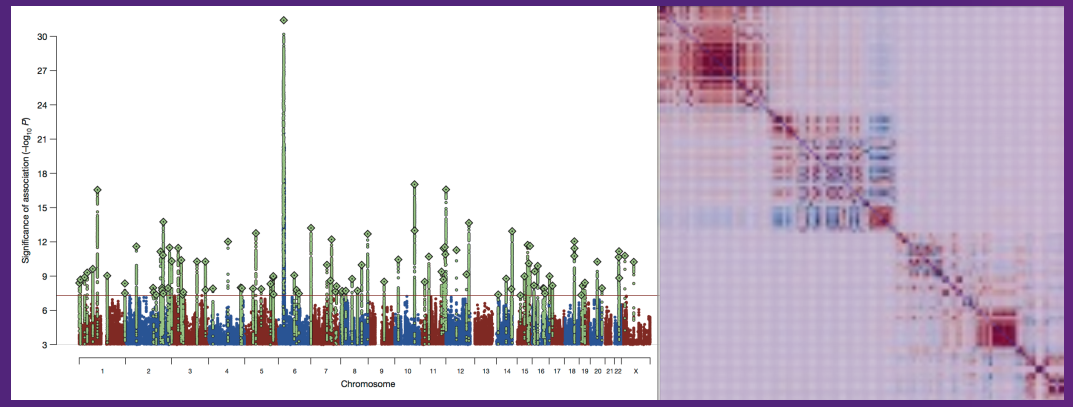
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE



Polygenic Prediction using Summary data

Jian Zeng

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
Institute for Molecular Bioscience
Program in Complex Trait Genomics

Polygenic prediction

- Best prediction methods take genetic values as random effect (e.g., BLUP and BayesR).
- These methods require individual-level genotype and phenotype data.
- Data are not publicly accessible, due to privacy and ethical considerations.
- Computationally demanding when numbers of individuals and SNPs are large.
- Use of GWAS summary-level data can address both problems.
- Methodology in human genetics has moved forward to use GWAS summary-level data.

Consensus of sharing GWAS summary data (in human genetics research community)

Has Become a standard to share and make publicly available the summary-level data when publishing a GWAS study.

Asking for more

Because of the usefulness of genome-wide association study (GWAS) data for mapping regulatory variation in the human genome, the journal now asks authors to report the co-location of trait-associated variants with gene regulatory elements identified by epigenetic, functional and conservation criteria. We also ask that authors publish or database the genotype frequencies or association *P* values for all SNPs investigated, whether or not they reached genome-wide significance.

—Nat Genet editorial, July 2012

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 do we need to perform summary-data-based polygenic prediction?

For simplicity, let's assume the genotypes of each SNP has been standardised with column mean zero and variance one when conducting GWAS.

In this case, the minimum data required are

- SNP marginal effect estimates
- GWAS sample size
- LD correlations among SNPs

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.

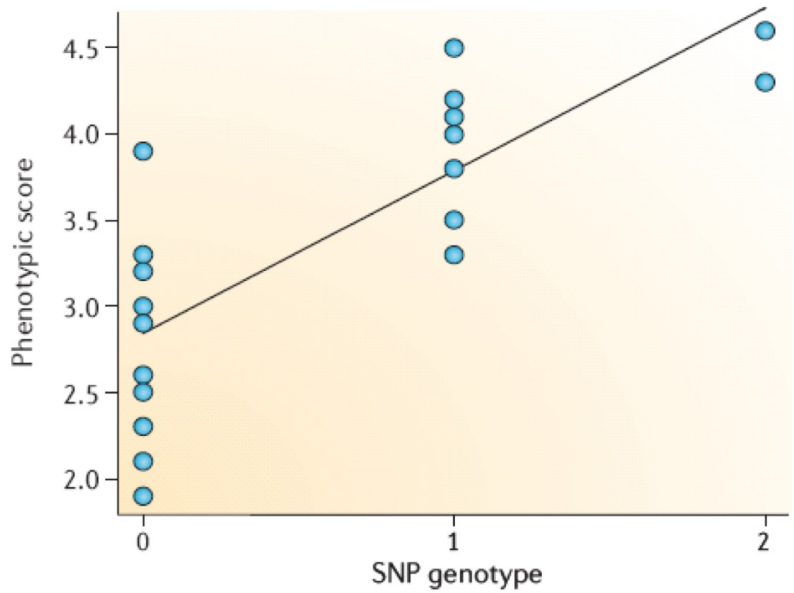
$$b_j = \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\mathbf{y}$$

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.

SNP marginal effect estimates

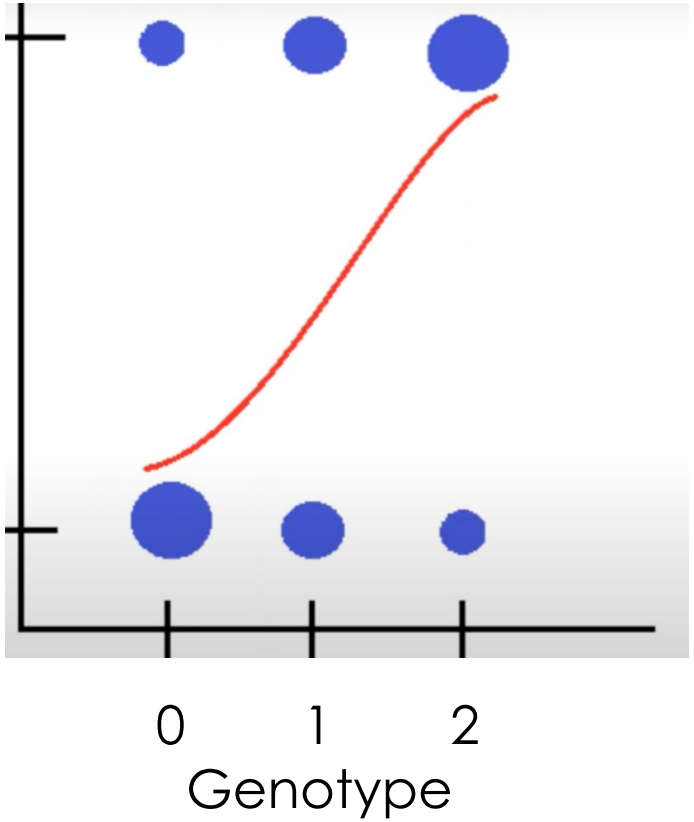
For diseases, GWAS is done using logistic regression

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + X_{ij}b_j$$

The SNP effect is log odds ratio (OR), i.e., difference in log odds for cases vs. controls

$$b_j = \log(OR)$$

Approximately equal to the b_j from the linear model when true effect size is small.



What do we need to perform summary-data-based polygenic prediction?

For simplicity, let's assume the genotypes of each SNP has been standardised with column mean zero and variance one when conducting GWAS.

