Bayesian Methods for PGS prediction

Jian Zeng

j.zeng@uq.edu.au

Polygenic prediction methodology

Polygenic score (PGS) is 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?

Clumping and P-value thresholding (C+PT)

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

Whole-genome regression approaches

Include all SNPs but adjust the effect sizes for LD.

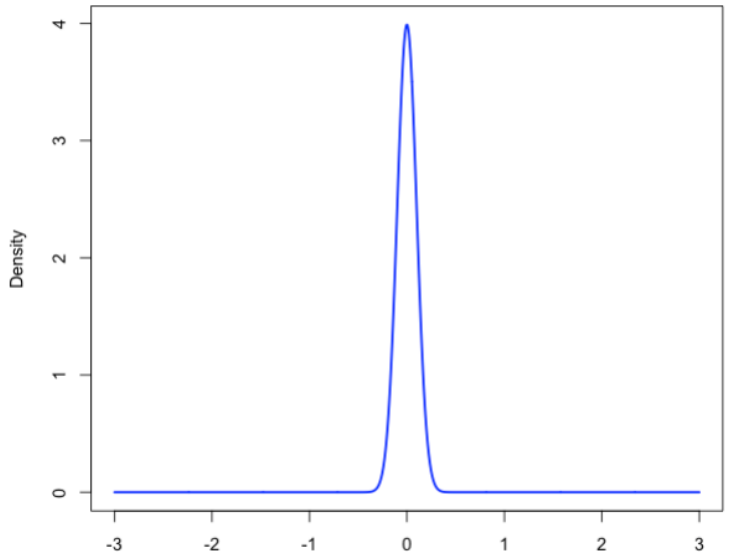
BLUP, Bayesian methods

Best Linear Unbiased Prediction (BLUP)

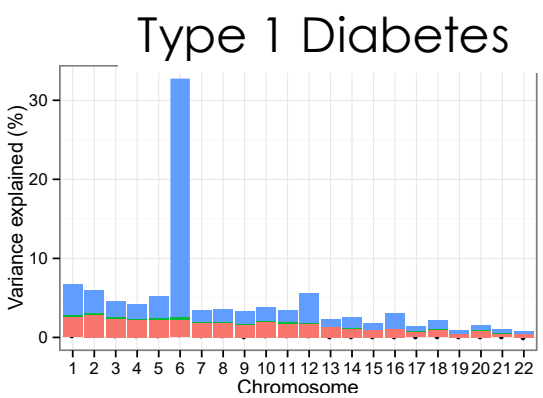
- Need to specify the shrinkage parameter

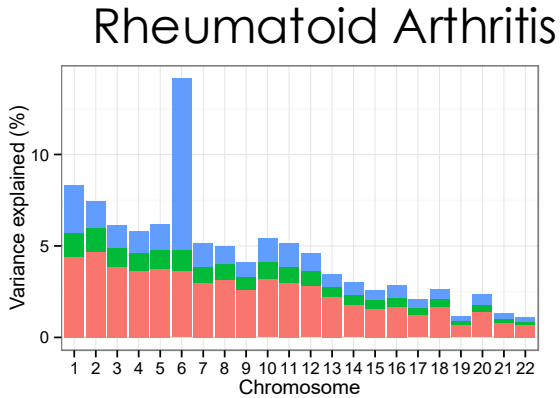
- Assumes SNPs effects are (*infinitesimal* model):
  - **all non-zero**
  - **all very small**
  - normally distributed

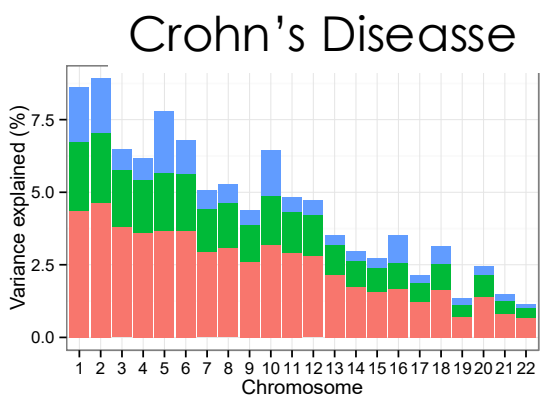
How realistic they are?

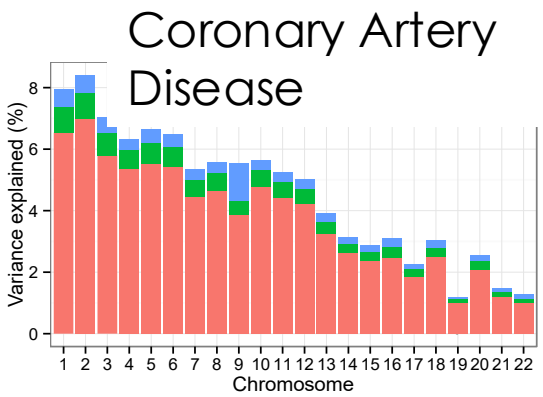


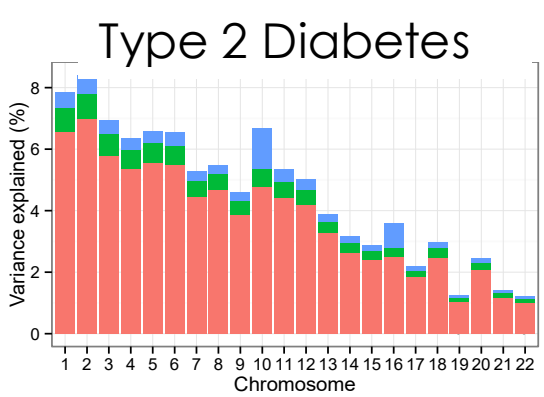
Do they look like *infinitesimal*?

