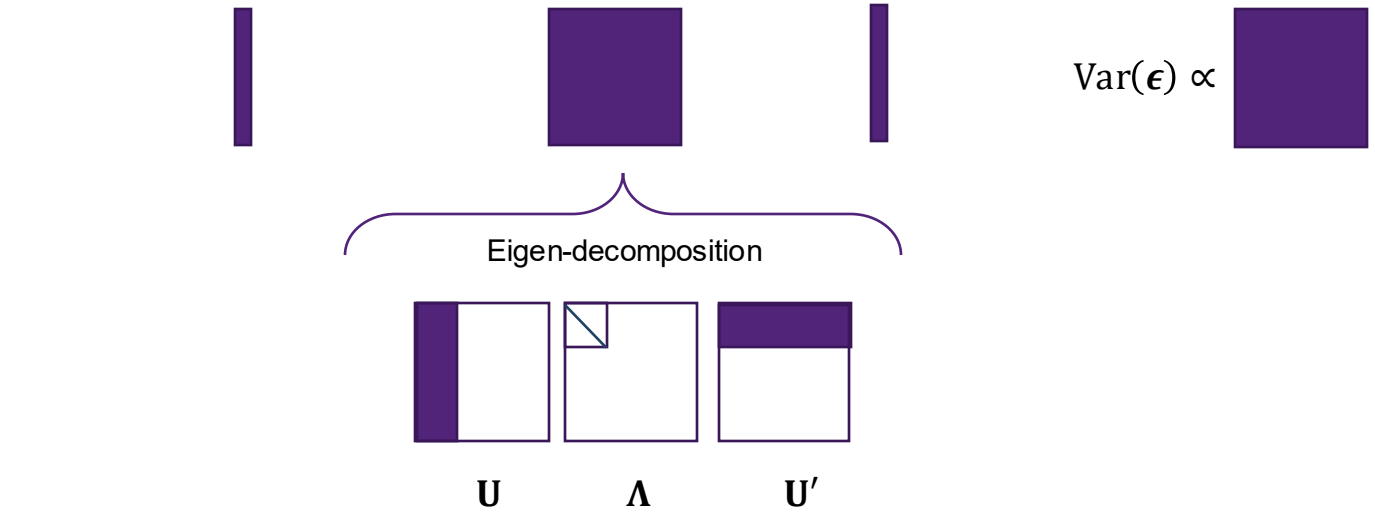


Low-rank model (fits 7M SNPs or more)

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



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

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

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

Practical Exercise Part 2: SBayesR

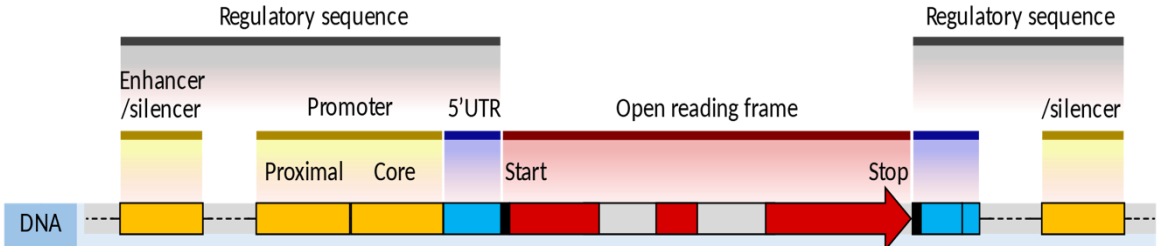
https://gctbhub.cloud.edu.au/data/teaching/

Incorporating functional genomic annotations

Functional genomic annotations

Functional genomic annotations provide orthogonal information useful for polygenic prediction.

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



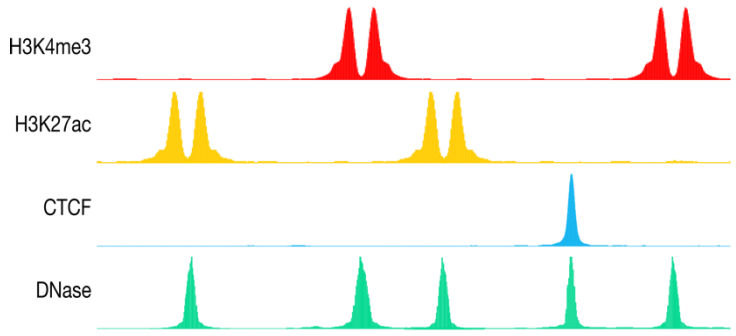


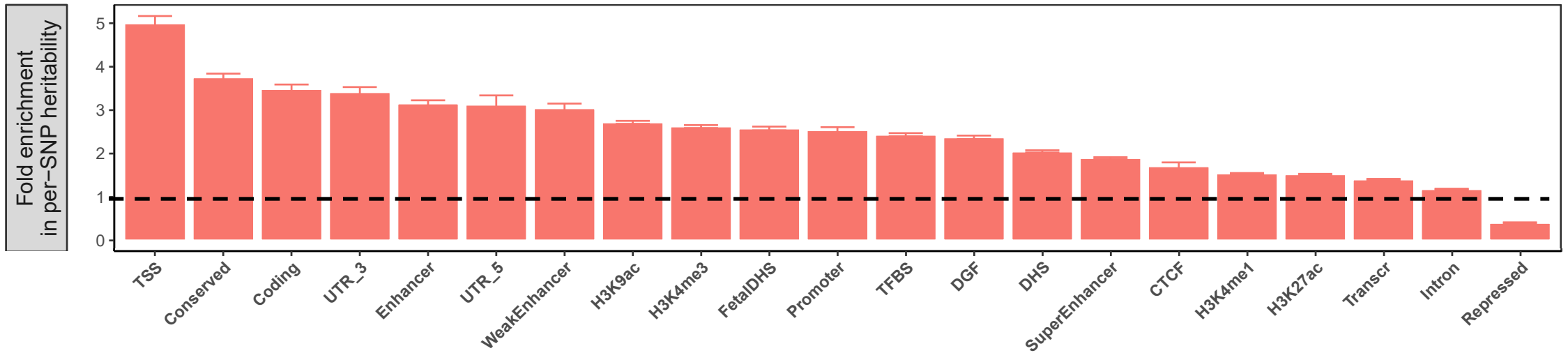
Image from ENCODE

Functional genomic annotations

Functional genomic annotations provide orthogonal information useful for polygenic prediction.