Then, the minimum data required are

- SNP marginal effect estimates
- GWAS sample size
- **LD correlations among SNPs**

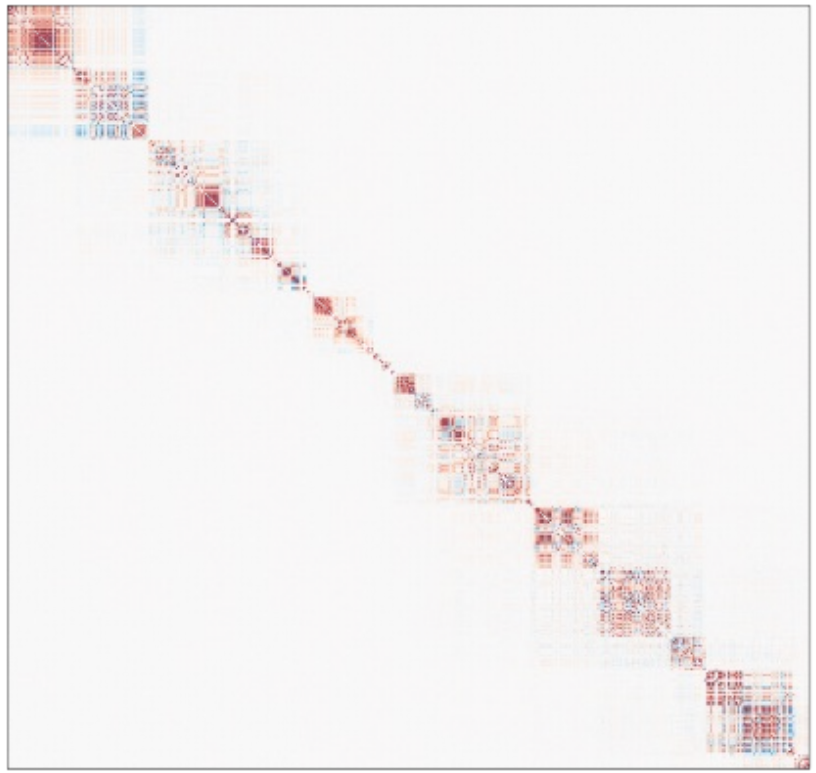
Linkage disequilibrium (LD) correlations

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



Use of summary data only - how does it work?

GWAS results and LD correlations are **sufficient statistics** for the estimation of SNP joint effects!

A statistic is **sufficient** if no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter.

e.g., $x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$ and we want to estimate μ and σ^2

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- $\sum_{i=1}^{n} x_i$ and n are sufficient statistics for μ

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left[\frac{\sum_{i=1}^{n} x_i}{n}\right]^2$$

- $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} x_i$ and n are sufficient statistics for σ^2

We don't need to know the value of each x!

How does it work?

BLUP

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

BLUP solutions:

$$[\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda]\widehat{\boldsymbol{\beta}} = \mathbf{X}'\boldsymbol{y}$$

$n\,\mathbf{R}$ $\qquad$ $n\,\mathbf{b}$

where $\lambda = \frac{\sigma_e^2}{\sigma_\beta^2}$

Recall

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

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

R (LD matrix), **b** (marginal effects) and n are sufficient statistics for the estimation of $\boldsymbol{\beta}$.

From individual- to summary-level model

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

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

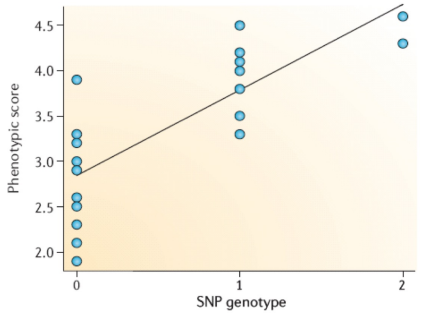
Multiply both sides by $\frac{1}{n}\mathbf{X}'$ gives

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

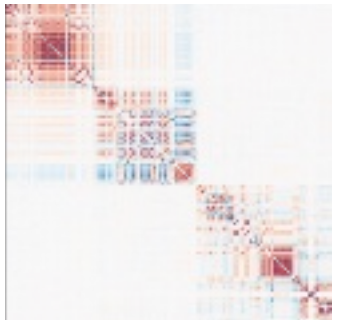
$$\mathbf{b} = \mathbf{R}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

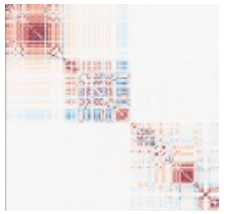
$Var(\boldsymbol{\epsilon}) = \frac{1}{n}\mathbf{R}\sigma_e^2$

GWAS marginal SNP effects



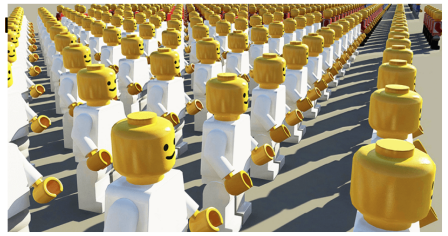
LD correlation matrix

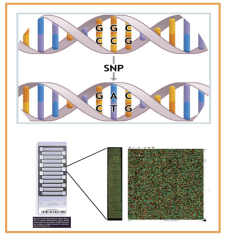




Individual-level data analysis

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



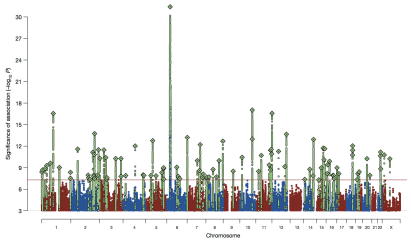


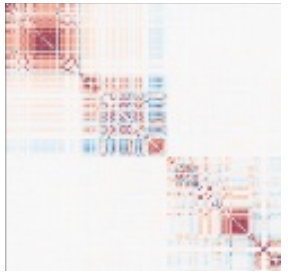
BLUP

Bayes

Summary-level data analysis

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





SBLUP

SBayes

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

BLUP vs. SBLUP

BLUP

- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$
- $Var(\mathbf{e}) = \mathbf{I}\sigma_e^2$
- $[\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda]\widehat{\boldsymbol{\beta}} = \mathbf{X}'\boldsymbol{y}$
- $\widehat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda]^{-1}\mathbf{X}'\boldsymbol{y}$

Genotype matrix

Phenotypes

Individual-level data
Summary-level data

SBLUP

- $\mathbf{b} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- $Var(\boldsymbol{\epsilon}) = \frac{1}{n}\mathbf{R}\sigma_e^2$
- $[n\mathbf{R} + \mathbf{I}\lambda]\widehat{\boldsymbol{\beta}} = n\mathbf{b}$
- $\widehat{\boldsymbol{\beta}} = [n\mathbf{R} + \mathbf{I}\lambda]^{-1}n\mathbf{b}$