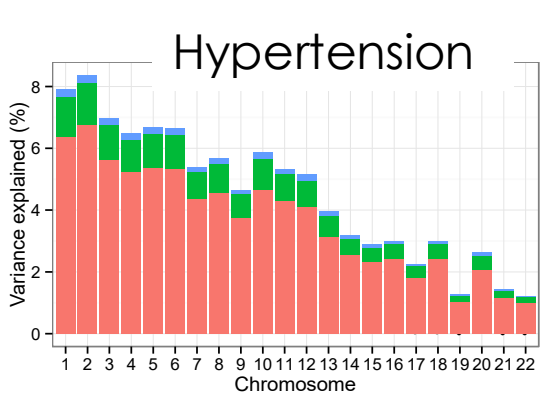


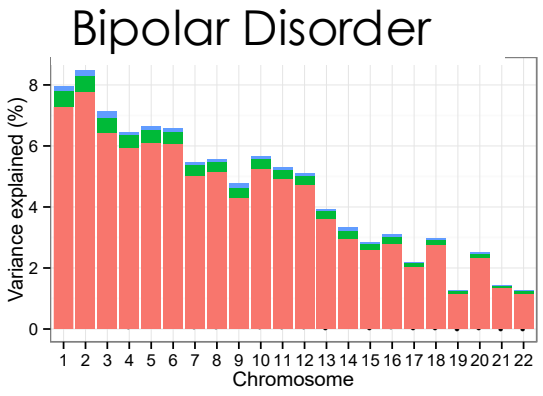


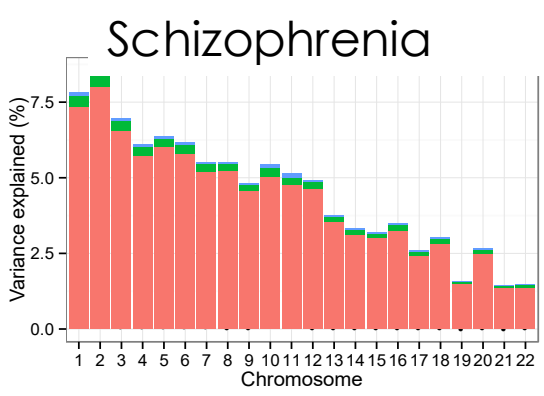












Mixture component

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

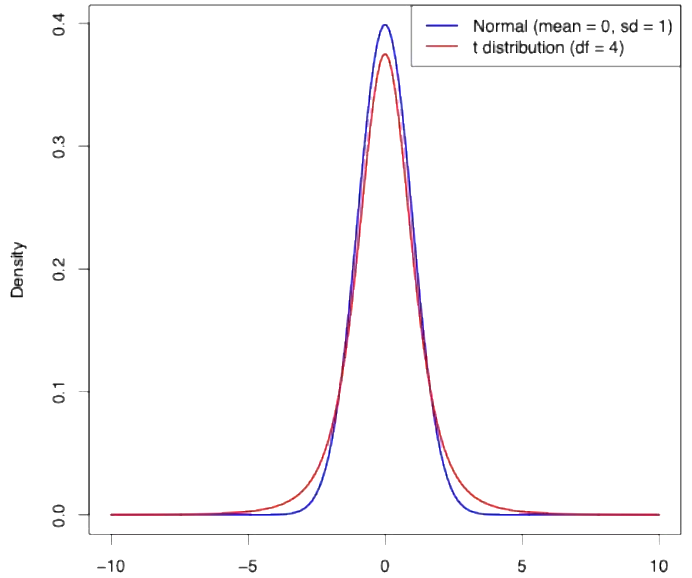
Gerhard Moser[1]*, Sang Hong Lee[1], Ben J. Hayes[2,3], Michael E. Goddard[2,4], Naomi R. Wray[1], Peter M. Visscher[1,5]

Many DNA variants contribute to genetic risk. Most have very small effects, but some have moderate to large effects.

Best Linear Unbiased Prediction (BLUP)

- Need to specify the shrinkage parameter

- Assumes SNPs effects are (*infinitesimal* model):
  - all non-zero
  - **all very small** → mostly very small but some larger
  - normally distributed → t distributed



BayesA (Meuwissen *et al*. *Genetics* 2001)

Bayesian methods

- Bayesian methods can estimate all parameters including SNP effects simultaneously

- Allow alternative assumptions regarding the distribution of SNP effects

What other distributions make sense?

Alternative distributions

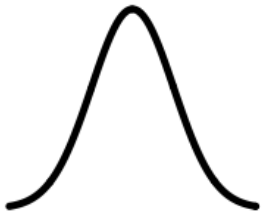
Assumption	Distribution of SNP effects	Method
Small number of moderate to large effects, many small effects	Students t	BayesA
Small number of moderate to large effects, many zero effects	Mixture, spike at zero, Students t	BayesB
Small number of small effects, many zero effects	Mixture, spike at zero,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

Bayesian alphabet models in animal breeding

Gianola. Priors in whole-genome regression: the bayesian alphabet returns. Genetics. 2013

Bayesian methods in human genetics

PRS-CS



Continuous shrinkage

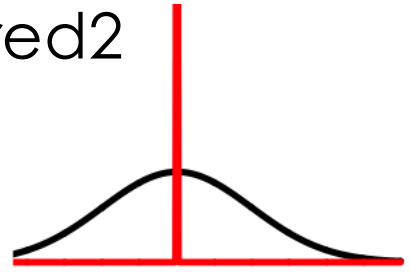
$$\beta_j \sim N(0, \phi\psi_j)$$

$$\psi_j \sim Ga(a, \delta_j)$$

$$\delta_j \sim Ga(b, 1)$$

Parameters a and b determine how aggressively to shrink small estimates and how much you don't shrink large ones

LDPred2

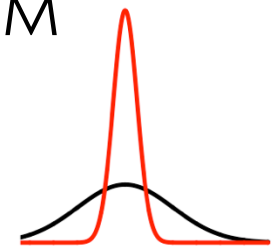


Spike-and-Slab

$$\beta_j \sim \begin{cases} N(0, \tau^2), & \pi \\ 0 & 1-\pi \end{cases}$$

π can be estimated from data, sparsity allowed, $\tau^2 = h^2/M\pi$

BSLMM

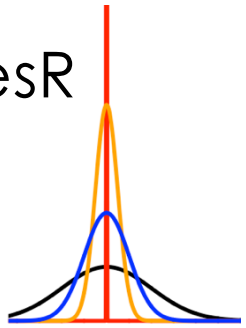


Normal mixture

$$\beta_j \sim \begin{cases} N(0, \sigma_b^2 + \sigma_u^2), & \pi \\ N(0, \sigma_u^2), & 1-\pi \end{cases}$$

σ_b^2 captures "sparse effects" as in the spike-and-slab model, and σ_u^2 captures small polygenic effects.

SBayesR

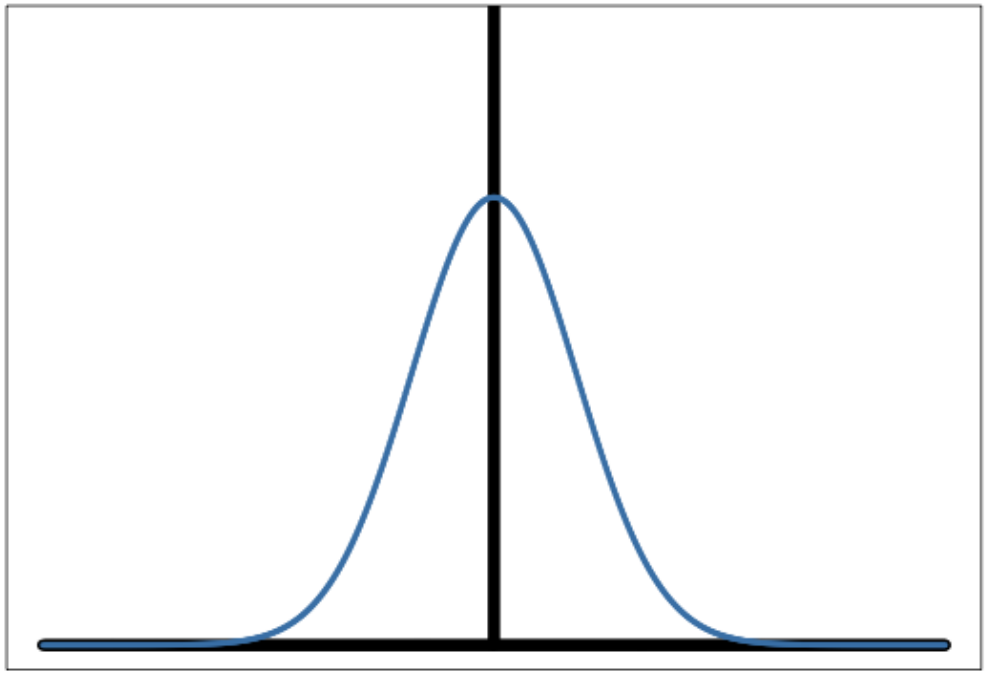


Flexible finite mixture of normal distributions, sparsity allowed

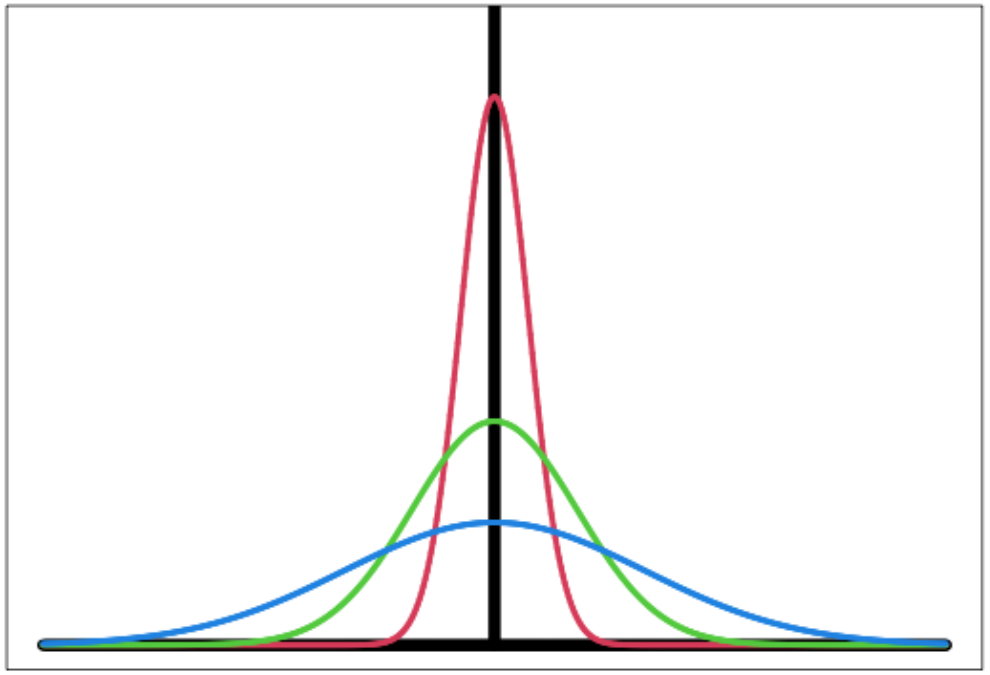
$$\beta_j \sim \begin{cases} 0, & \pi_1 \\ N(0, \gamma_2\sigma_b^2), & \pi_2 \\ \cdots & \\ N(0, \gamma_C\sigma_b^2), & 1 - \sum_{c=1}^{C-1} \pi_c \end{cases}$$