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

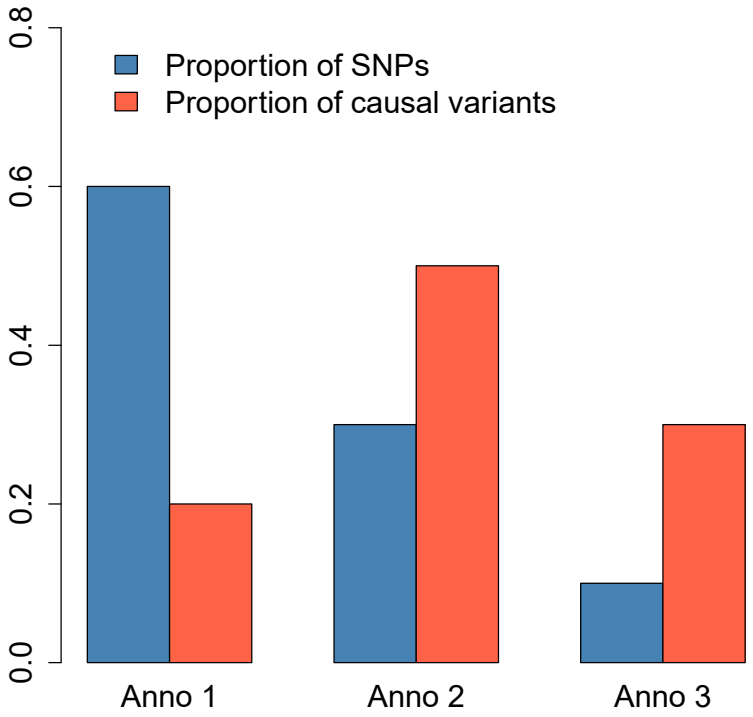
Zeng et al 2021 Nature Communications



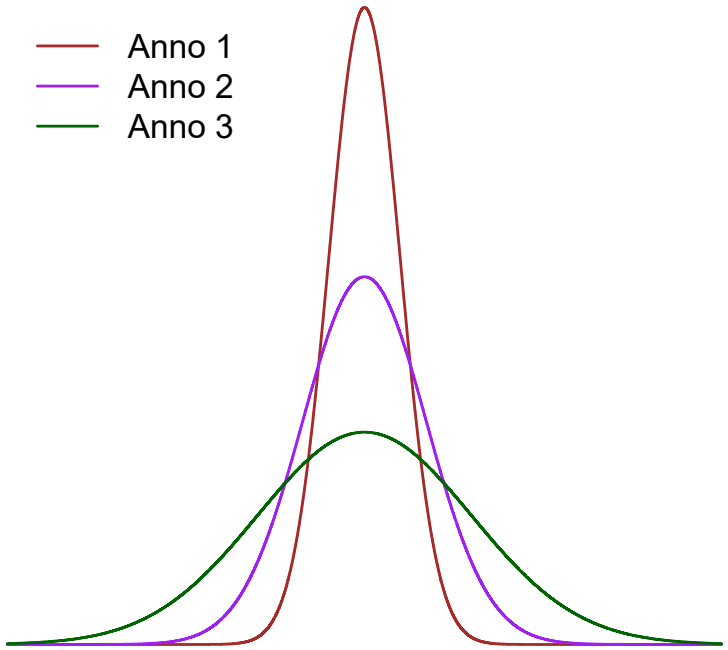
Opportunities/challenges

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



Literature

The University Of Queensland AUSTRALIA

nature communications

Explore content · About the journal · Publish with us

nature › nature communications › articles › article

Article | Open Access | Published: 18 October 2021

Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets

Carla Márquez-Luna, Steven Gazal, Po-Ru Loh, Samuel S. Kim, Nicholas Furlotte, Adam Auton, 23andMe Research Team & Alkes L. Price

LDpred-funct

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Leveraging functional annotations in genetic risk prediction for human complex diseases

Yiming Hu, Qiongshi Lu, Ryan Powles, Xinwei Yao, Can Yang, Fang Fang, Xinran Xu, Hongyu Zhao

AnnoPred

Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data

Jianxin Shi, Ju-Hyun Park, Jubao Duan, Sonja T. Berndt, Winton Moy, Kai Yu, Lei Song, William Wheeler, Xing Hua, Debra Silverman, Montserrat Garcia-Closas, Chao Agnes Hsiung, Jonine D. Figueroa, […], Nilanjan Chatterjee [view all]

P+T-funct-LASSO

Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits

I. M. MacLeod, P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes & M. E. Goddard

BMC Genomics **17**, Article number: 144 (2016) | Cite this article

6209 Accesses | **146** Citations | **9** Altmetric | Metrics

BayesRC

nature genetics

Explore content · About the journal · Publish with us

nature › nature genetics › articles › article

Article | Published: 07 April 2022

Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores

Omer Weissbrod, Masahiro Kanai, Huwenbo Shi, Steven Gazal, Wouter J. Peyrot, Amit V. Khera, Yukinori Okada, The Biobank Japan Project, Alicia R. Martin, Hilary K. Finucane & Alkes L. Price