GWAS sample size

LD correlation matrix

Marginal SNP effects

SBayes

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

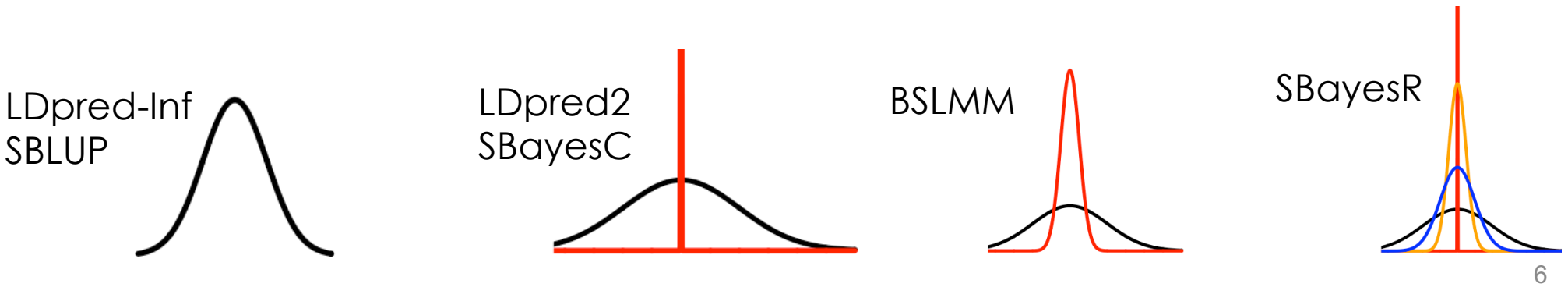
$$Var(\boldsymbol{\epsilon}) = \frac{1}{n}\mathbf{R}\sigma_e^2$$

GWAS marginal SNP effects

LD correlation matrix

SNP joint effects

Prior distribution for each SNP effect



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12653-0 OPEN

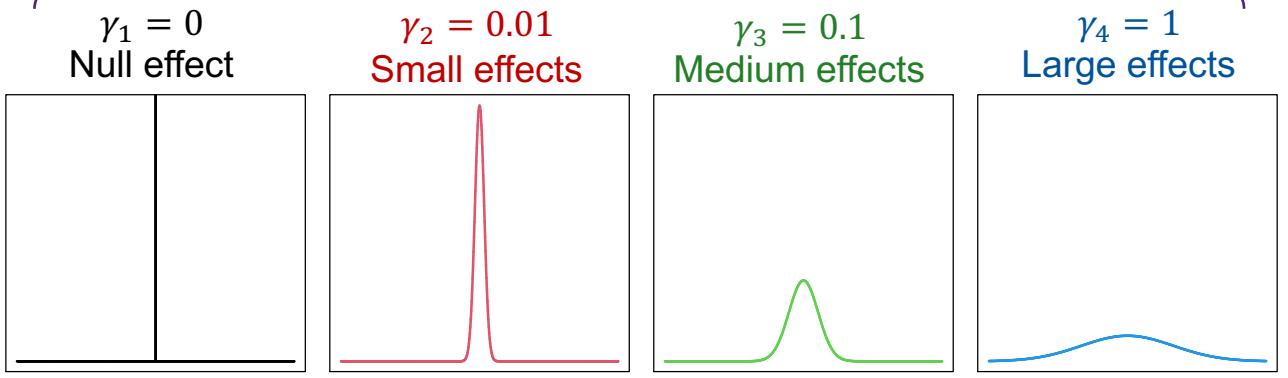
SBayesR

Improved polygenic prediction by Bayesian multiple regression on summary statistics

Luke R. Lloyd-Jones[1,9*], Jian Zeng[1,9*], Julia Sidorenko[1,2], Loïc Yengo[1], Gerhard Moser[3,4], Kathryn E. Kemper[1], Huanwei Wang[1], Zhili Zheng[1], Reedik Magi[2], Tõnu Esko[2], Andres Metspalu[2,5], Naomi R. Wray[1,6], Michael E. Goddard[7], Jian Yang[1,8*] & Peter M. Visscher[1*]

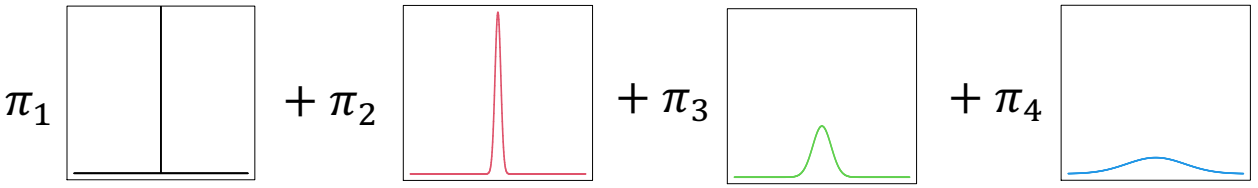
Each SNP effect has a mixture distribution:

$$\boldsymbol{\pi} \sim Dirichlet\,(\mathbf{1}) \qquad \beta_j \sim \sum_{k=1}^{4} \pi_k\, N\left(0, \gamma_k \sigma_\beta^2\right) \qquad \sigma_\beta^2 \sim \chi_\upsilon^{-1}(\upsilon, \tau^2)$$

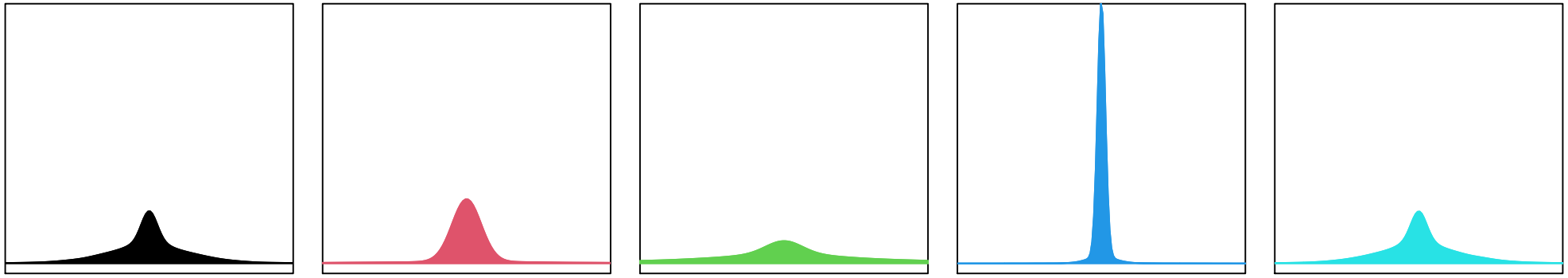


Account for various SNP effect distributions

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



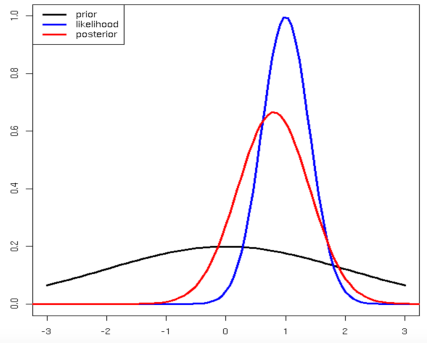
SNP 1 SNP 2 SNP 3 SNP 4 SNP 5



The posterior distribution of SNP effects

$$\textbf{Posterior} \propto \textbf{Likelihood} \times \textbf{Prior}$$

$$f(\boldsymbol{\beta}|\textbf{Summary data}) \propto f(\textbf{Summary data}|\boldsymbol{\beta}) \times f(\boldsymbol{\beta})$$