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

BayesC

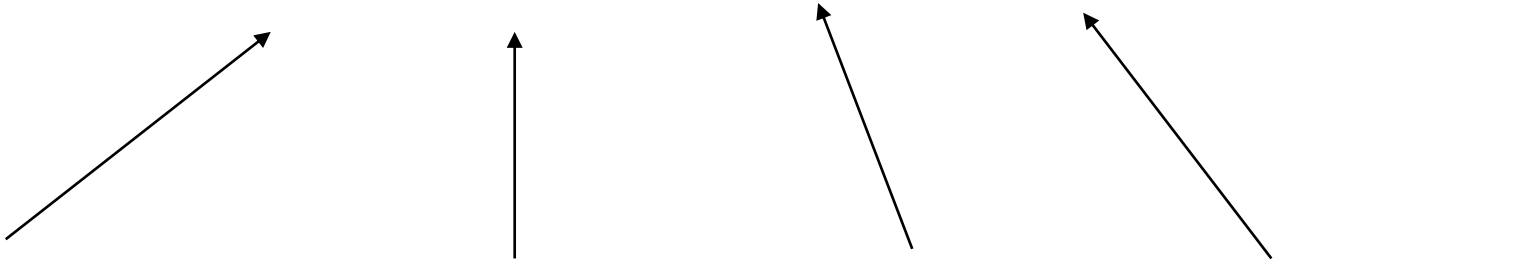


BayesR



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

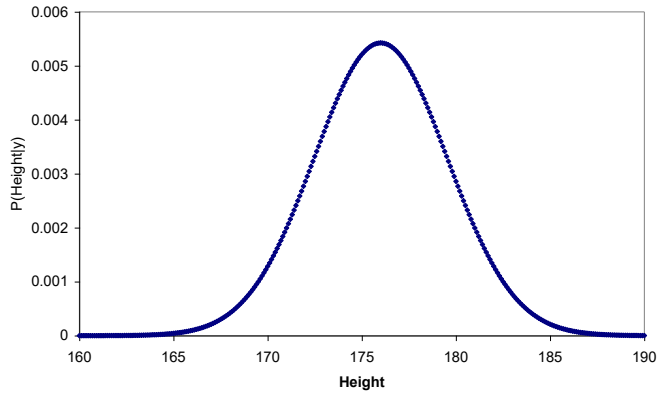
Example: estimate average height

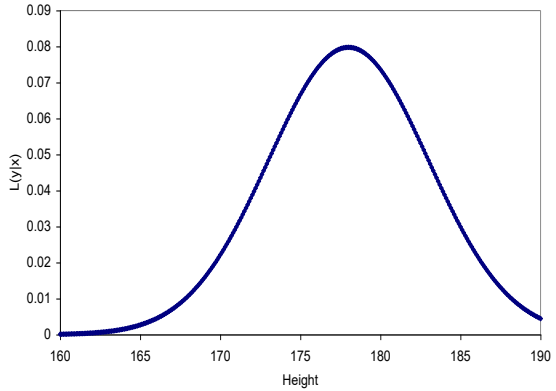
$$P(x \mid y) \propto P(y \mid x)P(x)$$

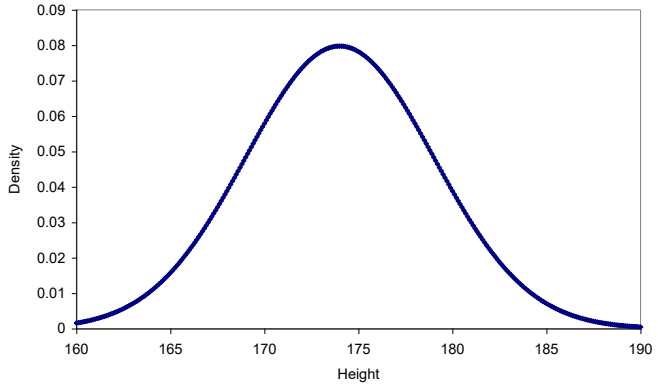
P(x|y) mean = 176cm

L(y|x) $\bar{y} = 178$, $s.e. = 5$

P(x) $\bar{x} = 174$, $s.e. = 5$







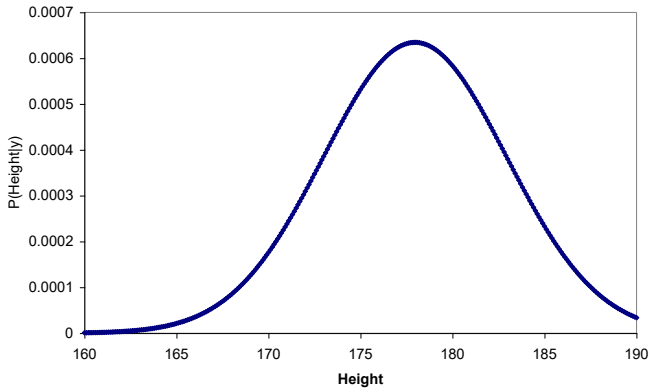
Less certainty about prior information? Use *less* informative (flat) prior

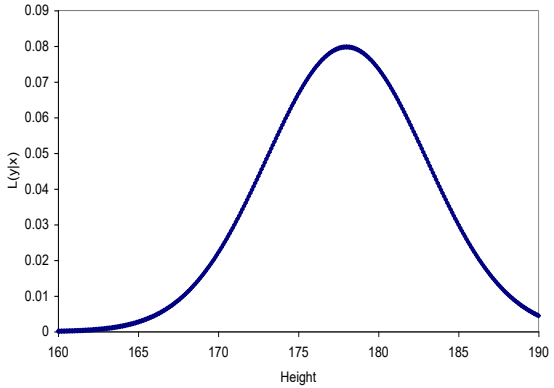
$$P(x \mid y) \propto P(y \mid x)P(x)$$

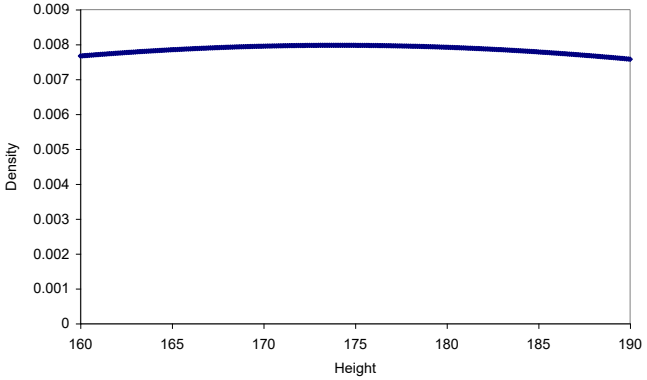
P(x|y) mean = 178cm

L(y|x) $\bar{y} = 178$, $s.e. = 5$

P(x)