Nature Genetics **54**, 450–458 (2022) | Cite this article

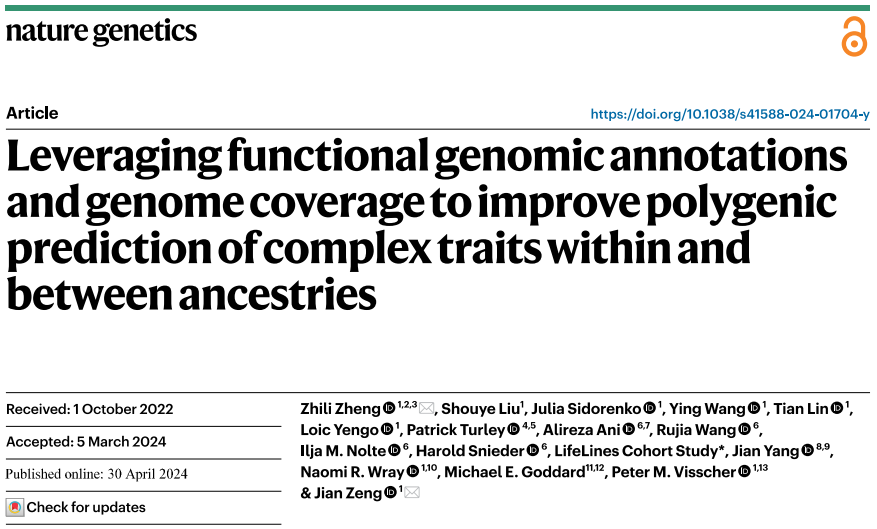
PolyPred

Gaps

Need new method that can

- simultaneously fit all SNPs and annotation data in a unified model
- account for variations in both causal variant proportion and causal effect distribution

Leveraging functional annotations for cross-ancestry prediction



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

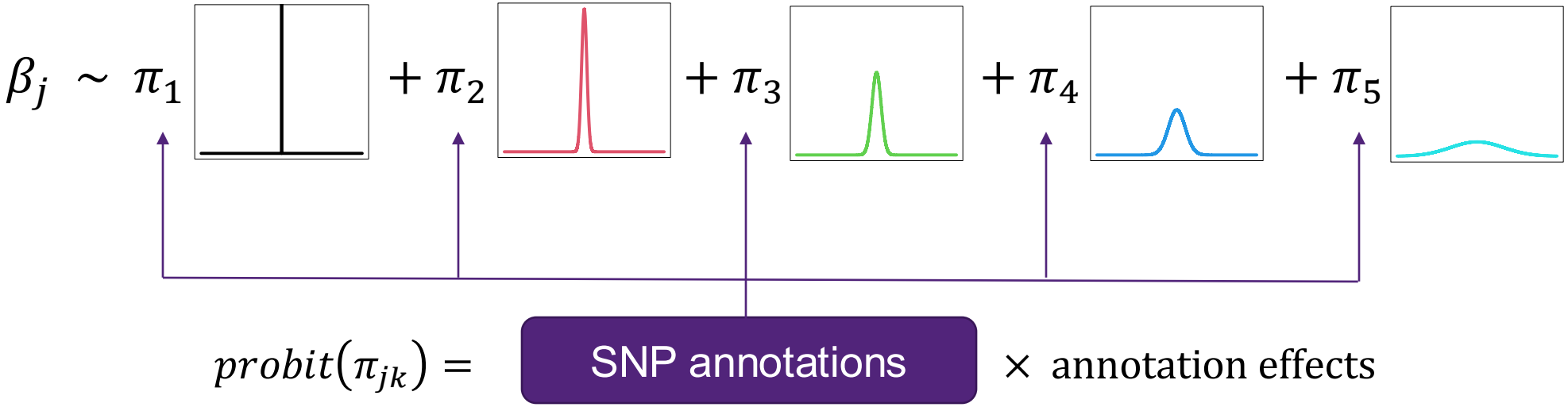
Accepted: 5 March 2024

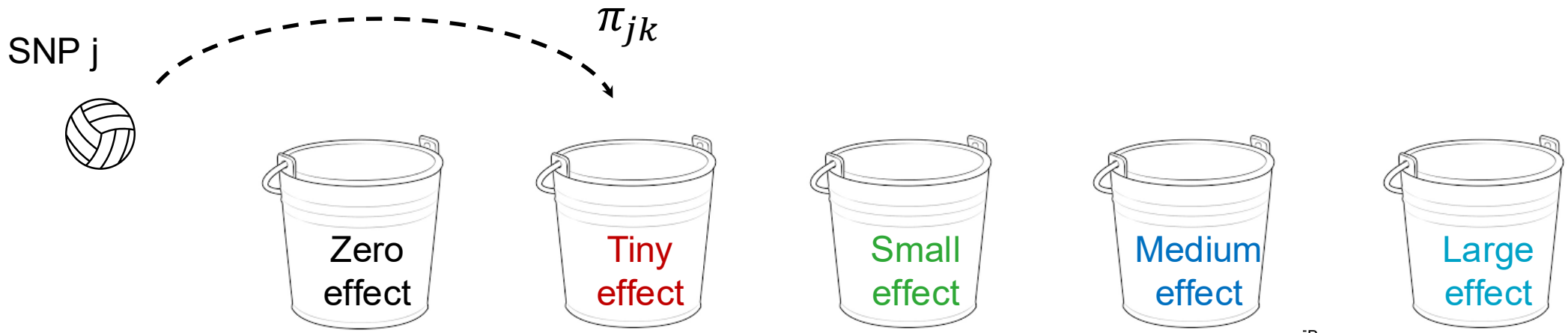
Published online: 30 April 2024

Check for updates

Zhili Zheng[1,2,3], Shouye Liu[1], Julia Sidorenko[1], Ying Wang[1], Tian Lin[1], Loic Yengo[1], Patrick Turley[4,5], Alireza Ani[6,7], Rujia Wang[6], Ilja M. Nolte[6], Harold Snieder[6], LifeLines Cohort Study*, Jian Yang[8,9], Naomi R. Wray[1,10], Michael E. Goddard[11,12], Peter M. Visscher[1,13] & Jian Zeng[1]

Incorporate functional annotations through a hierarchical prior:

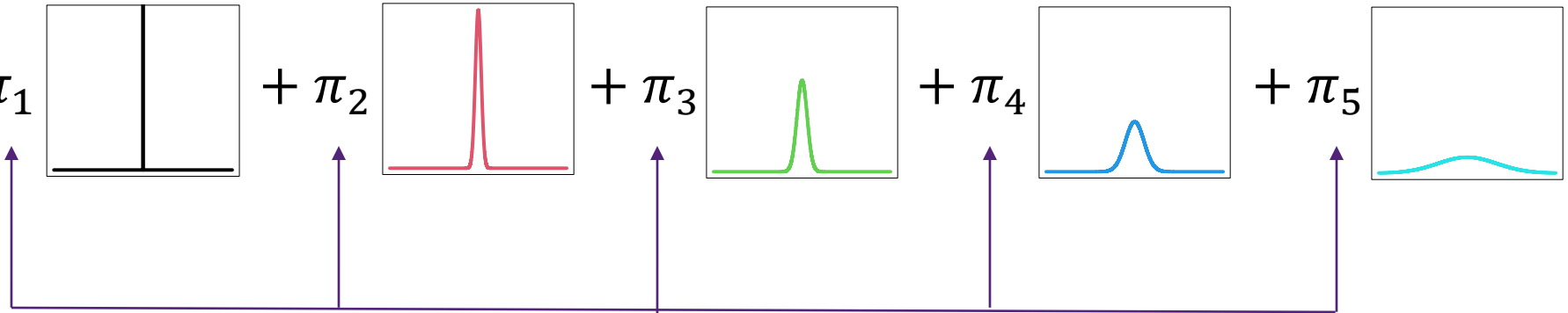




SBayesRC

Incorporate functional annotations through a hierarchical prior:

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



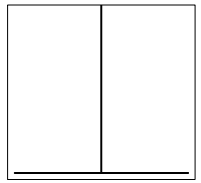
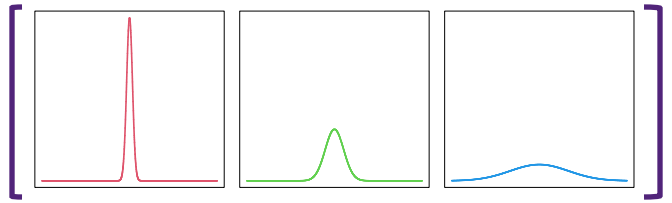
$$probit(\pi_{jk}) = \text{SNP annotations} \times \text{annotation effects}$$

Assumption	Pros	Cons
• Annotation effects are additive at the GLM scale.	• Estimation of conditional effects. • Allow annotation overlap. • Interpretation.	• # annotation effect parameters x 5. • $\pi_{j1} + \pi_{j2} + \pi_{j3} + \pi_{j4} + \pi_{j5} = 1$.

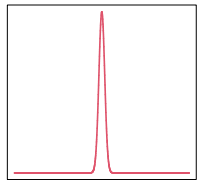
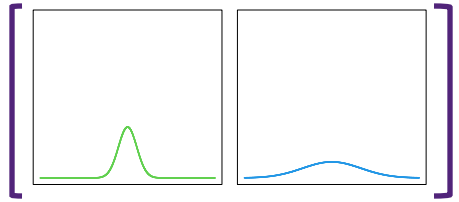
Reparameterisation of annotation effects

Suppose 4 components for simplicity

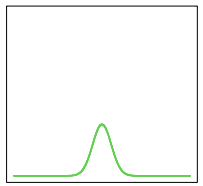
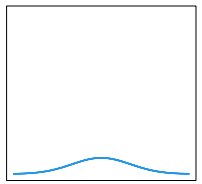
- A set of 2-component independent models:
- For all SNPs

$$\beta_j \sim (1 - p_2)$$  $$+ \ p_2$$ 

- For SNPs with nonzero effects (conditional on non-null SNPs)

$$\beta_j \sim (1 - p_3)$$  $$+ \ p_3$$ 

- For SNPs with at least medium effects (conditional on non-small-effect SNPs)

$$\beta_j \sim (1 - p_4)$$  $$+ \ p_4$$ 

p_2, p_3, p_4 are independent!

Reparameterisation of annotation effects

- Probit link function:

$$\Phi^{-1}(p) = \sum \text{SNP annotation} \times \text{annotation effect}$$

where Φ is the CDF of the standard normal distribution.

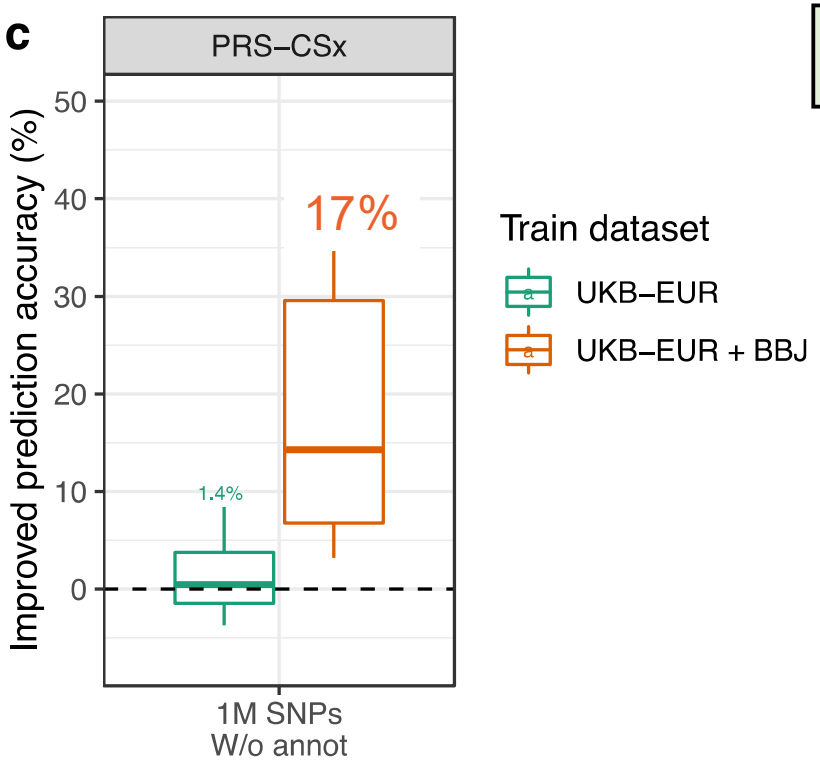
- It is straightforward to compute $p = \Phi(\cdot)$