$$\boldsymbol{\beta}|\mathbf{b} \sim \mathrm{N}(\mathbf{C}^{-1}\mathbf{r}, \mathbf{C}^{-1}\sigma_e^2)$$

where

Individual-level data

$$\mathbf{r} = \mathbf{X}'\mathbf{y}$$

$$\mathbf{C} = \mathbf{X}'\mathbf{X} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_\beta^2}$$

$$\mathbf{G} = diag\{\gamma_j\}$$

Summary-level data

$$\mathbf{r} = n\mathbf{b}$$

$$\mathbf{C} = n\mathbf{R} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_\beta^2}$$

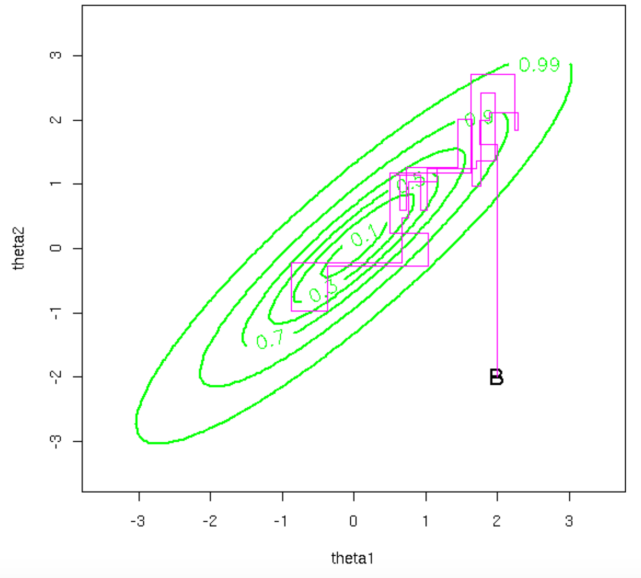
$$\mathbf{G} = diag\{\gamma_j\}$$

Single-site Gibbs sampling

Full conditional distribution for β_j

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

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

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

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

Comparison between individual and summary level algorithms

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 $l_{jc} = \mathbf{x}_j'\mathbf{x}_j + \frac{\sigma_\varepsilon^2}{\sigma_c^2}$ for each of the C classes
    Calculate the log densities of given $\delta_j = c$ using $\log(\mathcal{L}_c) = -\frac{1}{2}\left[\log\left(\frac{\sigma_c^2 l_{jc}}{\sigma_\varepsilon^2}\right) - \frac{r_j^2}{\sigma_\varepsilon^2 l_{jc}}\right] + \log(\pi_c)$, 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 $N\left(\frac{r_j}{l_{jc}}, \frac{\sigma_\varepsilon^2}{l_{jc}}\right)$
    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 $\widetilde{S}_\varepsilon^2 = \frac{SSE + \nu_\varepsilon S_\varepsilon^2}{n + \nu_\varepsilon}$ 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 $h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$
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 $l_{jc} = \mathbf{x}_j'\mathbf{x}_j + \frac{\sigma_e^2}{\sigma_c^2}$ for each of the C classes
    Calculate the log densities of given $\delta_j = c$ using $\log(\mathcal{L}_c) = -\frac{1}{2}\left[\log\left(\frac{\sigma_c^2 l_{jc}}{\sigma_e^2}\right) - \frac{r_j^2}{\sigma_e^2 l_{jc}}\right] + \log(\pi_c)$, 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 $N\left(\frac{\mathbf{r}_j}{l_{jc}}, \frac{\sigma_e^2}{l_{jc}}\right)$
    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 $h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$
od

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

Lloyd-Jones & Zeng et al. 2019 NC supplement

Relax the assumption of standardised genotypes

- Insofar, derivations are based on standardised genotypes.
- GWAS are usually performed using unstandardised genotypes (allele counts; X_j^*)
- b_j^* from the GWAS using X_j^* is called per-allele effect
- Need to rescale GWAS effects by $b_j = s_j b_j^*$ where s_j is the genotype SD
- Because

$$y = X_j^* b_j^* + e$$

$$= \frac{X_j^*}{s_j} \times s_j b_j^* + e$$

$$= X_j b_j + e$$

How to find s_j

We need more information and make assumptions!

Method 1:

- Use (minor) allele frequency p_j and assume Hardy-Weinberg Equilibrium (HWE)
- $s_j = \sqrt{2p_j(1-p_j)}$
- Cons: allele frequency from the GWAS sample may not be available

Method 2:

- Use GWAS effect standard error SE_j and assume b_j contribute to negligible variance
- $SE_j = \sqrt{\frac{1}{\mathbf{x}_j'\mathbf{x}_j}\hat{\sigma}_e^2} = \sqrt{\frac{1}{ns_j^2}\left(\sigma_y^2 - s_j^2 b_j^2\right)} \quad \rightarrow \quad s_j = \sqrt{\frac{\sigma_y^2}{nSE_j^2 + b_j^2}} \approx \sqrt{\frac{\sigma_y^2}{nSE_j^2}}$
- Cons: b_j may be large for a major QTL.

How to find s_j

In GCTB, we use

Method 3:

- Use p_j and SE_j and estimate σ_y^2
- σ_y^2 = the median value of $2p_j(1-p_j)\left[nSE_j^2 + b_j^2\right]$ across SNPs
- $s_j = \sqrt{\frac{\sigma_y^2}{nSE_j^2 + b_j^2}}$
- The median value of per-SNP σ_y^2 is robust to allele frequency errors.