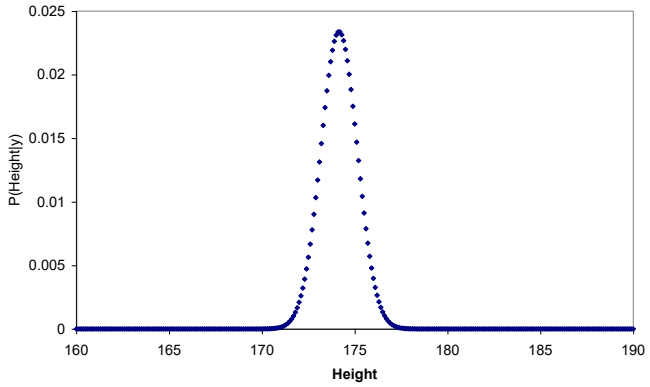
More certainty about prior information? Use *more* informative prior

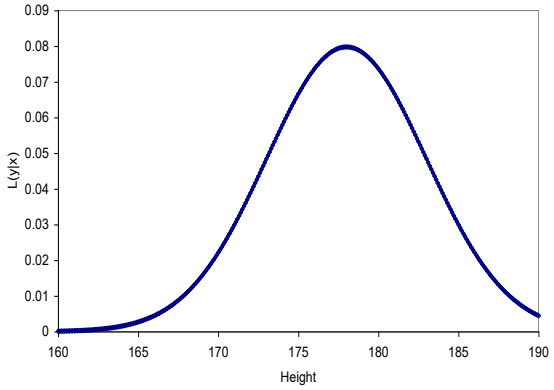
$$P(x \mid y) \propto P(y \mid x)P(x)$$

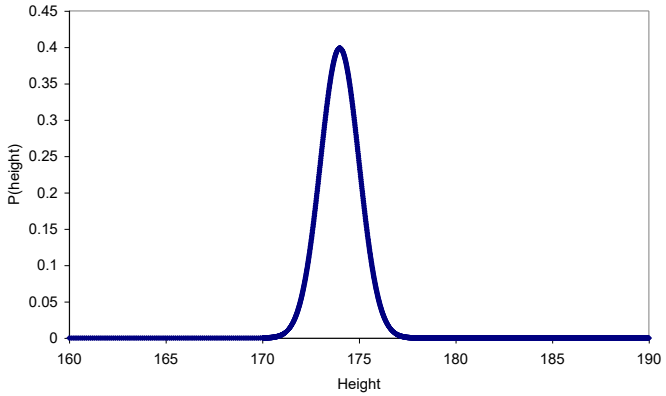
P(x|y) mean = 174.5cm

L(y|x) $\bar{y} = 178$, $s.e. = 5$

P(x)



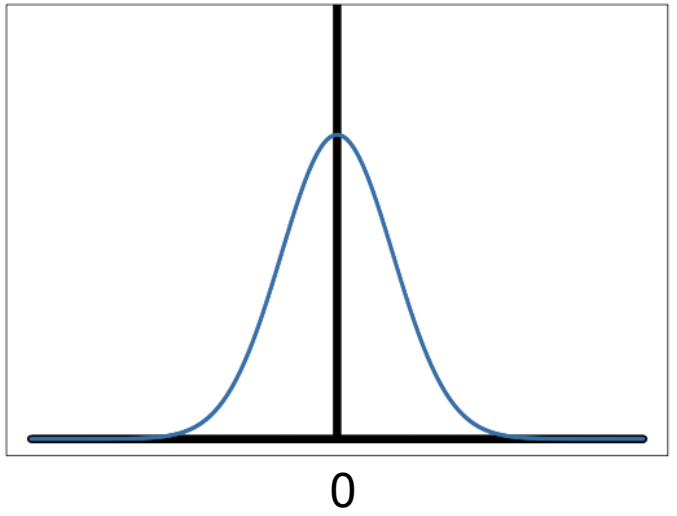




BayesC

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

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



Using Bayes theorem, the posterior distribution of SNP effects

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

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

Posterior inference on SNP effects

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

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

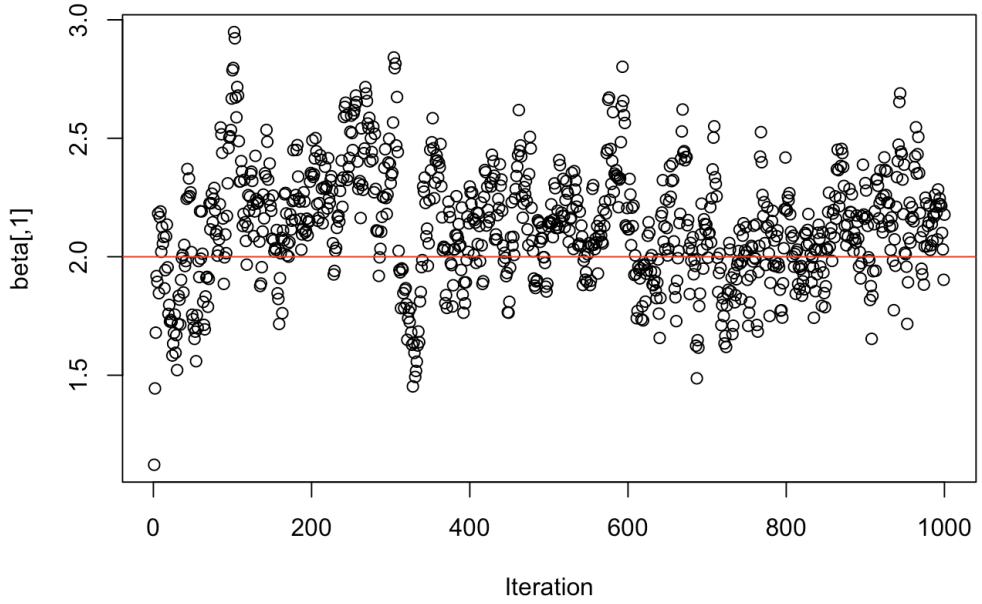
MCMC sampling

- 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 $\boldsymbol{\beta}$

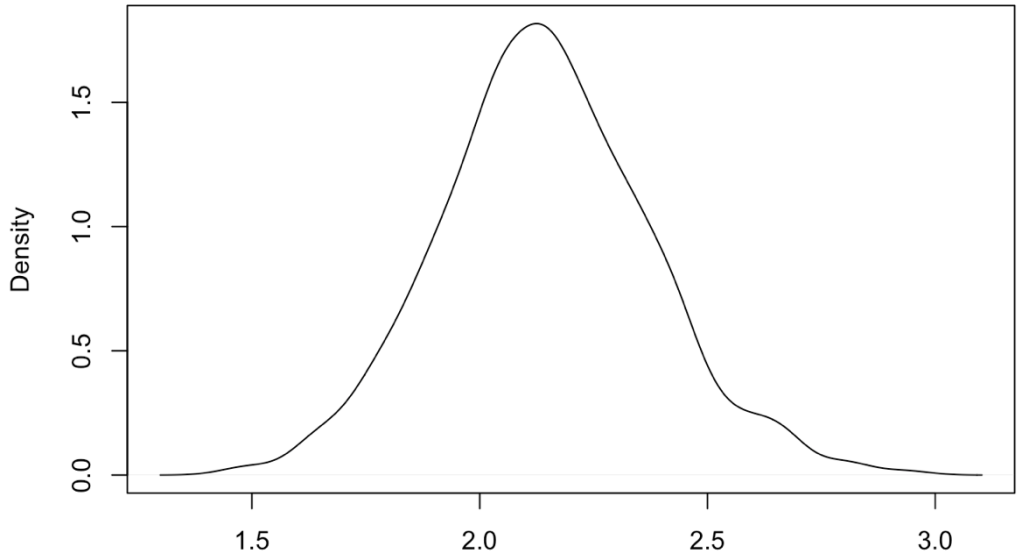
Samples reconstruct posterior distributions of parameters