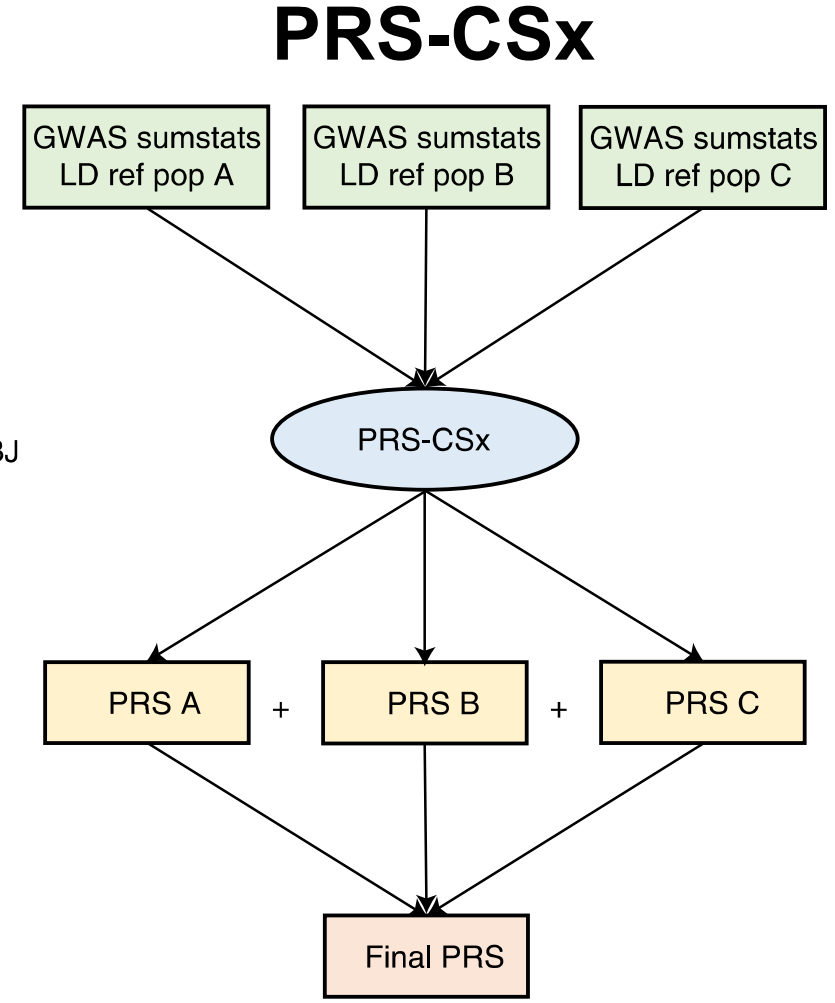
and $\pi_1 = 1 - p_2;\ \pi_2 = (1 - p_3)p_2;\ \pi_3 = (1 - p_4)p_3p_2;\ \pi_4 = p_2p_3p_4$

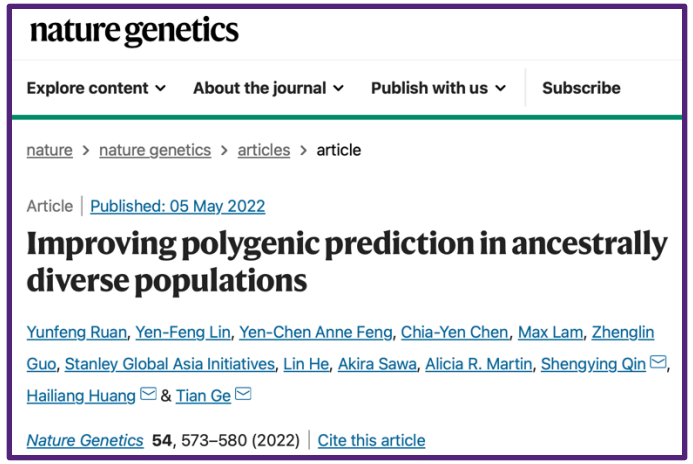
- Assume a normal prior distribution for each annotation effect.
- Gibbs sampling for all parameters.

Trans-ancestry prediction

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



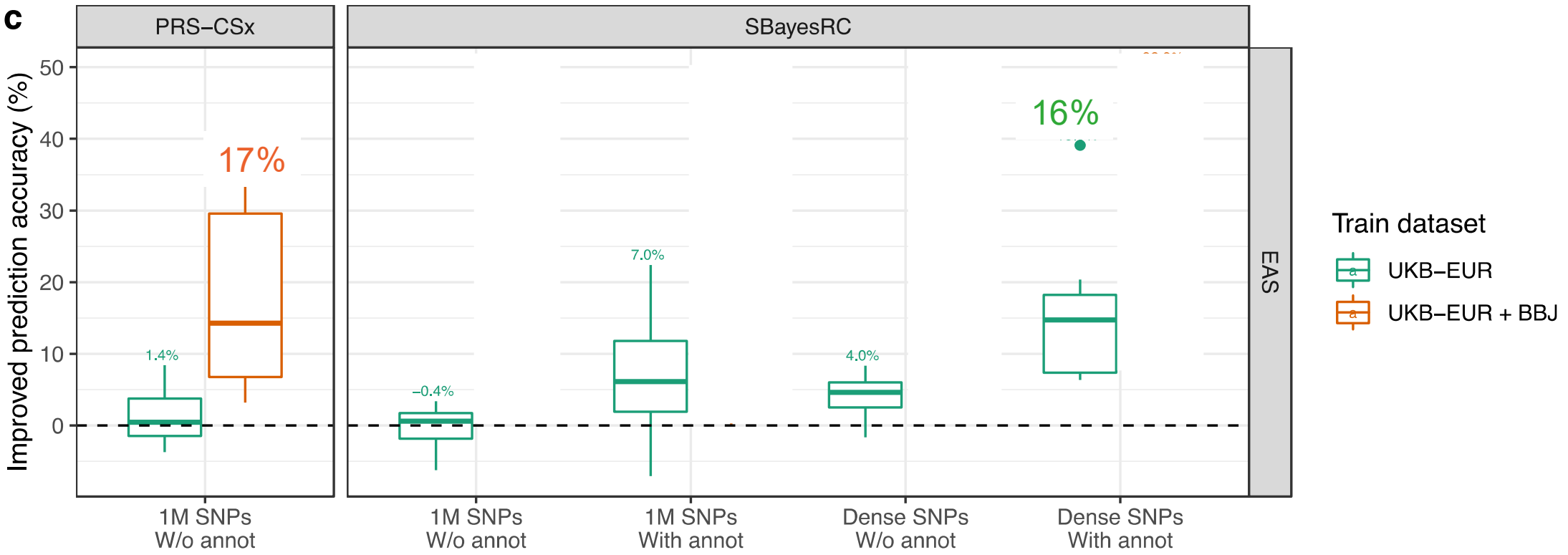




How important is functional annotation data compare to another GWAS dataset from the target ancestry?