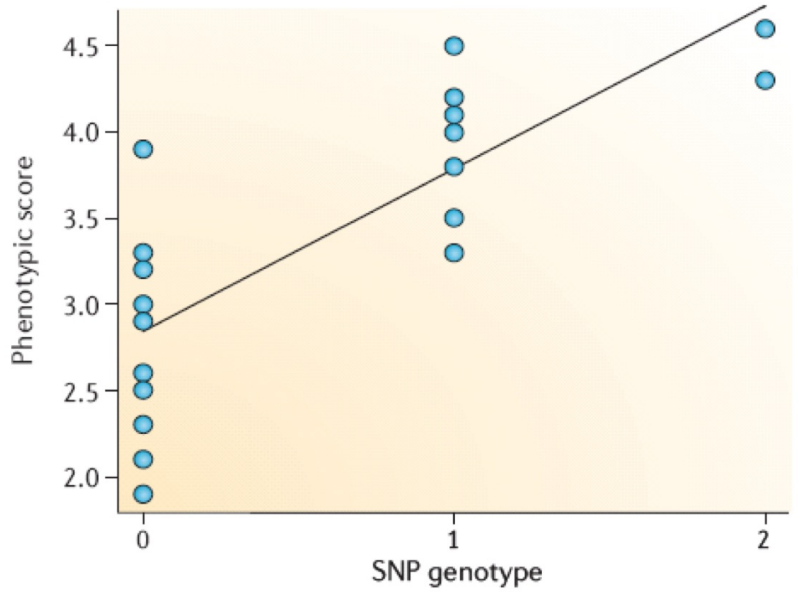
Critical information from GWAS summary data

- Marginal SNP effects
- GWAS sample size
- Standard errors
- Allele frequencies

Other information critical to quality control (QC)

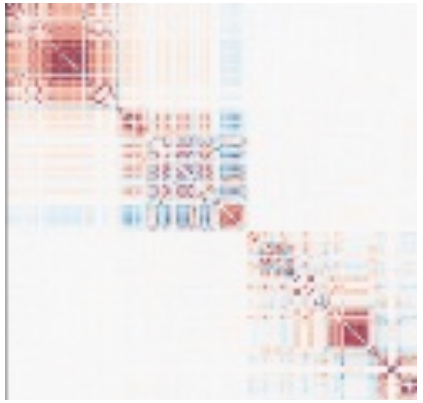
Which allele is the **effect allele** in GWAS?

e.g., A1 allele



A2A2 A1A2 A1A1

Need to match with the allele used to calculate the LD matrix in the reference sample



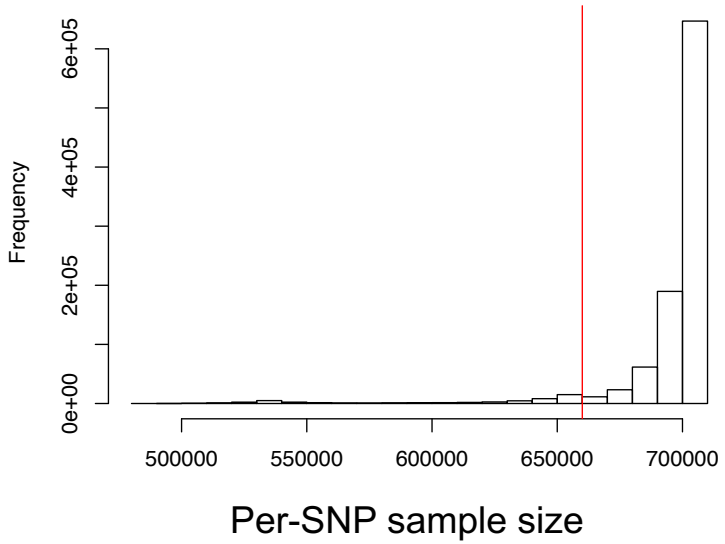
Other information critical to quality control (QC)

Per-SNP sample size

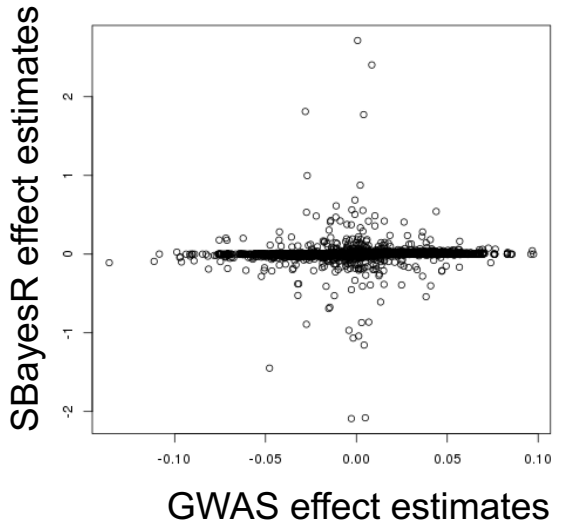
Heterogeneity in per-SNP sample size (usually due to meta-analysis) may result in a convergence problem in MCMC.

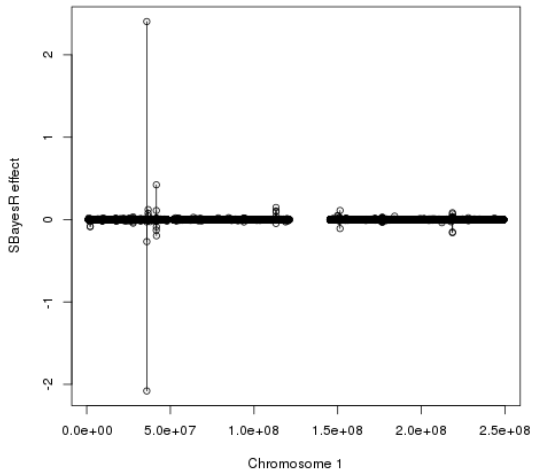
We recommend to visualise the per-SNP sample size distribution and remove the outliers.