MCMC sampling

Trace plot



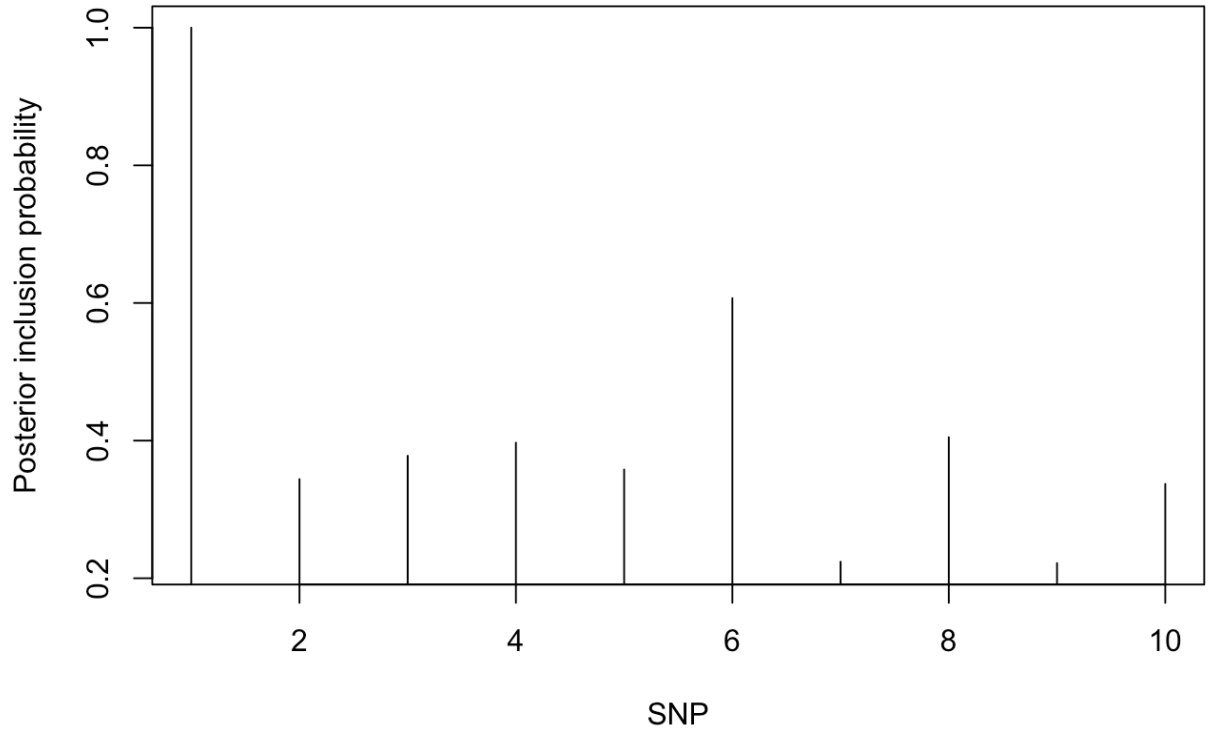
Posterior distribution



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

As a method of fine-mapping

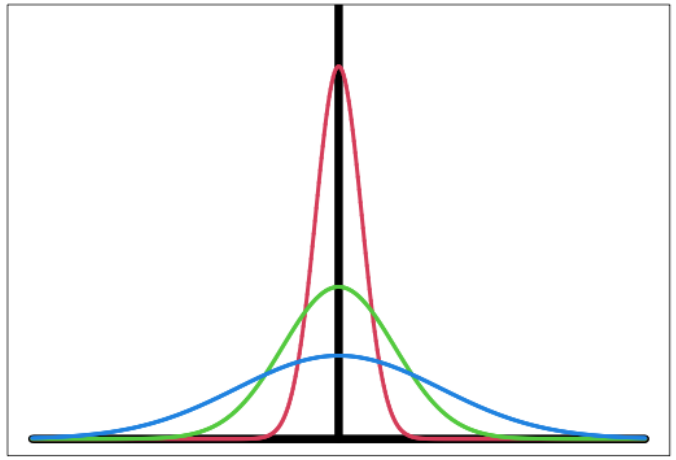
Posterior inclusion probability (PIP):
probability that the SNP is included in the model with a nonzero effect.



BayesR

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

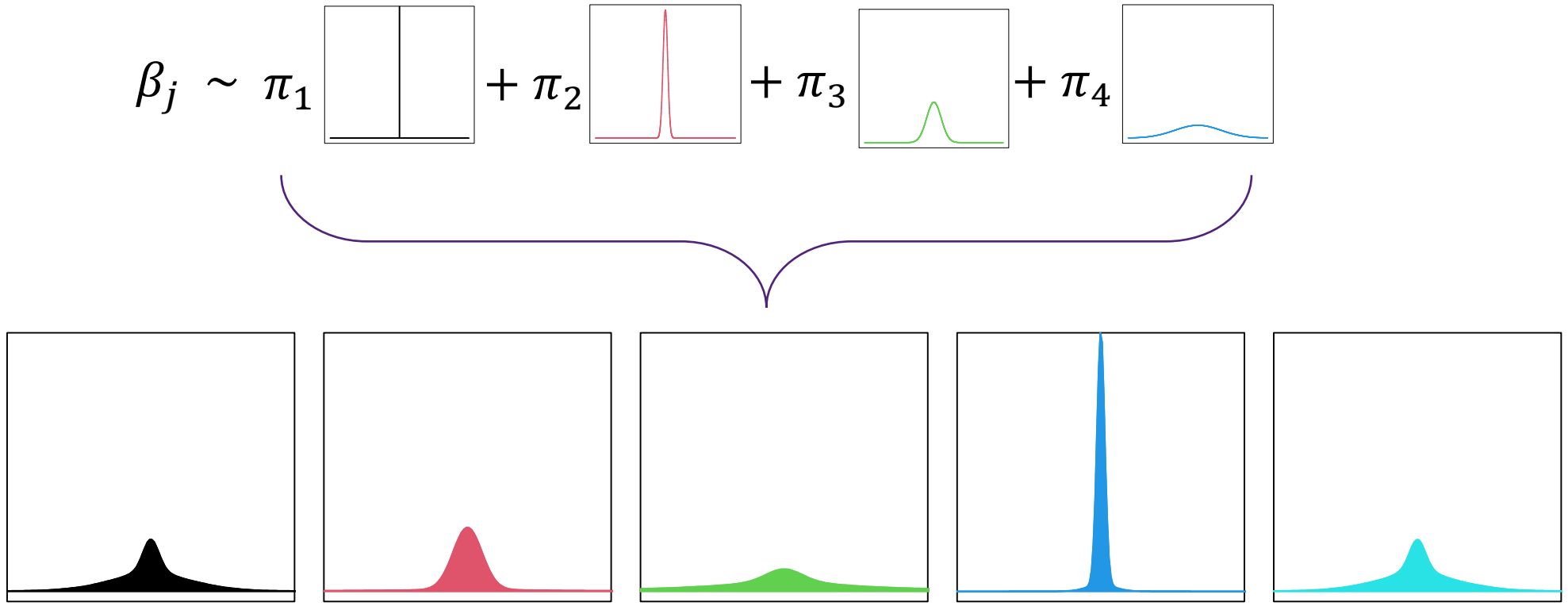
$$\beta_j|\pi, \sigma^2_\beta = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0, \gamma_2\sigma^2_\beta) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C\sigma^2_\beta) & \text{with probability } 1 - \sum_{c=1}^{C-1} \pi_c, \end{cases}$$



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

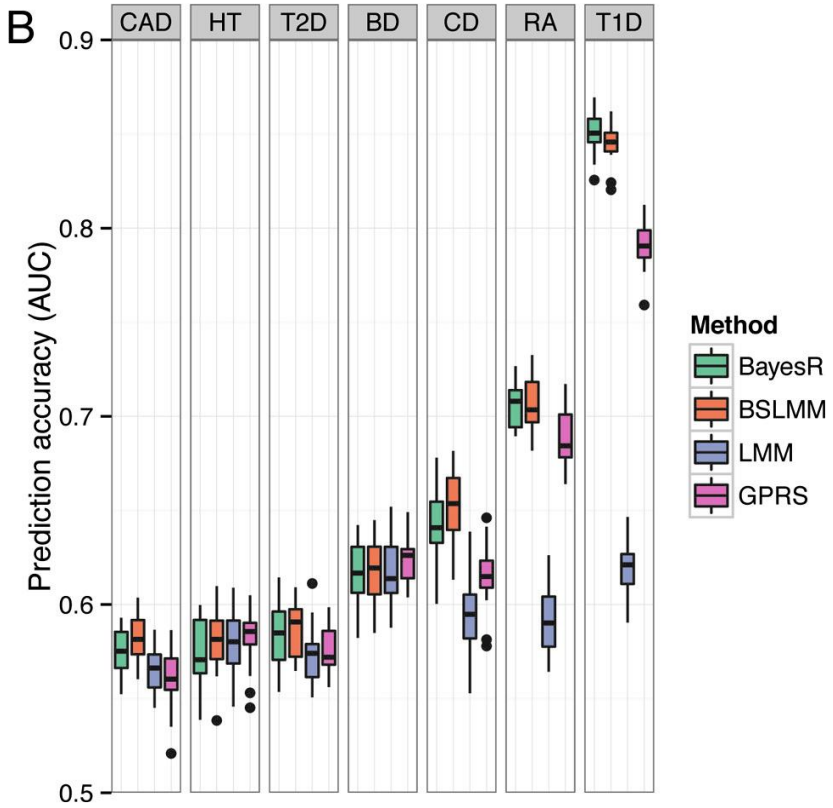
BayesC is a special case of BayesR with two components

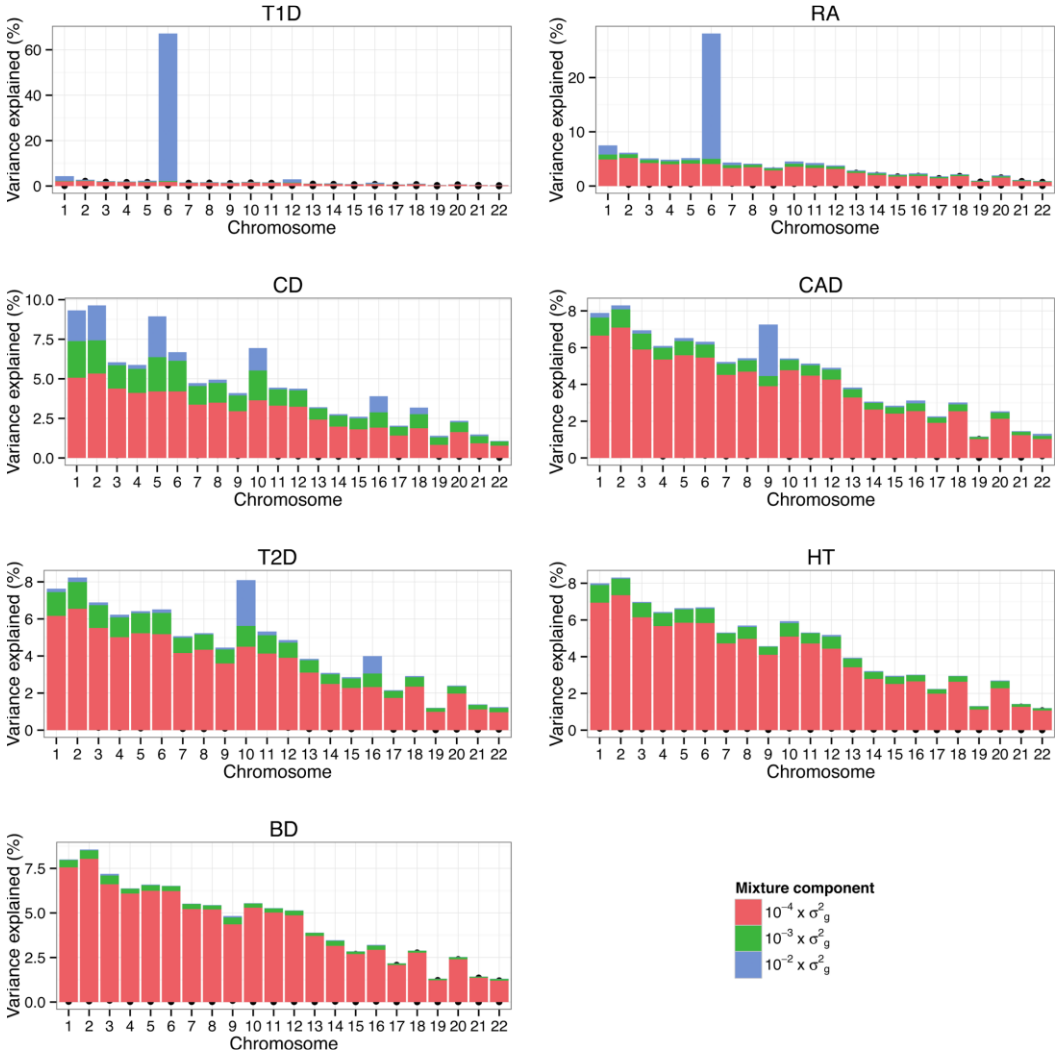
Why use multi-normal mixture?



Account for almost any distribution!

Prediction of disease risk using BayesR





Moser et al PLoS Genetics 2015

Methodology based on summary statistics

- Individual-level data are often not publicly accessible.
- Computationally demanding with large # individuals/SNPs.
- Methodology in human genetics has moved forward to use GWAS sumstats only.

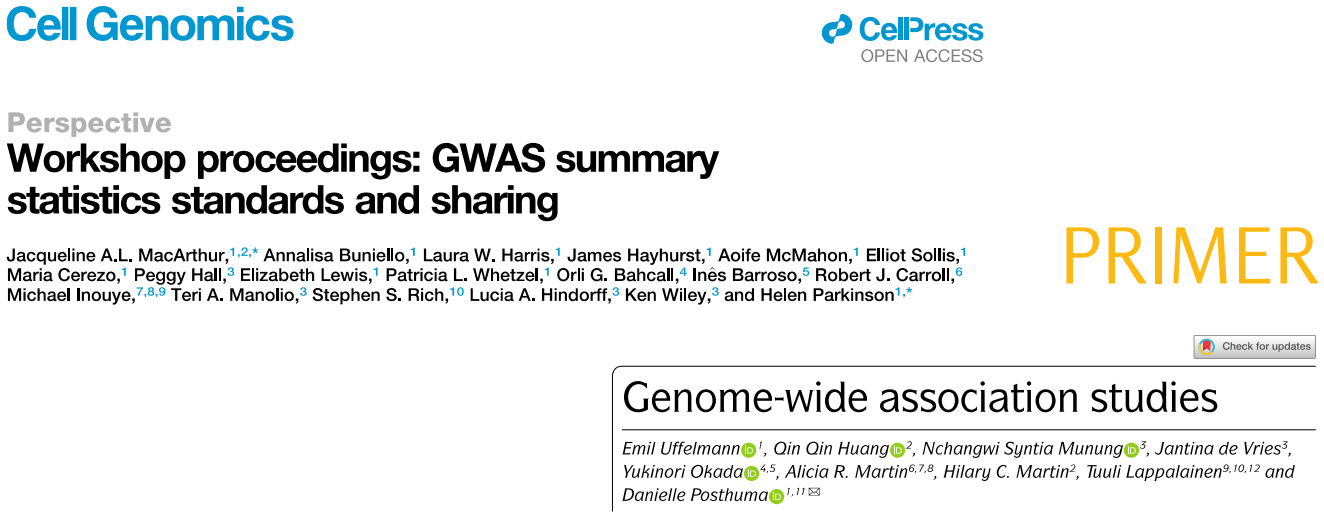


Table 3 | **Databases of GWAS summary statistics**

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

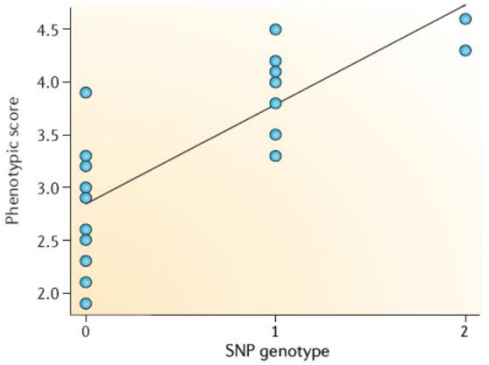
For a comprehensive list of genetic data resources, see REF.[13]. GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism.

What are the minimum data required?

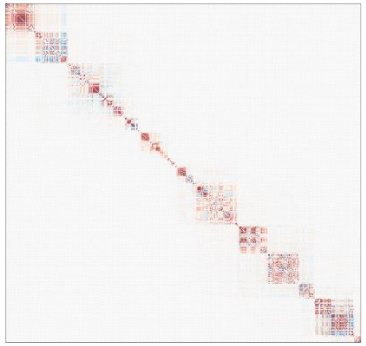
Essentially, the sumstats-based methods take GWAS marginal effect estimates as input data, and **re-estimate** SNP **joint effects**.

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

} GWAS sumstats

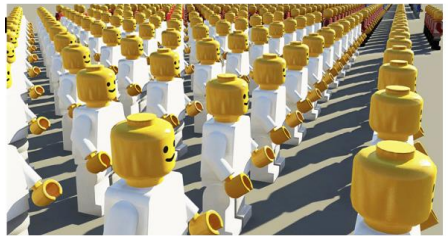


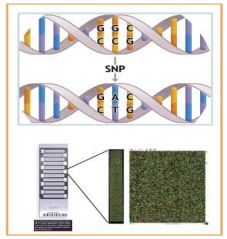
- LD correlations among SNPs → LD matrix



Individual-level model

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



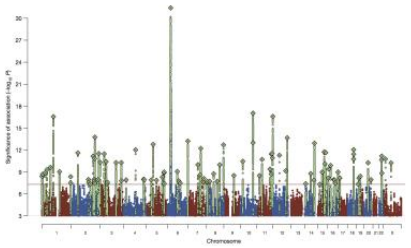


BLUP

Bayes

Summary-level model

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



SBLUP

SBayes

Covariates, such as age and sex, are accounted for when running GWAS.

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

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

SBayes

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

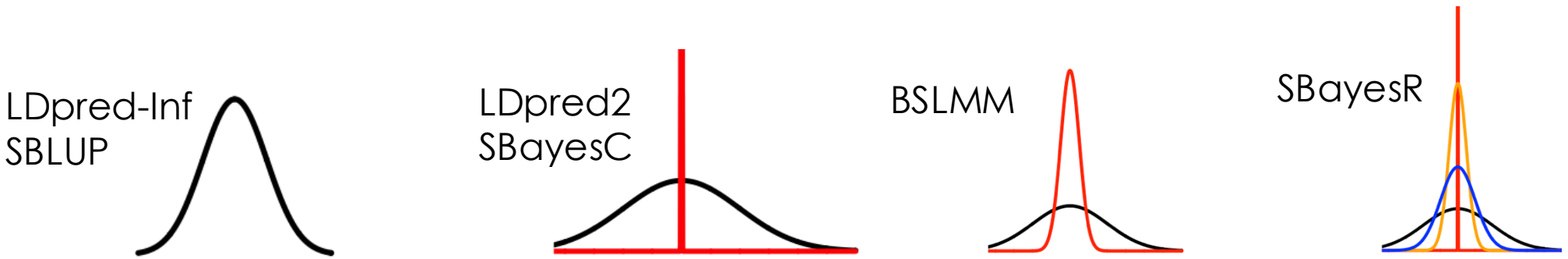
$$Var(\boldsymbol{\epsilon}) = \frac{1}{n}\mathbf{R}\sigma_e^2$$

SNP marginal effects from GWAS

LD correlation matrix

SNP joint effects

Prior distribution for each SNP effect

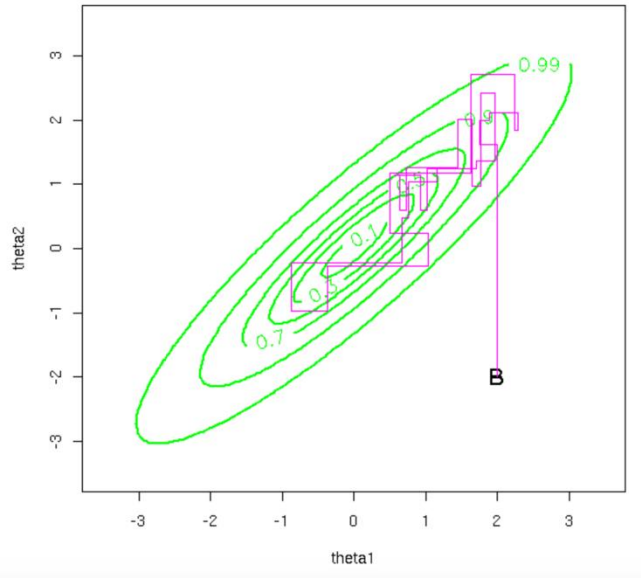


MCMC sampling

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

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

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

Potential issue

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

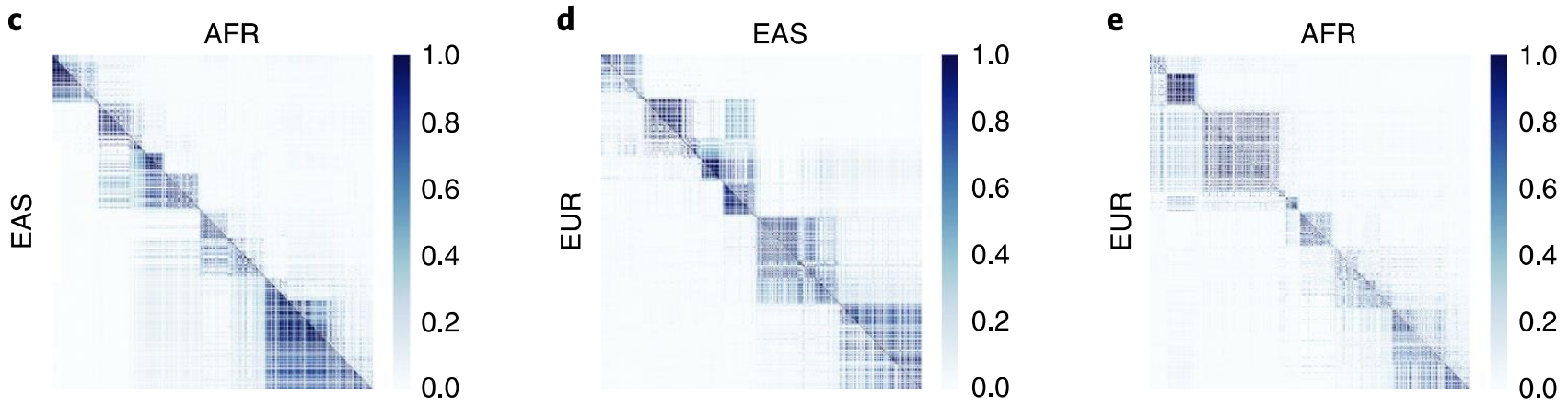
- In practice, SBayes is approximation to Bayes when LD is estimated from a reference sample.

- 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

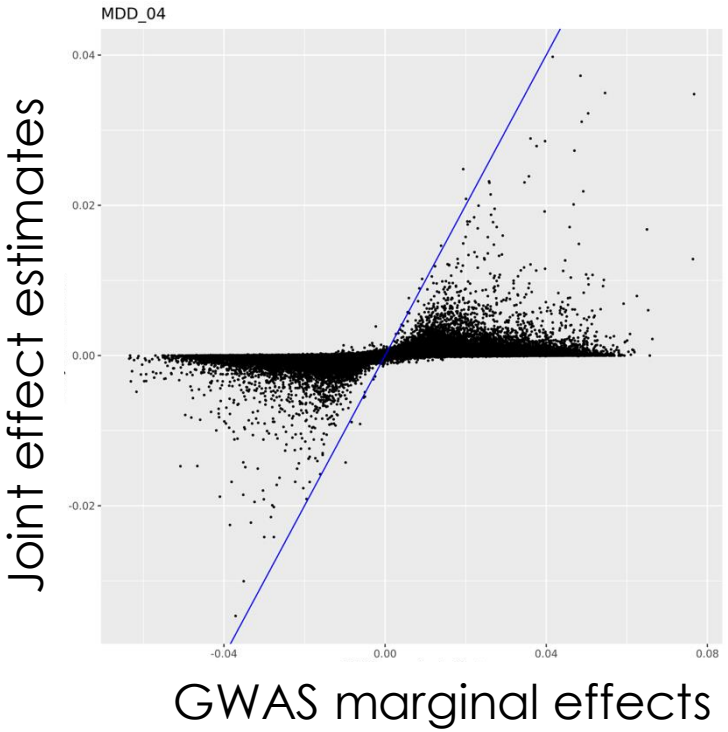


Failure to meet this assumption can result in a convergence issue!

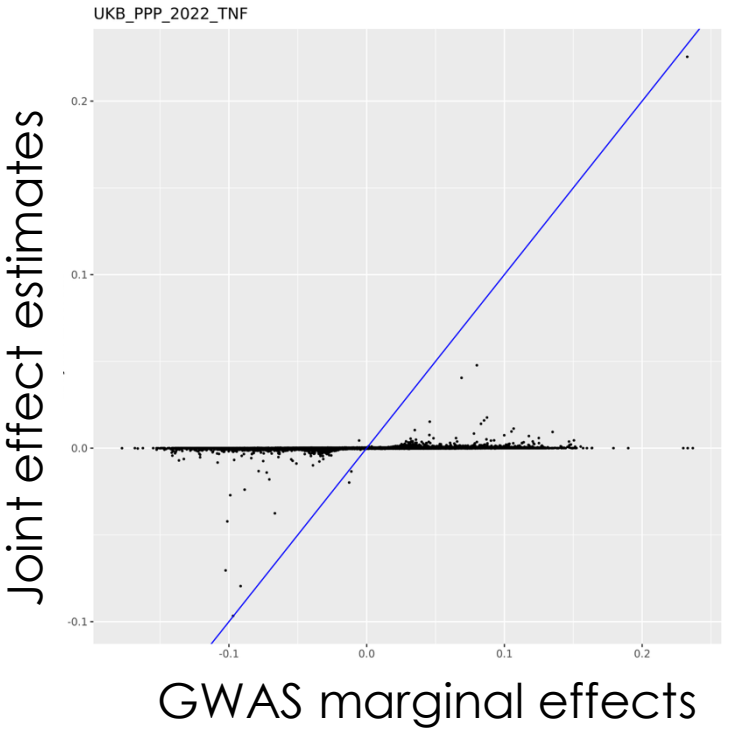
Always good to check SNP effect estimates

Estimated joint effect size vs. GWAS marginal effect size

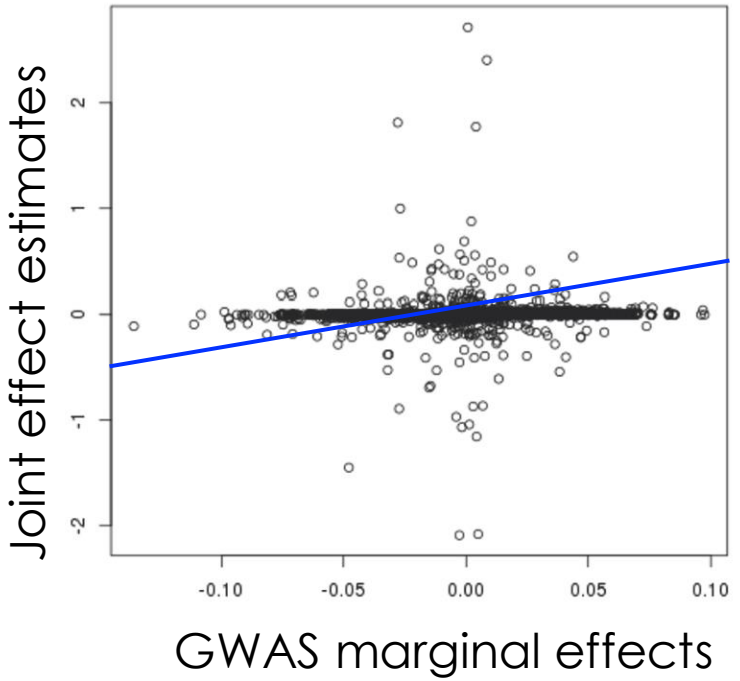
Most common



Presence of large effects



Bad convergence!

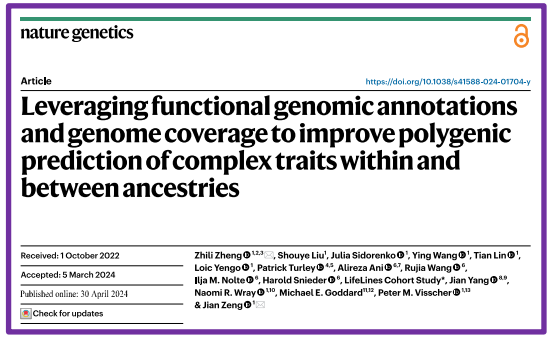


How do different methods handle this issue?

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

- Force a strong minimum shrinkage to SNP effects
  - e.g., PRS-CS

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

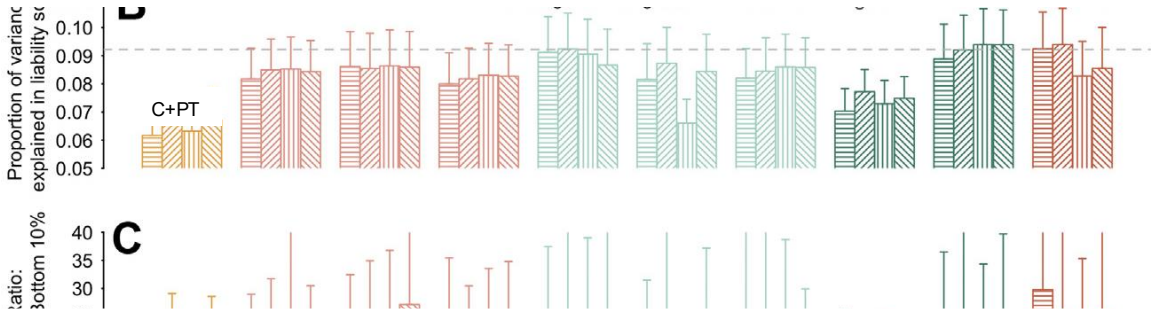


Archival Report

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 (C+PT)
- Mixture models > non-mixture (infinitesimal) models

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	$f(\beta) = \mathbf{y}^T\mathbf{y} + (1 - s)\beta^T\mathbf{X}_r^T\mathbf{X}_r\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + s\beta^T\beta + 2\lambda\|\beta\|_1^1$ $\mathbf{X}_r$: $n \times m$ 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	$\beta_j \mid \pi, \sigma_\beta^2 \sim \begin{cases} 0, & \text{with probability of } \pi_1 \\ N(0, \gamma_2\sigma_\beta^2), & \text{with probability of } \pi_2 \\ \vdots \\ N(0, \gamma_c\sigma_\beta^2), & \text{with probability of } 1 - \sum_{c=1}^{C-1}\pi_c \end{cases}$ $\sigma_\beta^2 \sim Inv - \chi^2\ (d.f. = 4)$ $\pi_i \sim Dir(\mathbf{1})$, 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

Summary

- Bayesian approach allows us to incorporate prior knowledge in estimation of SNP effects.
- Markov chain Monte Carlo (MCMC) is a technique to draw samples from a posterior distribution for Bayesian inference.
- Bayesian methods can have an advantage when:
  - Causal variants of moderate to large effect on the trait (e.g. T1D)
  - Very large numbers of SNP → set some SNP effects to zero
- Sumstats-based methods scale to large GWAS but rely on accurate LD.
- These methods represent the current state of the art in polygenic prediction.

Recommended reading

1. Meuwissen THE, *et al*. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, Volume 157, Issue 4, 1 April 2001, Pages 1819–1829. (**Founding paper of Bayesian methods for genomic prediction; BayesA and BayesB**)

2. Habier D, *et al*. Extension of the bayesian alphabet for genomic selection. BMC *Bioinformatics*. 2011 May 23;12:186. (**BayesC**)

3. Moser G, *et al*. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* 2015 Apr 7;11(4):e1004969. (**BayesR**)

4. Gianola D. Priors in whole-genome regression: the bayesian alphabet returns. *Genetics*. 2013 Jul;194(3):573-96. (**Review of Bayesian alphabet models**)

5. Lloyd-Jones LR, *et al*. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun* 10, 5086 (2019). (**SBayesR; MCMC details in Supplementary**)

6. Ni G, *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol. Psychiatry* 90, 611–620 (2021). (**Overview of methods**)