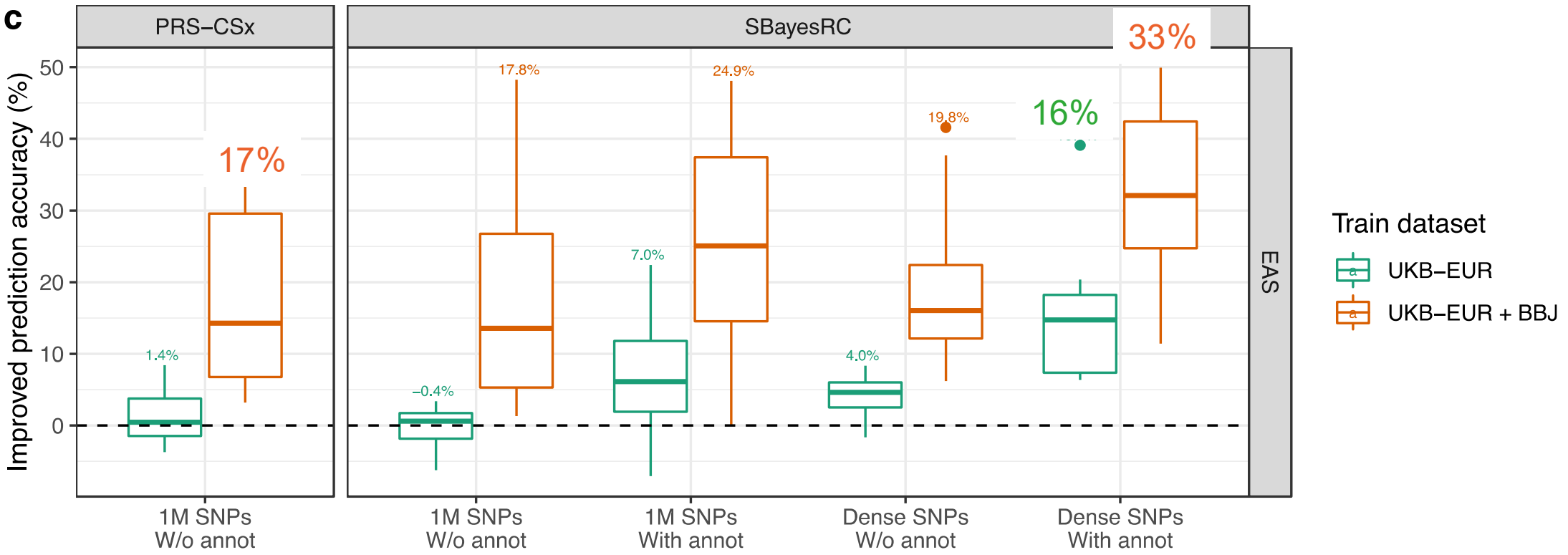
Trans-ancestry prediction

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



Trans-ancestry prediction

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



Practical Exercise Part 3: SBayesRC

https://gctbhub.cloud.edu.au/data/teaching/

Summary and challenges

Summary

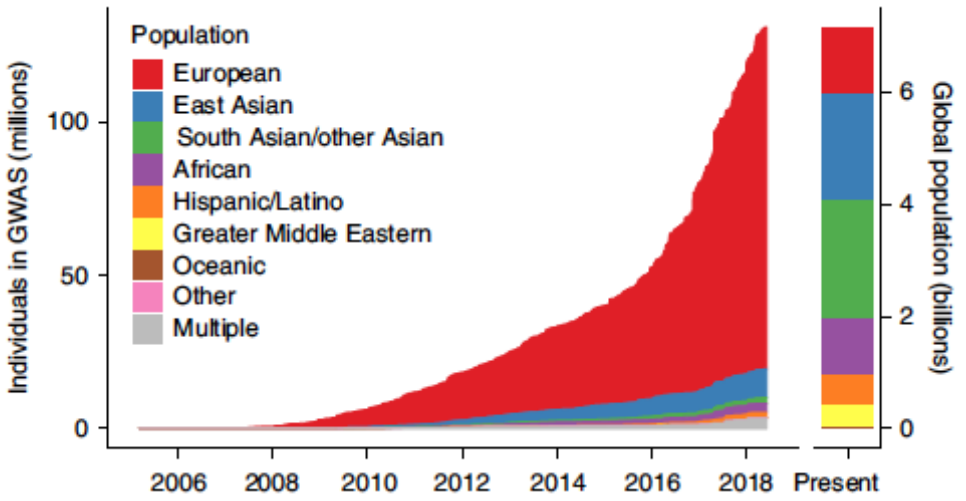
- PGS are imperfect genetic predictors with inherently limited accuracy.
- Being improved with more data and better methods.
- Bayesian approach allows us to incorporate prior knowledge in estimation of SNP effects.
- State-of-the-art Bayesian methods utilize GWAS summary statistics, which unleash the power of large GWAS sample size but may encounter convergence issue.
- Different methods mainly differ in prior assumption and implementation.
- Functional annotations with genome coverage provide orthogonal information to GWAS data and therefore boost prediction accuracy.