Influence of heterogeneity in per-SNP sample size

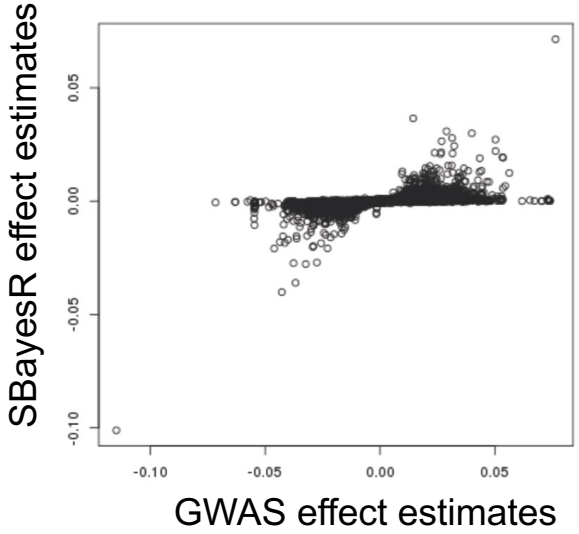


Abnormal





Normal



https://cnsgenomics.com/software/gctb/#FAQ

Critical information from GWAS summary data

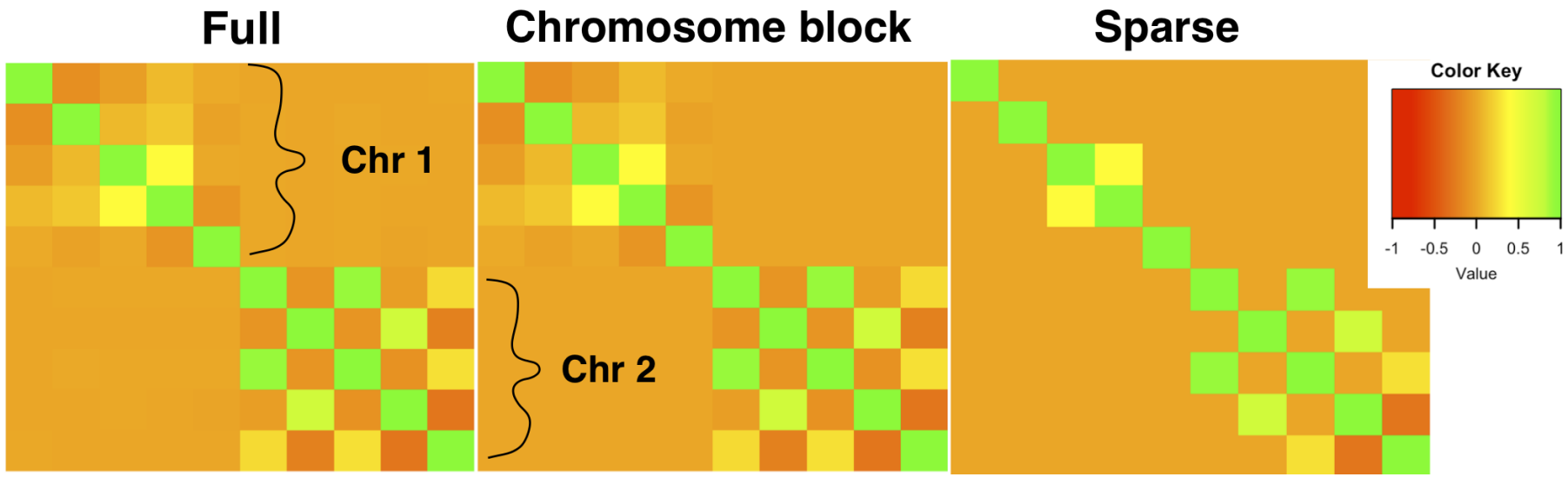
- Marginal SNP effects
- (Per-SNP) GWAS sample sizes
- Standard errors
- Effect alleles and alternate alleles (A1 and A2)
- Effect allele frequencies

Input file (.ma)

```
SNP A1 A2 freq b se p N
rs1001 A G 0.8493 0.0024 0.0055 0.6653 129850
rs1002 C G 0.0306 0.0034 0.0115 0.7659 129799
rs1003 A C 0.5128 0.0045 0.0038 0.2319 129830
```

LD matrix from a reference sample

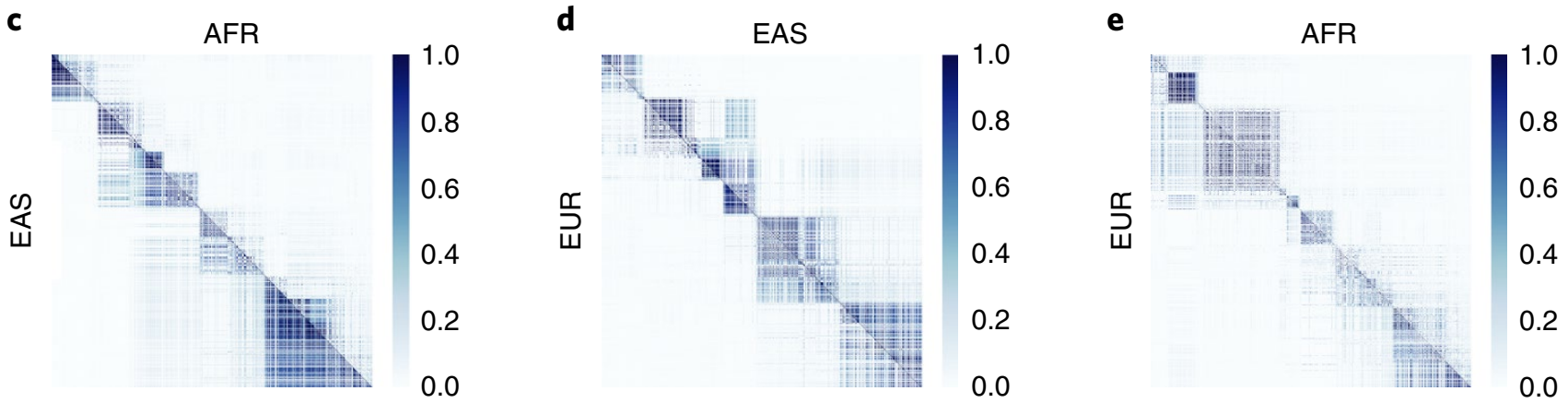
- Often cannot use genome-wide full LD matrix from the GWAS sample.
- Use a reduced (banded, sparse, or shrunk) LD matrix from a reference sample.



Implicit assumptions

LD reference population matches with GWAS population in genetics

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



Implicit assumptions

LD reference population matches with GWAS population in genetics

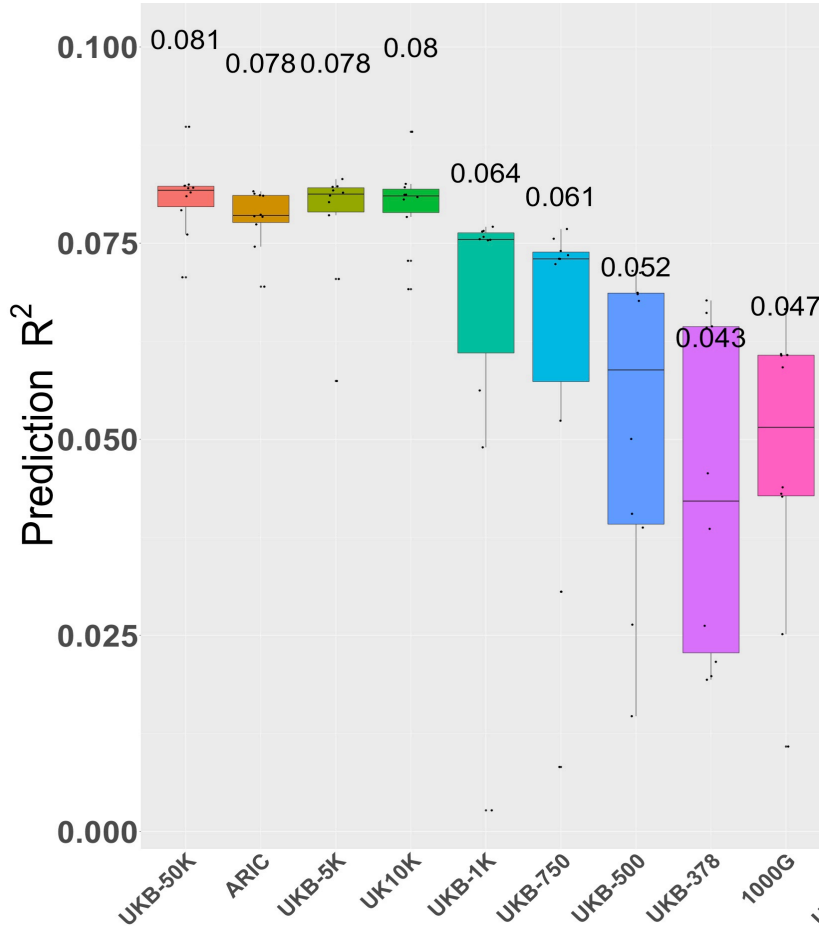
- No systematic differences in LD → same ancestry and population structure
- Minimum sampling variance in LD → LD ref sample size cannot be too small

GWAS data are collected on the same set of individuals

- Often an issue in GWAS meta-analysis
- Consistent genotyping platforms or imputation panels across cohorts
- Remove SNP outliers in per-SNP sample size

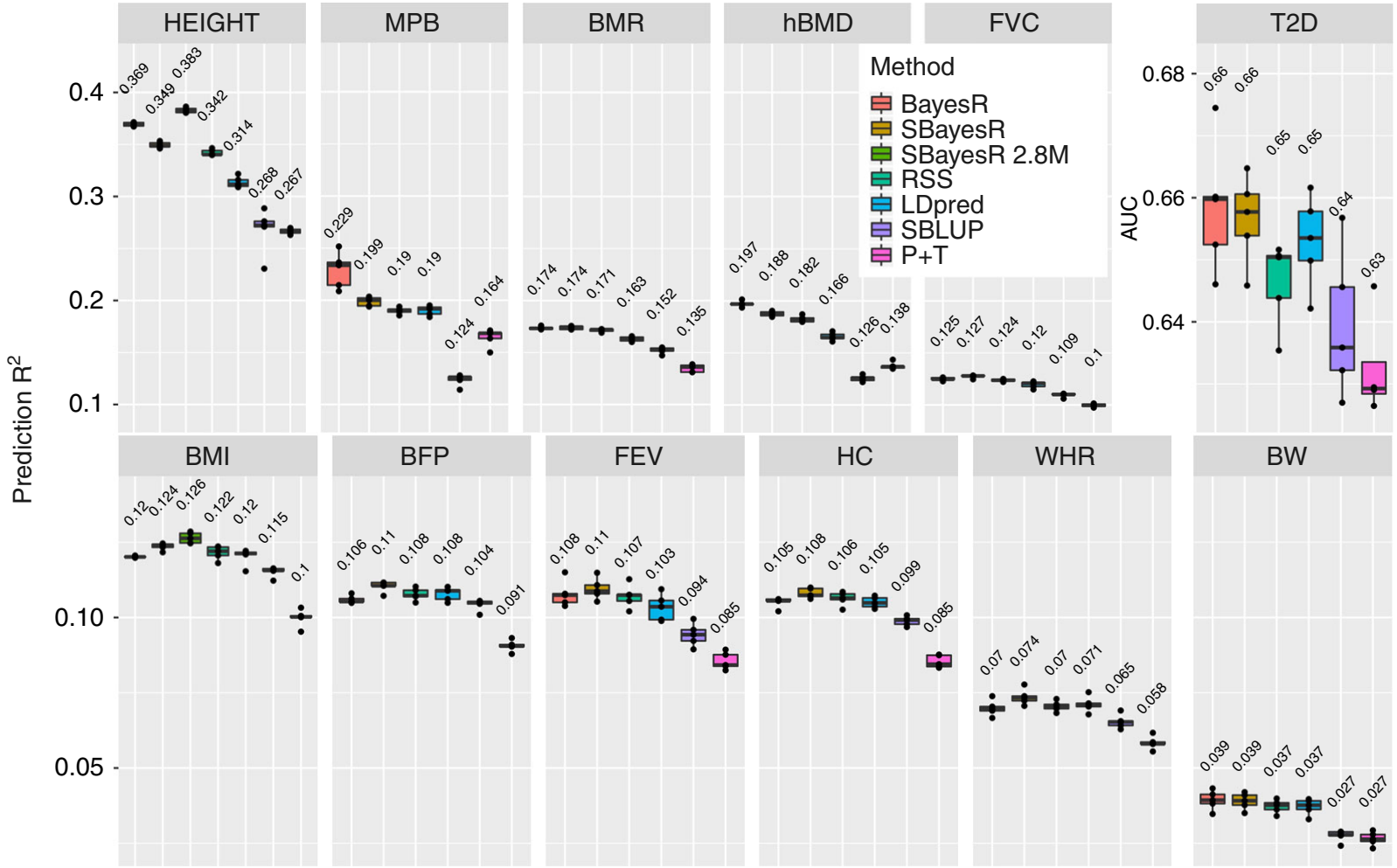
Violation these assumptions can cause model misspecification, resulting in attenuated prediction accuracy or even failure to reach convergence.

Influence of choice of LD reference



Lloyd-Jones & Zeng et al. 2019 NC

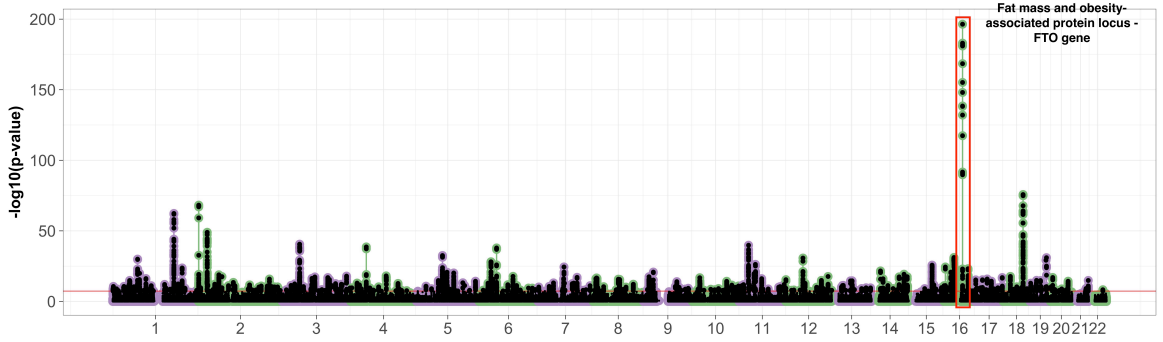
Method comparison



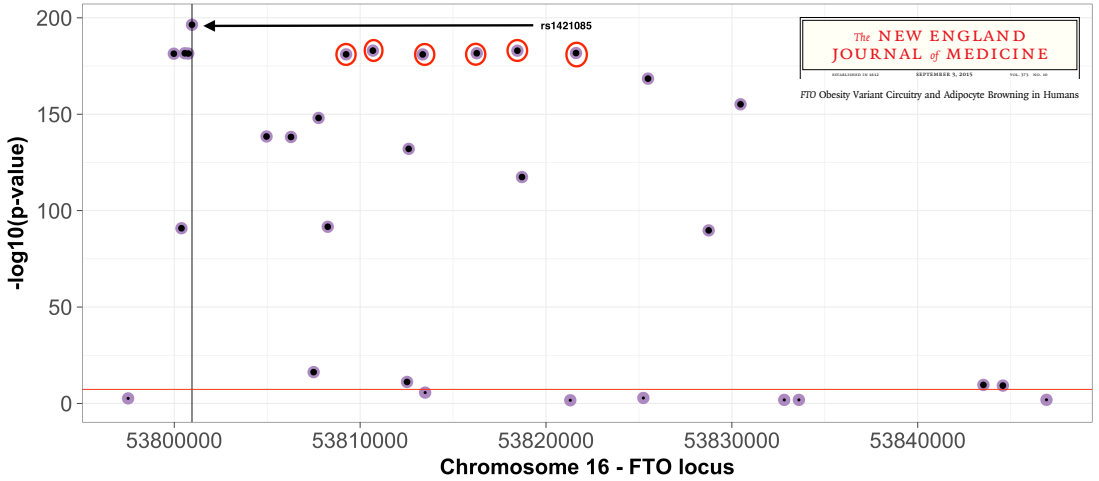
Lloyd-Jones & Zeng et al. 2019 NC

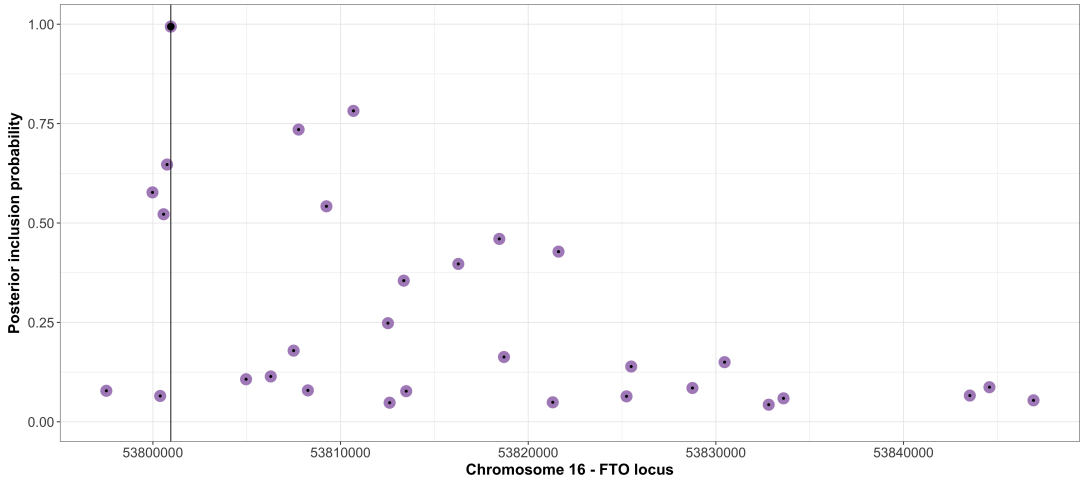
Fine-mapping

Real data - body mass index



Real data - FTO - GWAS



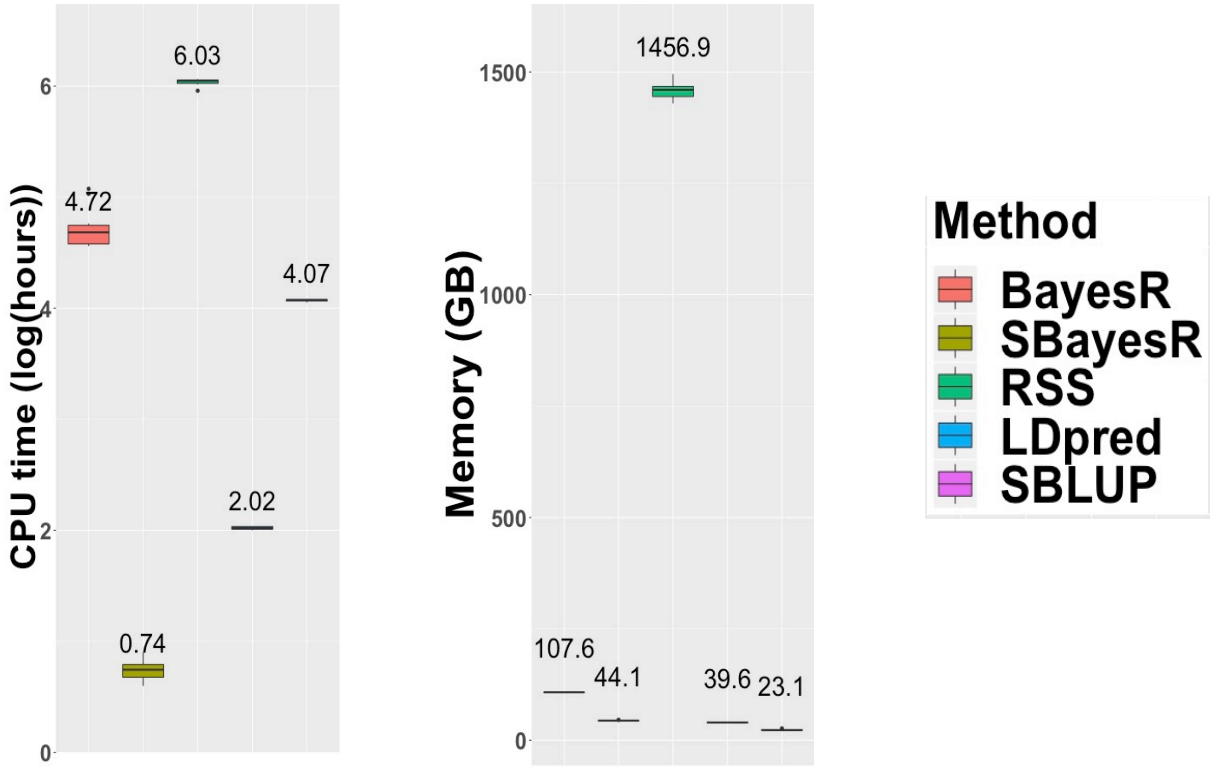


Computational efficiency

100k individuals

1M SNPs

Resource consumption for summary-data-based methods is independent of sample size once GWAS summary statistics are obtained.



Summary

- Summary-level methods unleash the full power of GWAS of large sample sizes for polygenic prediction.
- Free from limitation of data accessibility.
- Computationally efficient.
- Only an approximation to the individual-level counterpart due to reduction in LD matrix.
- Flexible to incorporate other information, such as functional annotations or omics data.

Practical 5: Polygenic prediction using summary data

https://cnsgenomics.com/data/teaching/GNGWS23/model5/Practical5_SBLUP_and_SBayes.html

Log into the cluster

cd to your working directory in scratch: cd /scratch/[your folder]

You will learn how to run MCMC with summary data using the toy example data set in R.

You will use GCTB to run SBayesR in the simulated data set based on real genotypes. Now we are able to leverage information from the full UK Biobank data without accessing to the individual genotypes and phenotypes.