Ancestry

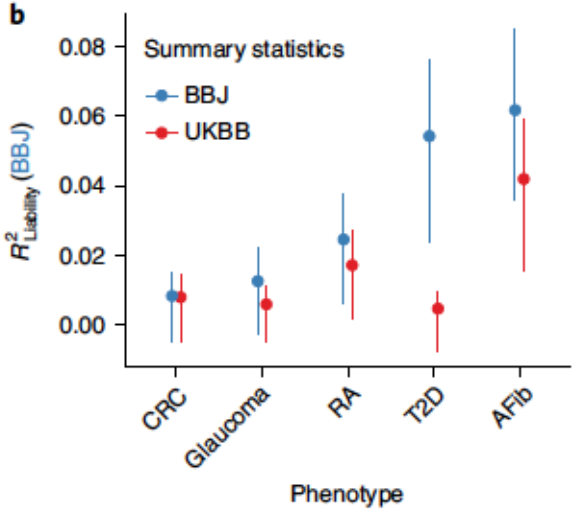
PERSPECTIVE
https://doi.org/10.1038/s41588-019-0379-x
nature genetics

Clinical use of current polygenic risk scores may exacerbate health disparities

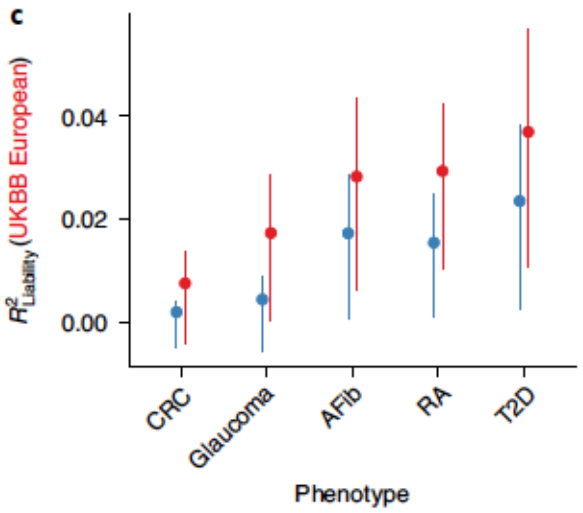
Alicia R. Martin[1,2,3*], Masahiro Kanai[1,2,3,4,5], Yoichiro Kamatani[5,6], Yukinori Okada[5,7,8], Benjamin M. Neale[1,2,3] and Mark J. Daly[1,2,3,9]

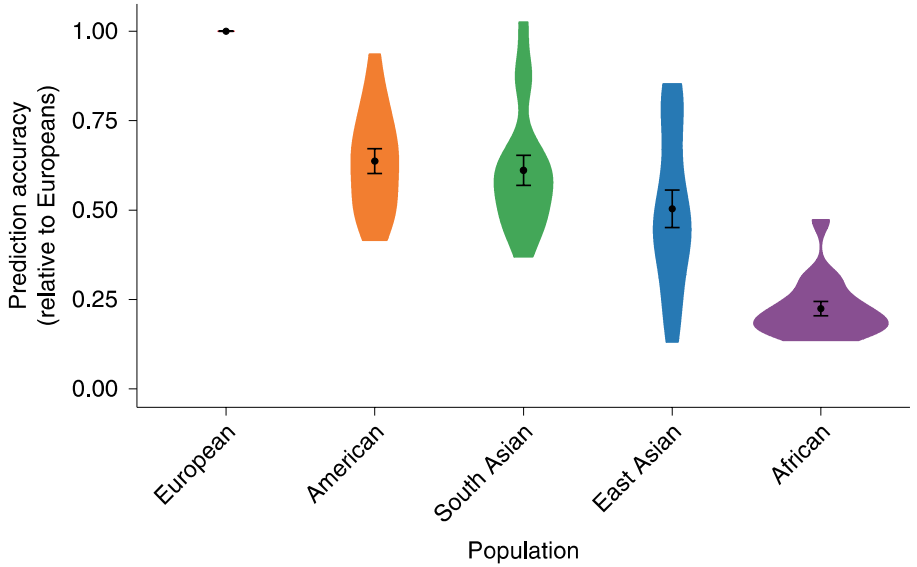


Predicted into Japanese

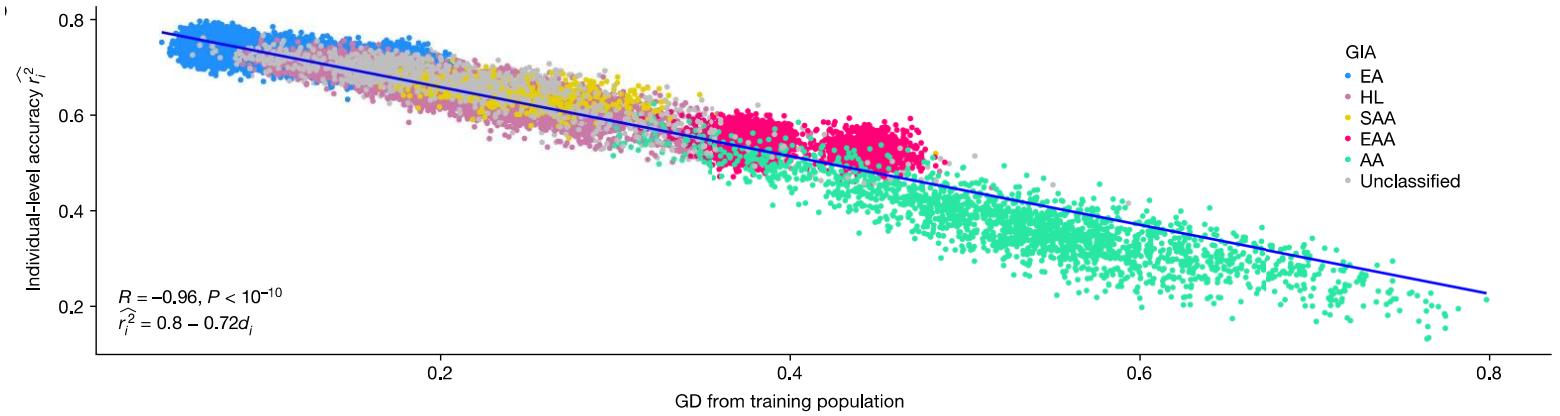


Predicted into European





Ancestry

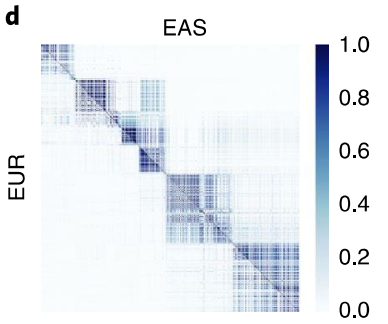


Issues

- Same causal variants
  - Different allele frequencies
  - LD differences
  - Different effect sizes
- Different causal variants
  - GxE
  - Different phenotype

In *general:*
We expect common causal variants to be shared across ancestries

But correlation structure differs



What is the maximum prediction accuracy we can get?

h_m^2 : True variance explained by the predictor depends on the SNP set - subscript m.

Variance explained by the predictor

$$R^2 = \frac{h_m^2}{1 + C}$$

C: captures the error in estimation

As C→ 0, R^2 → h_m^2

We want C to be as small as possible:

- C decreases as Discovery sample N increases
- C decreases as the number of SNPs in the SNP set m decreases

$$C \approx \frac{m}{N h_m^2}$$

Catch-22

As m gets smaller, h_m^2 also gets smaller

How to optimise m and h_m^2 to get max R^2 ?

Wray et al (2019) Complex trait prediction from genome data. Genetics

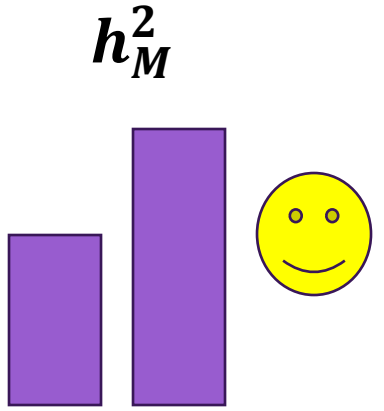
How about whole genome sequencing?

Maximum depends on maximising h_m^2

We use GWAS data so the maximum h_m^2 is the SNP-based heritability

Theoretical maximum depends on the heritability of the trait

$$R^2 \approx \frac{h_m^2}{1 + \frac{m}{Nh_m^2}}$$

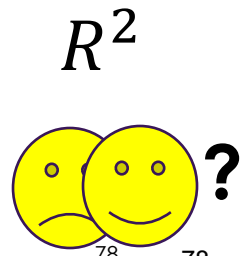


With whole genome sequencing the variance captured by all measured SNPs will increase

But the number of SNPs that we have estimate effect sizes for increases much more

Need MASSIVE discovery sample sizes for WGS associations

Also… rare variants are less likely to be shared across populations

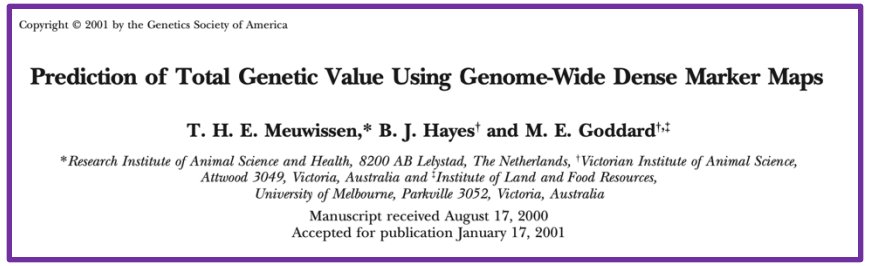


Wray et al (2019) Complex trait prediction from genome data. Genetics

78

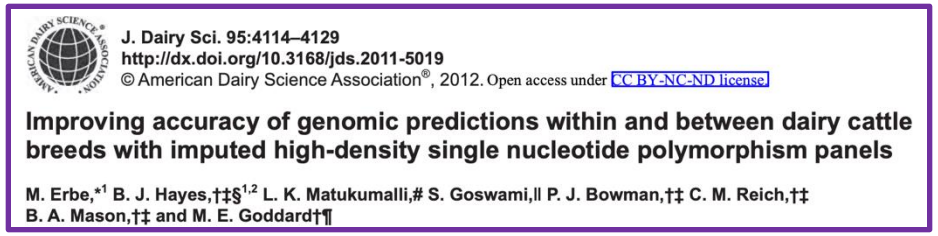
Reference

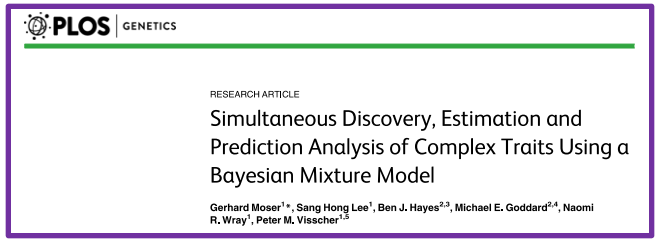
BayesA, BayesB:



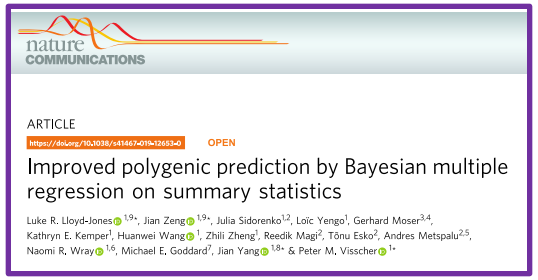
BayesC:

BayesR:

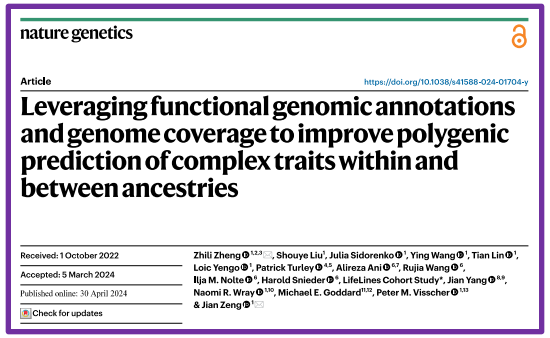




SBayesR:



SBayesRC:



LDpred2:

PRS-CS:

MegaPRS: