THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

PGS Prediction using GWAS summary statistics

Jian Zeng

j.zeng@uq.edu.au

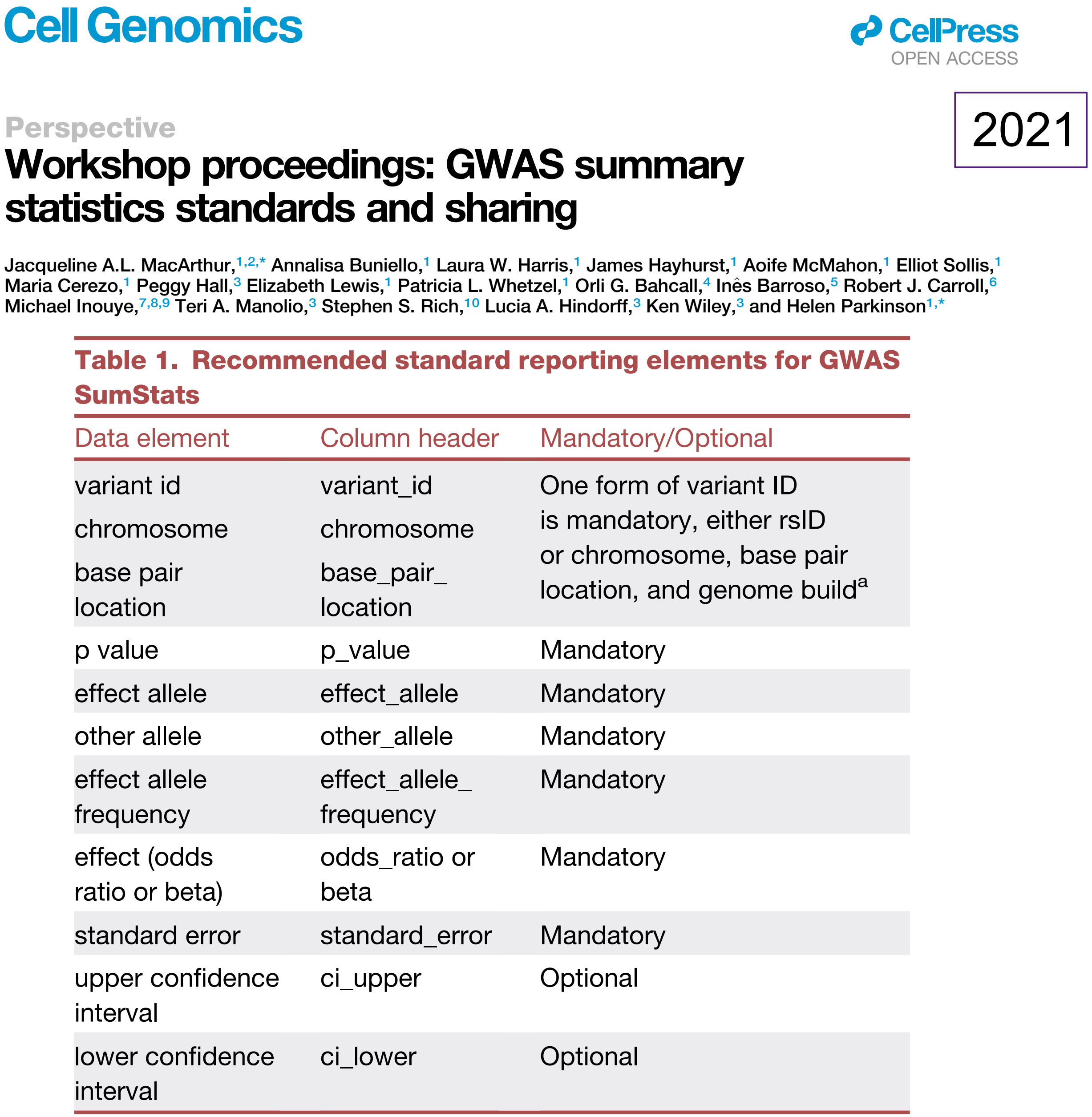
THE UNIVERSITY OF QUEENSLAND AUSTRALIA | Institute for Molecular Bioscience

Program in Complex Trait Genomics

Motivation

- Best prediction methods take genetic values as random effect (e.g., BLUP and BayesR).
- 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.

Sumstats



Cell Genomics

CellPress OPEN ACCESS

2021

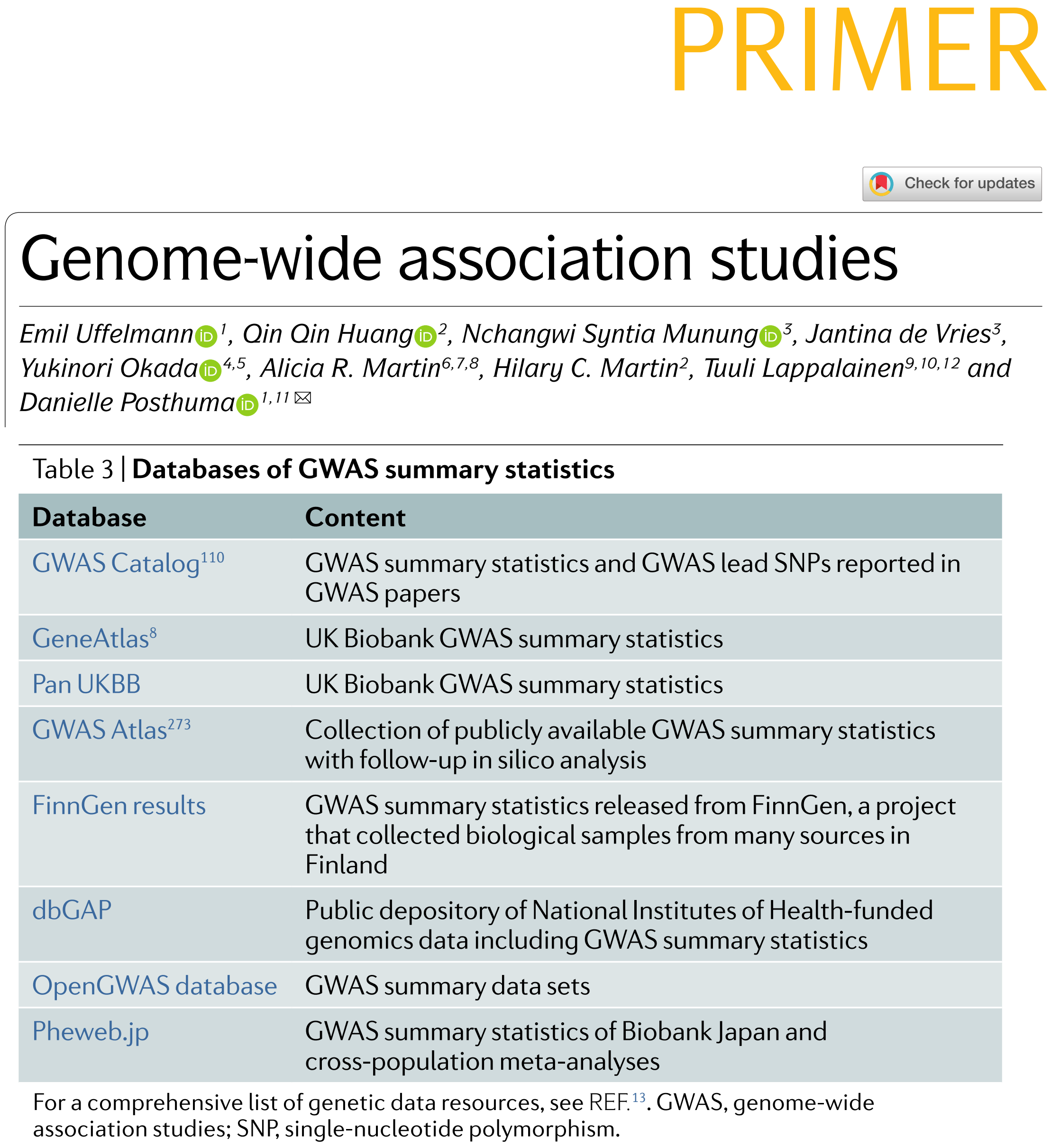
Perspective

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.

Sumstats for PGS prediction

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.

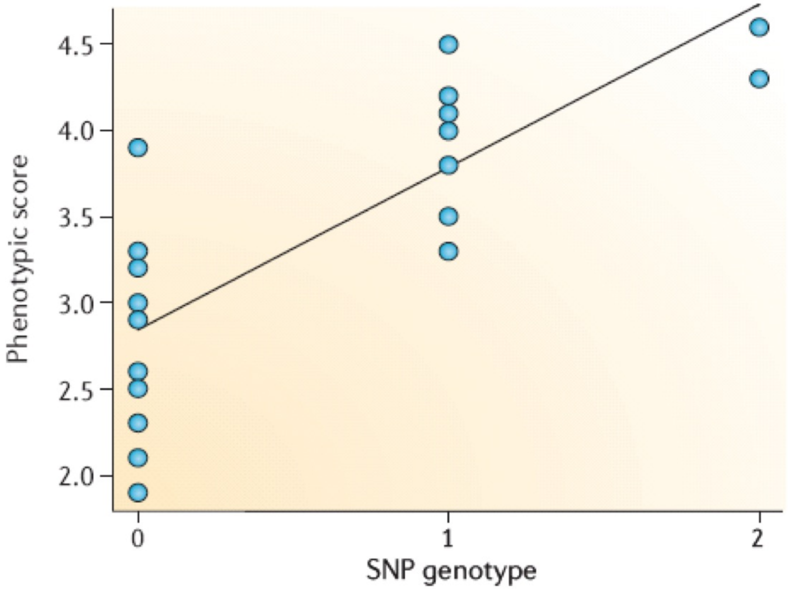
$$b_j = \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\mathbf{y}$$

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.

Sumstats for PGS prediction

SNP marginal effect estimates

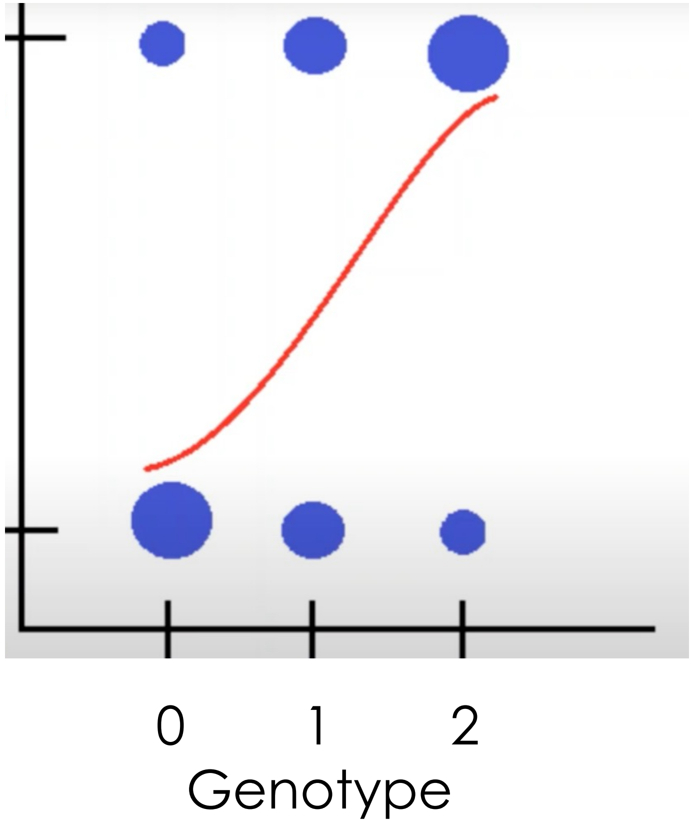
For diseases, GWAS is done using logistic regression

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

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

$$b_j = \log(OR)$$

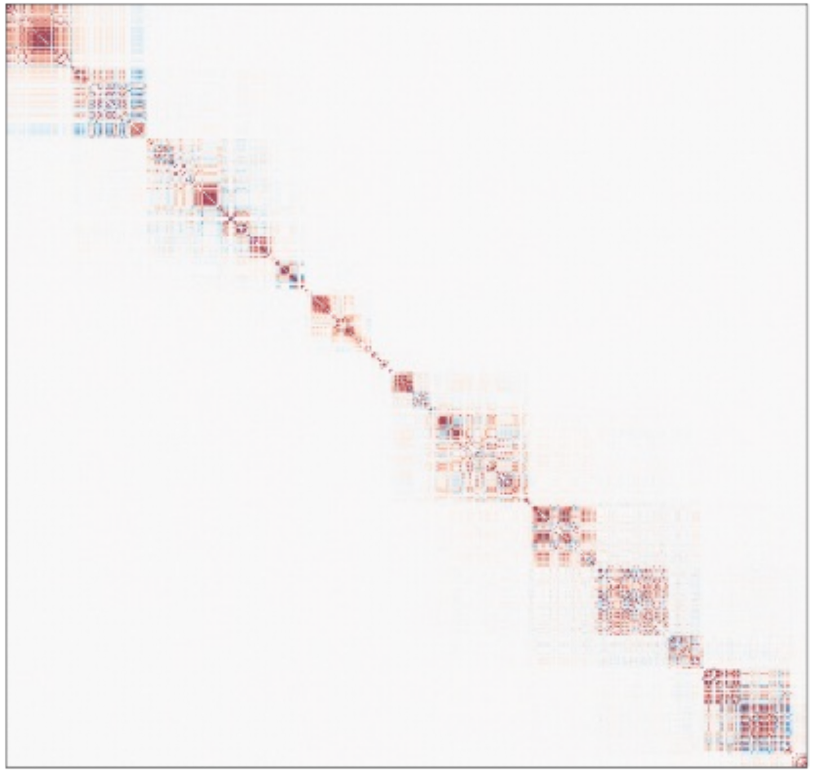
Approximately equal to the b_j from the linear model when true effect size is small.



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

The principle of sumstats-based methods

Principle of sumstats-based methods

Use of summary data only - how does it work?

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

Sufficient statistics

A statistic is **sufficient** if no other statistics 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!

Sumstats for PGS prediction

For simplicity, let’s assume that when running GWAS,

- the genotypes of each SNP are standardised with column mean zero and variance one.
- the phenotypes are standardised with mean zero and variance one.

We will come back to deal with this assumption later.

Principle of sumstats-based methods

BLUP

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

BLUP solutions:

$$\widehat{\boldsymbol{\beta}} = [\underset{\substack{\uparrow \\ n\,\mathbf{R}}}{\mathbf{X}'\mathbf{X}} + \mathbf{I}\lambda]^{-1}\underset{\substack{\uparrow \\ n\,\mathbf{b}}}{\mathbf{X}'\boldsymbol{y}}$$

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 $\boldsymbol{\beta}$.

Compare BLUP and SBLUP

BLUP

- Model:

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

- Estimator:

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda]^{-1}\mathbf{X}'\boldsymbol{y}$$

Genotype matrix

Phenotypes

SBLUP (sumstats-based BLUP)

- Model:

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

- Estimator:

$$\hat{\boldsymbol{\beta}} = [n\mathbf{R} + \mathbf{I}\lambda]^{-1}n\mathbf{b}$$

GWAS sample size

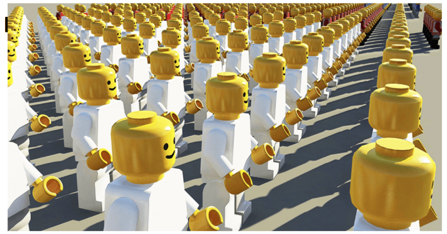
LD correlation matrix

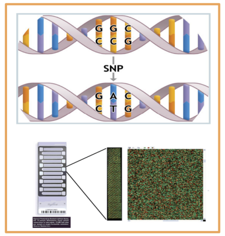
GWAS effects

From individual- to summary-level model

Individual-level data analysis

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





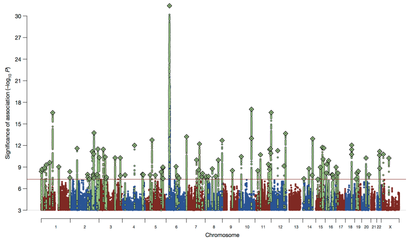
BLUP

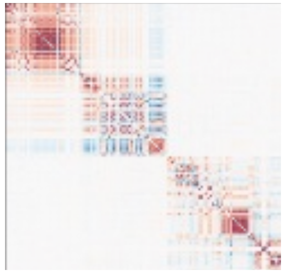
Bayes

→

Summary-level data analysis

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





SBLUP

SBayes

Covariates, such as age and sex, are accounted for when 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

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

Sumstats-based Bayesian methods

SBayes

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

SNP marginal effects from GWAS

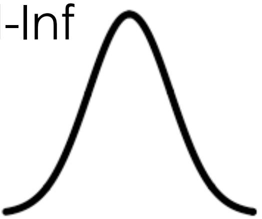
LD correlation matrix

SNP joint effects

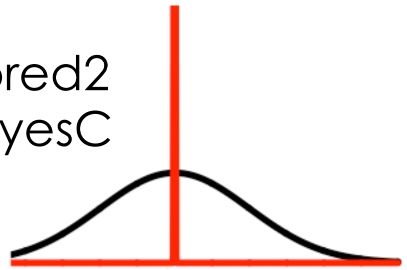
$$Var(\boldsymbol{\epsilon}) = \frac{1}{n}\mathbf{R}\sigma_e^2$$

Prior distribution for each SNP effect

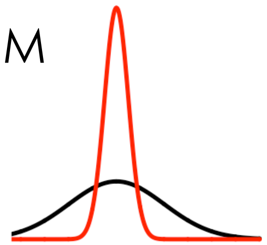
LDpred-Inf
SBLUP



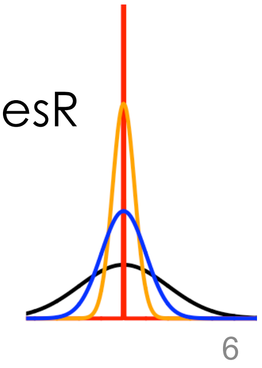
LDpred2
SBayesC



BSLMM



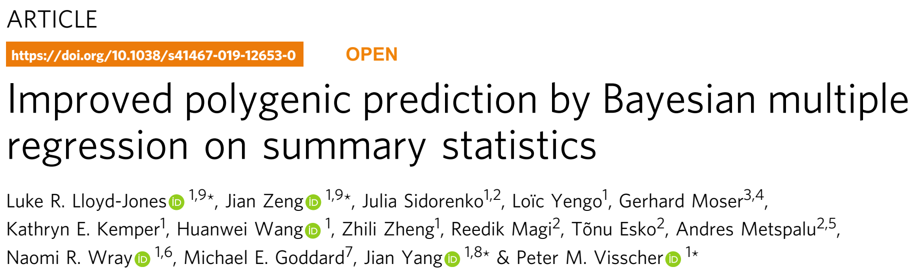
SBayesR



Sumstats-based BayesR

nature
COMMUNICATIONS

SBayesR



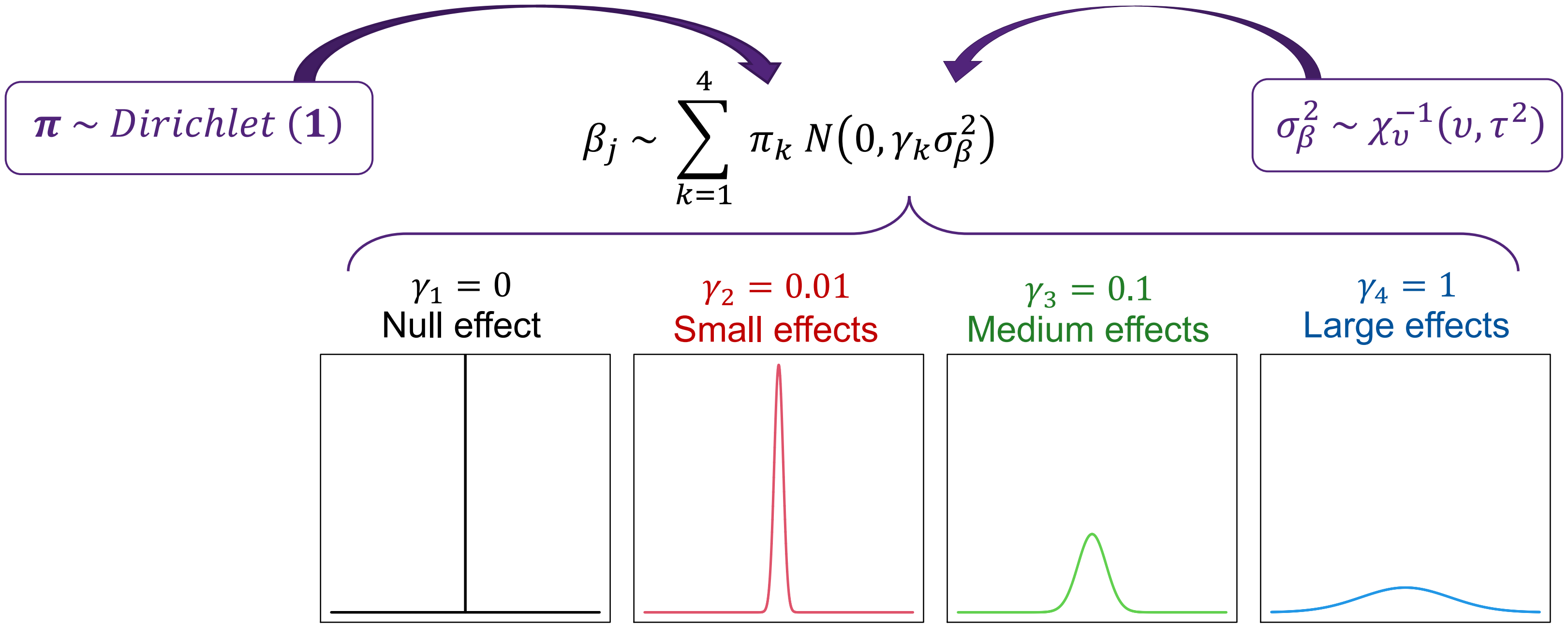
ARTICLE

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

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:



$\boldsymbol{\pi} \sim Dirichlet\ (\mathbf{1})$

$$\beta_j \sim \sum_{k=1}^{4} \pi_k\, N\left(0, \gamma_k \sigma_\beta^2\right)$$

$\sigma_\beta^2 \sim \chi_v^{-1}(v, \tau^2)$

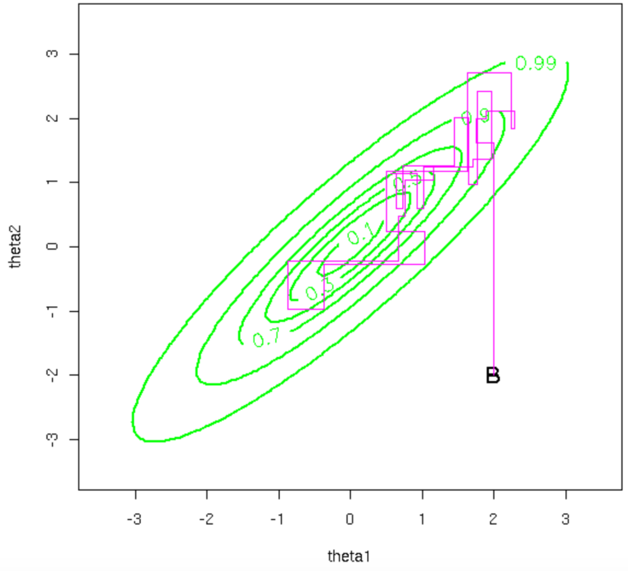
Compare BayesR and SBayesR algorithm

Gibbs sampling

Full conditional distribution for β_j, if it is nonzero,

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

where



Individual-level data

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

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

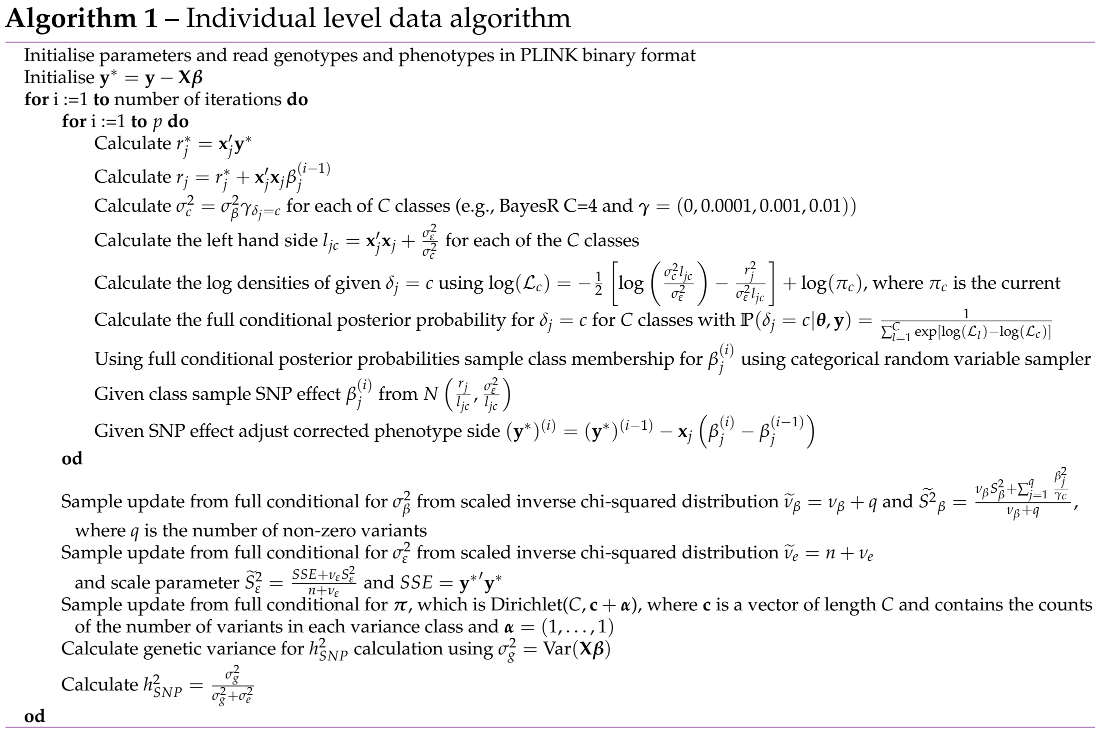
Summary-level data

$$r_j = n b_j - \sum\nolimits_{k \neq j} R_{jk} \beta_k$$

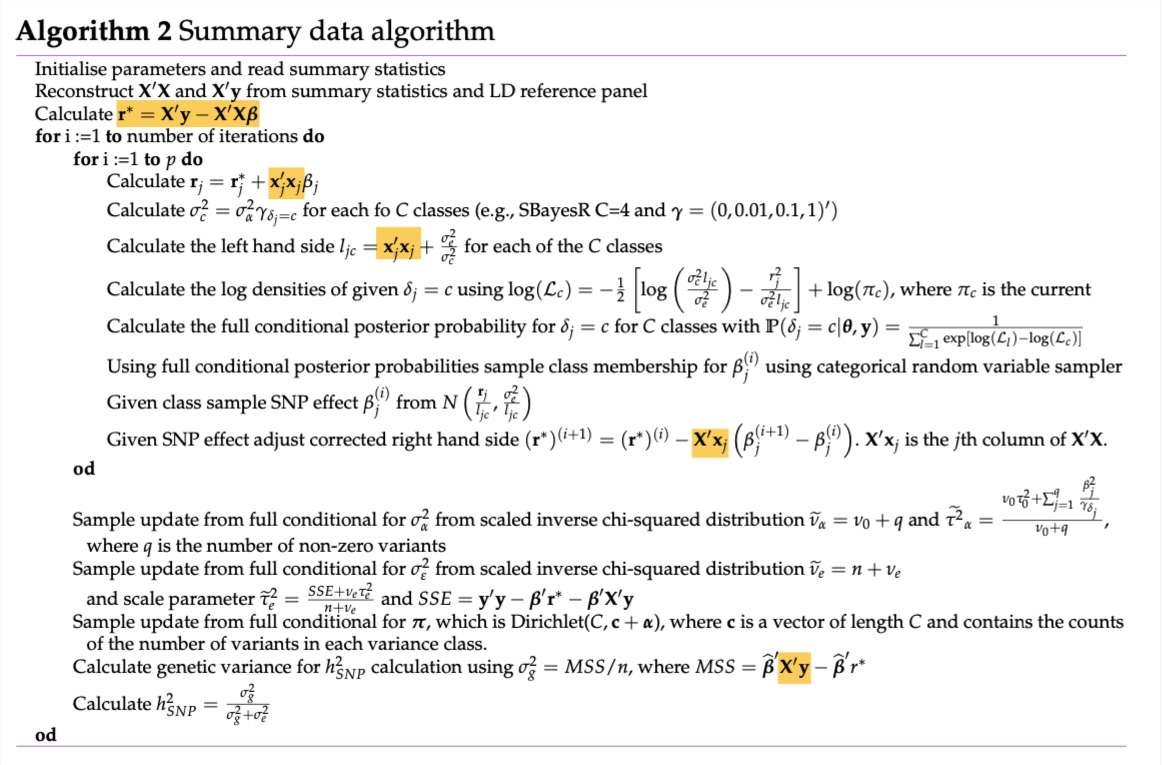
$$C_j = n + \frac{\sigma_e^2}{\gamma_j \sigma_\beta^2}$$

Compare BayesR and SBayesR algorithm

*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 $l_{jc} = \mathbf{x}_j'\mathbf{x}_j + \frac{\sigma_\varepsilon^2}{\sigma_c^2}$ 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 $\pi_c$ 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 $\sigma_\beta^2$ 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 $\sigma_\varepsilon^2$ 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 $h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$
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 $\pi_c$ 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 $j$th column of $\mathbf{X}'\mathbf{X}$.
    od
    Sample update from full conditional for $\sigma_\alpha^2$ 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 $\sigma_\varepsilon^2$ 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 $h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$
od
```

Lloyd-Jones & Zeng et al. 2019 NC supplement

Scaling GWAS effects

Now we deal with the condition of unstandardised genotypes/phenotypes:

- Typically, GWAS are performed using allele counts (0/1/2) as genotypes (X_j^{cnt})
- often with unstandardised phenotypes ($\mathrm{Var}(y) \neq 1$).

The solutions is to ‘scale’ the GWAS marginal effects before the analysis and ‘unscale’ the estimated joint effects after the analysis.

Scaling GWAS effects

Let σ_j be the SD of genotypes for SNP j and σ_y be the SD of phenotypes.

The genotypic value

$$g_j = X_j^{cnt} b_j^{cnt} = \frac{X_j^{cnt}}{\sigma_j} \times \sigma_j b_j^{cnt}$$

$$\frac{g_j}{\sigma_y} = X_j \ \frac{\sigma_j}{\sigma_y} b_j^{cnt} = X_j \ s_j b_j^{cnt} = X_j \boxed{b_j}$$

This is in the SD units

All we need to do is to get

$$b_j = s_j \boxed{b_j^{cnt}}$$

Output from GWAS

where s_j can be estimated by

$$s_j = \sqrt{\frac{1}{nSE_j^2 + b_j^2}}$$

Summary (1)

- Minimum data required for sumstat-based methods are
  - ➢ GWAS effects, standard errors, GWAS sample size, LD matrix

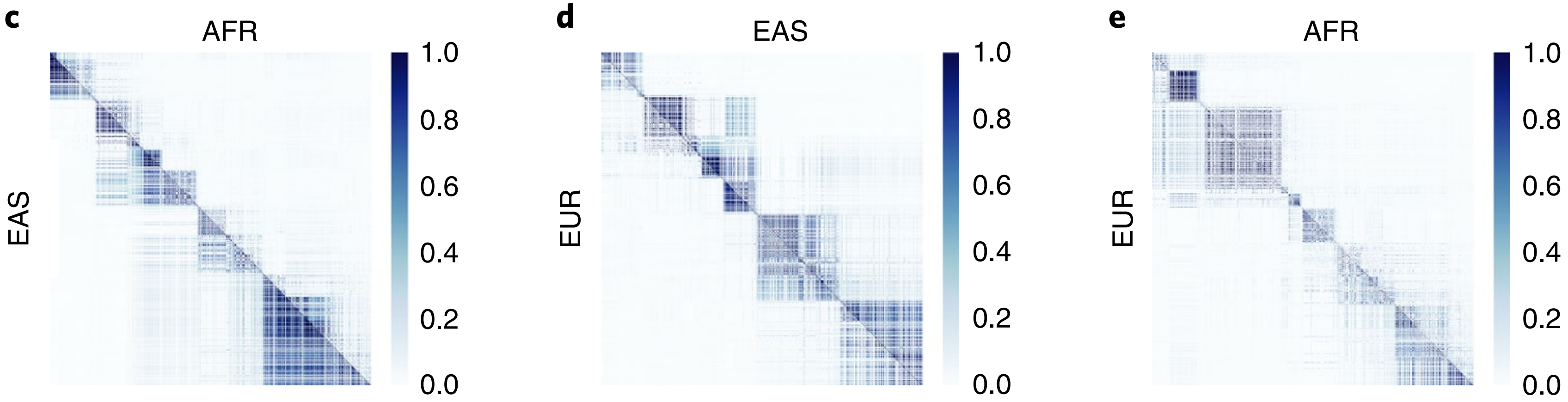
- In principle, SBayes and Bayes are equivalent methods when **same data** are used.

- 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 and population structure
- Minimum sampling variance in LD → LD ref sample size cannot be too small



LD decays to zero between distant SNPs

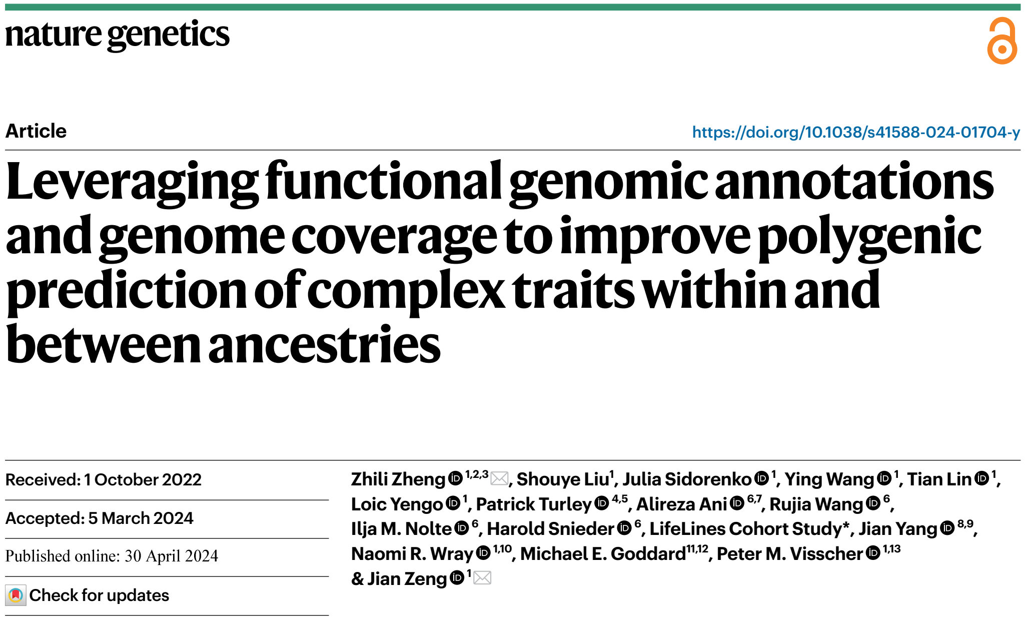
- Can use sparse or block-wide LD matrices

Regulation of LD matrix

Lloyd-Jones et al (2019) used chromosome-wide shrunk LD matrices.

Zheng et al (2024) used eigen-decomposed matrices from LD blocks.

- More robust to LD heterogeneity → better prediction performance
- Faster → allows us to fit multi-million SNPs simultaneously



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

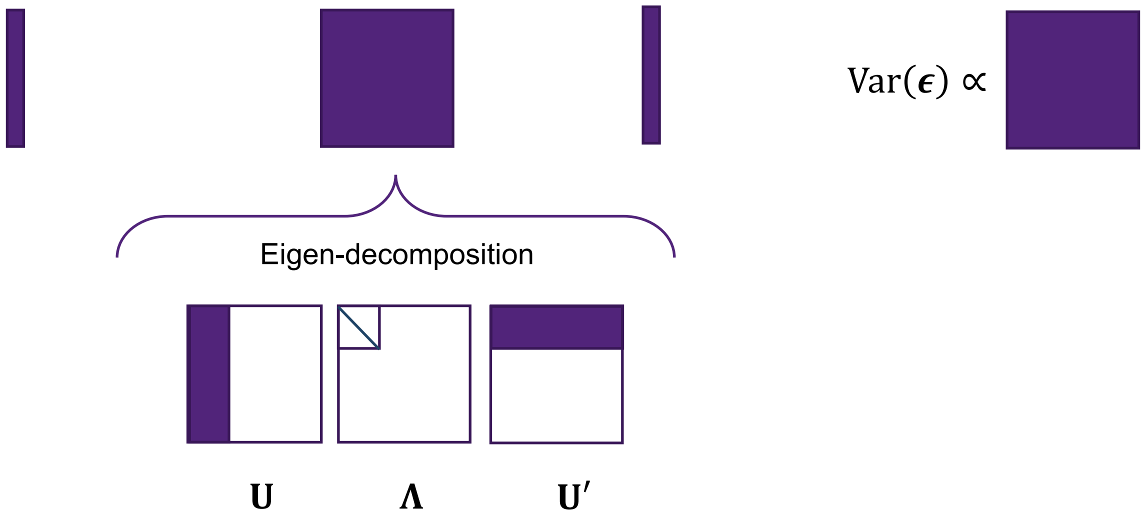
Zhili Zheng[1,2,3], Shouye Liu[1], Julia Sidorenko[1], Ying Wang[1], Tian Lin[1], Loic Yengo[1], Patrick Turley[4,5], Alireza Ani[6,7], Rujia Wang[6], Ilja M. Nolte[6], Harold Snieder[6], LifeLines Cohort Study*, Jian Yang[8,9], Naomi R. Wray[1,10], Michael E. Goddard[11,12], Peter M. Visscher[1,13]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

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



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

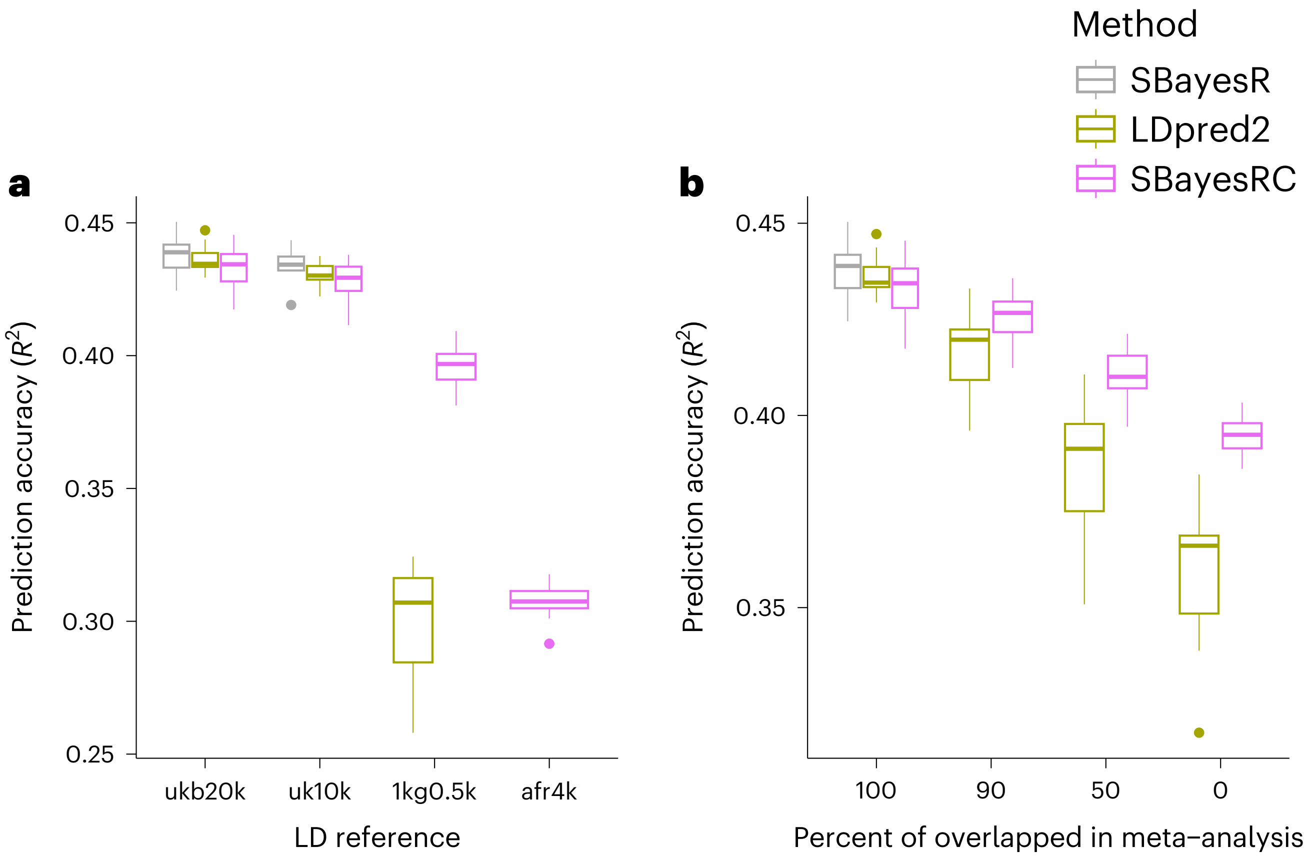
$$\mathbf{w} = \mathbf{Q}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Var($\boldsymbol{\varepsilon}$) ∝

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

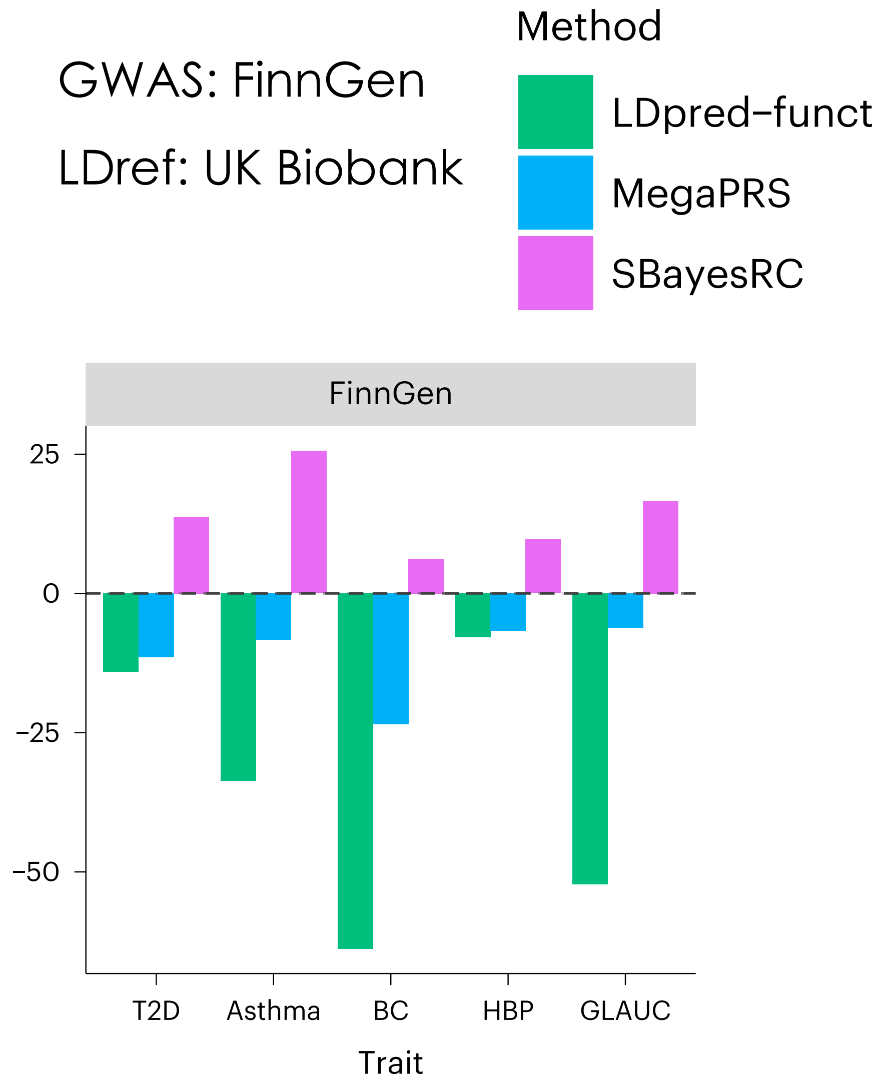
Low-rank model

Improved robustness



GWAS: FinnGen

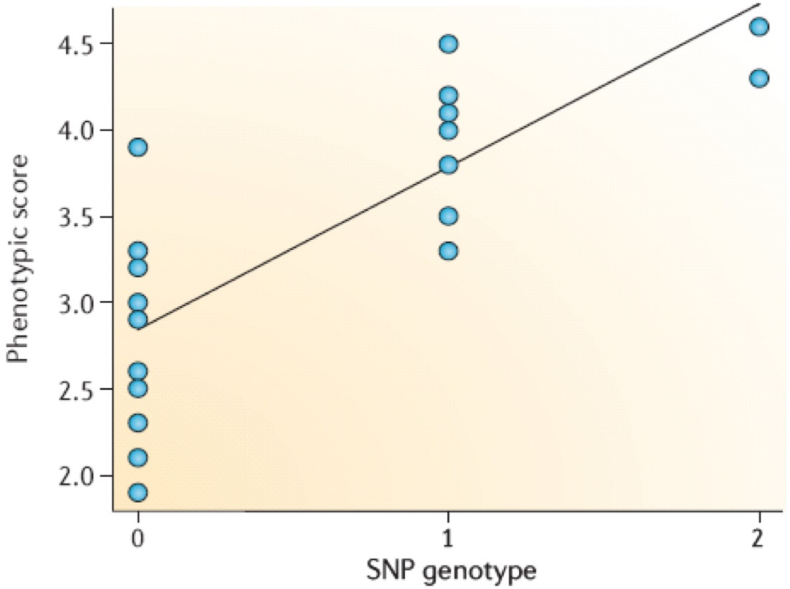
LDref: UK Biobank



Other information critical to quality control (QC)

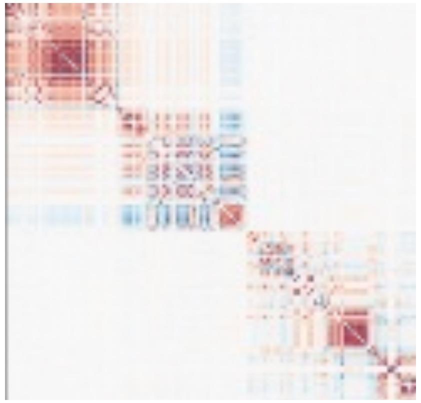
Which allele is the **effect allele** in GWAS?

e.g., A1 allele



A2A2 A1A2 A1A1

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



Other critical information

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.

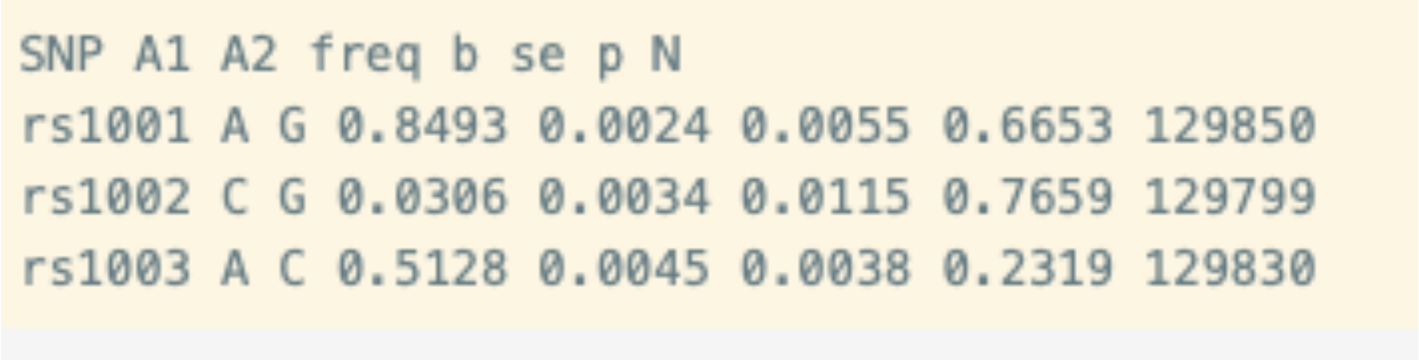
We recommend to visualise the per-SNP sample size distribution and remove the outliers.

Other critical information

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
```

Summary (2)

- Minimum data required for sumstat-based methods are
  - ➢ GWAS effects, standard errors, GWAS sample size, LD matrix

- Other information are critical/useful to quality control.

- SBayes an approx. to Bayes when LD is estimated from a reference sample, but unleashes the power of large GWAS sample size.

- Matrix regulation/factorisation can better model LD.

Incorporating functional annotations

Functional genomic annotations

Functional genomic annotations provide orthogonal information useful for polygenic prediction.

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

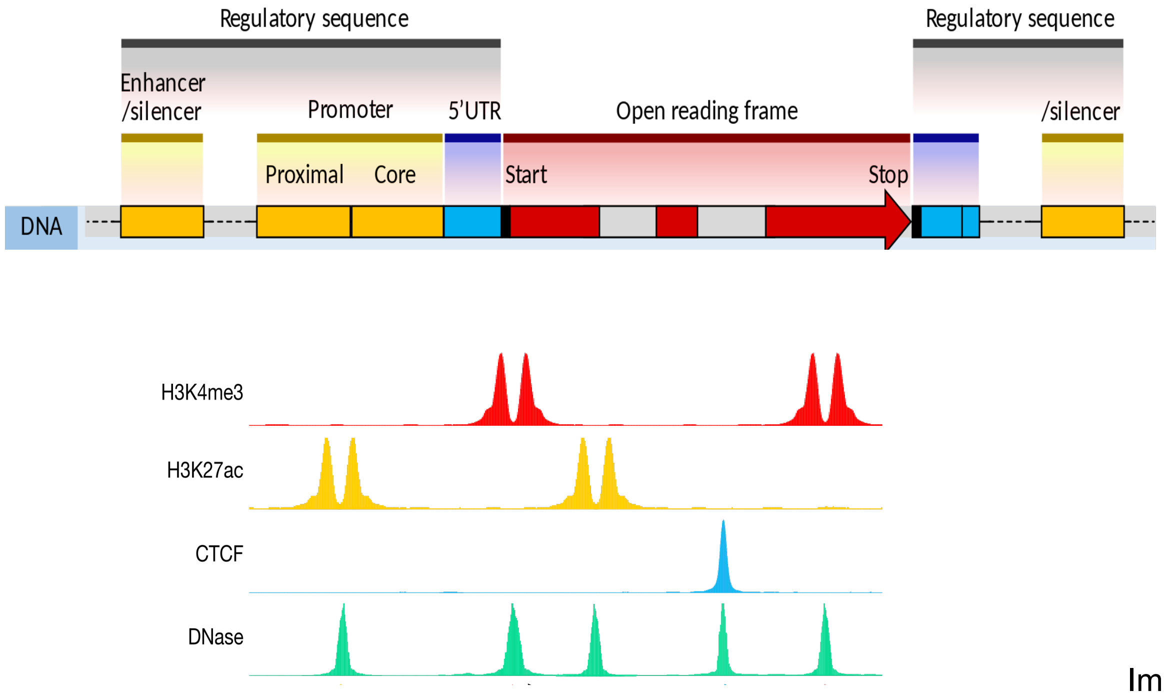


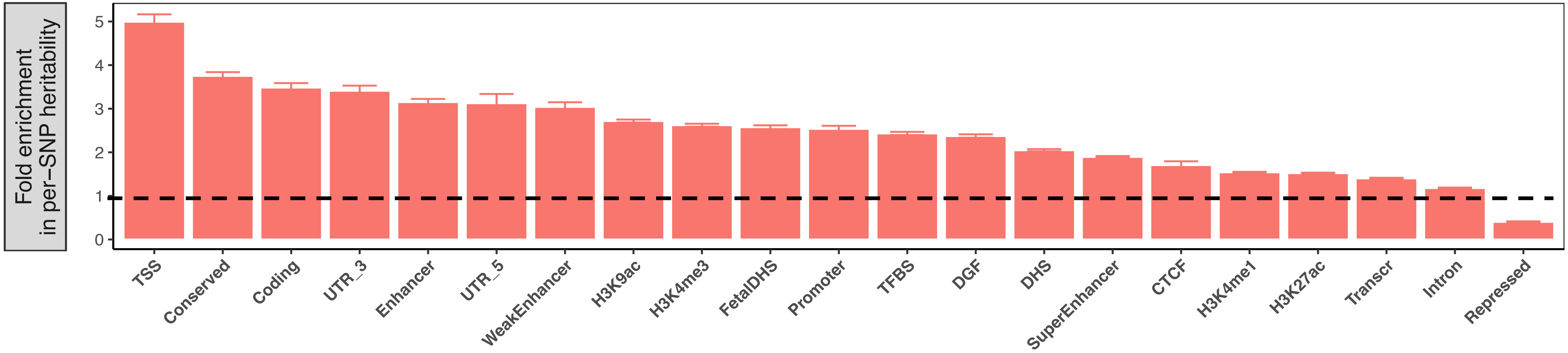
Image from ENCODE

Functional genomic annotations

Functional genomic annotations provide orthogonal information useful for polygenic prediction.

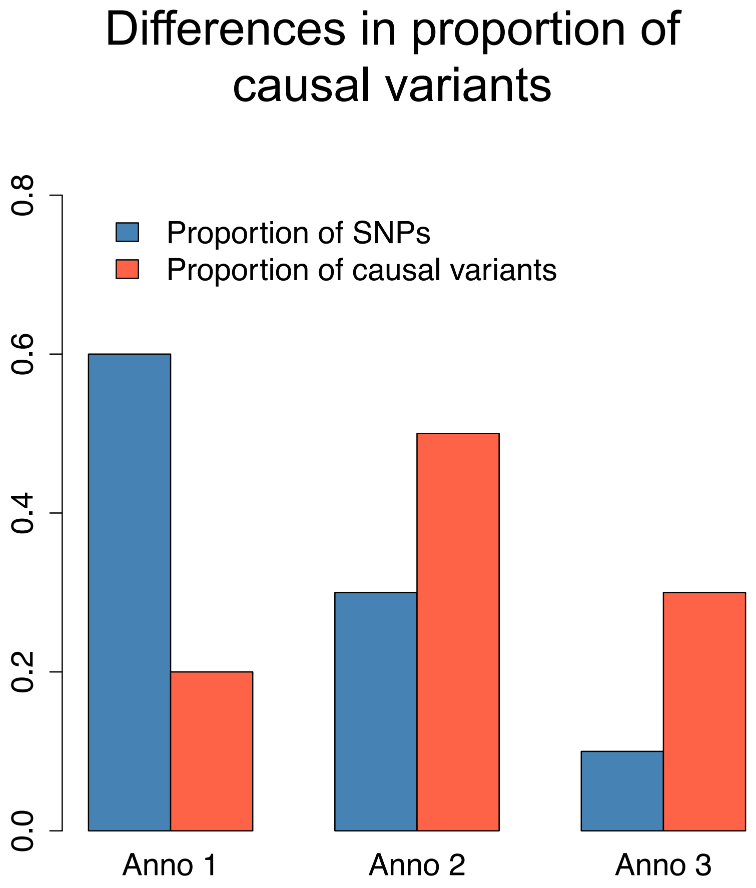
- Chromatin states
- Biological functions
- Molecular quantitative trait loci (xQTL)
- ……

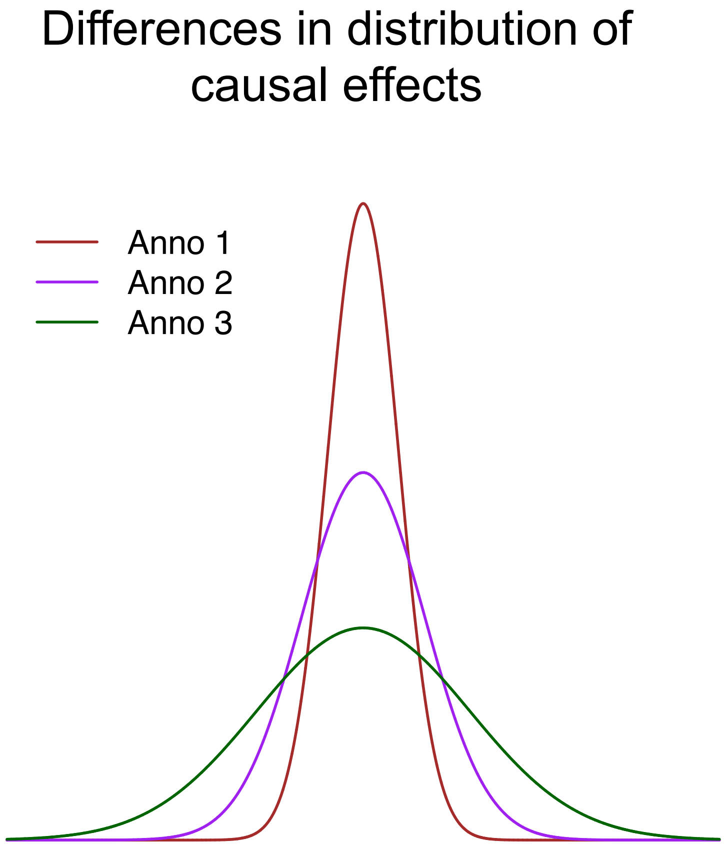
Zeng et al 2021 Nature Communications



Opportunities/challenges

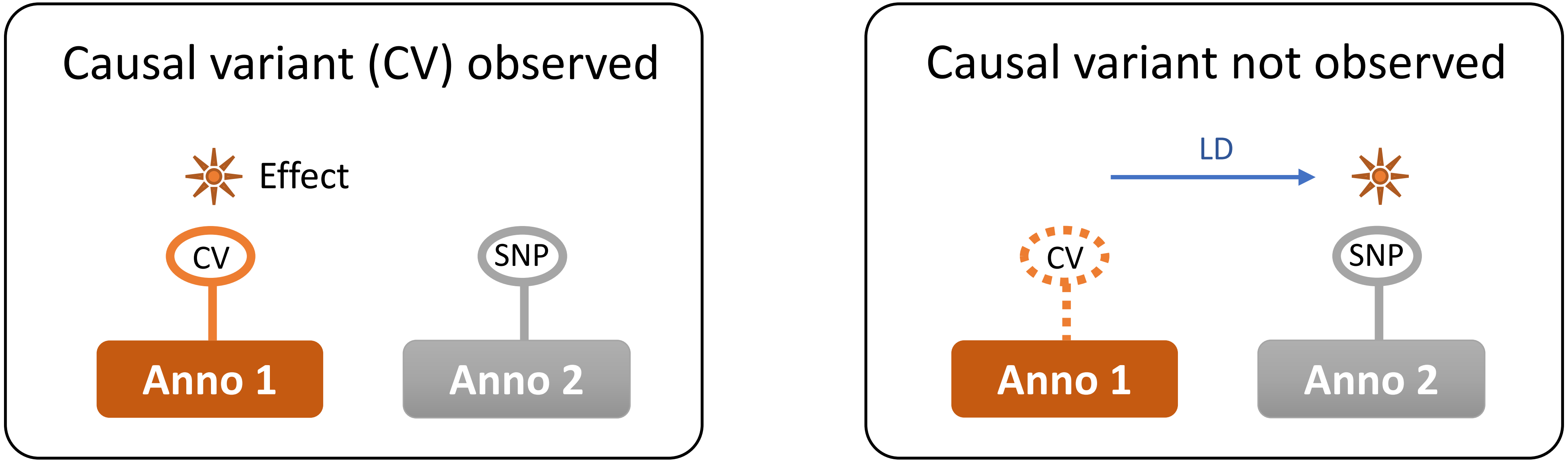
Functional annotations are informative on both the presence of causal variants and the distribution of causal effect sizes.





Opportunities/challenges

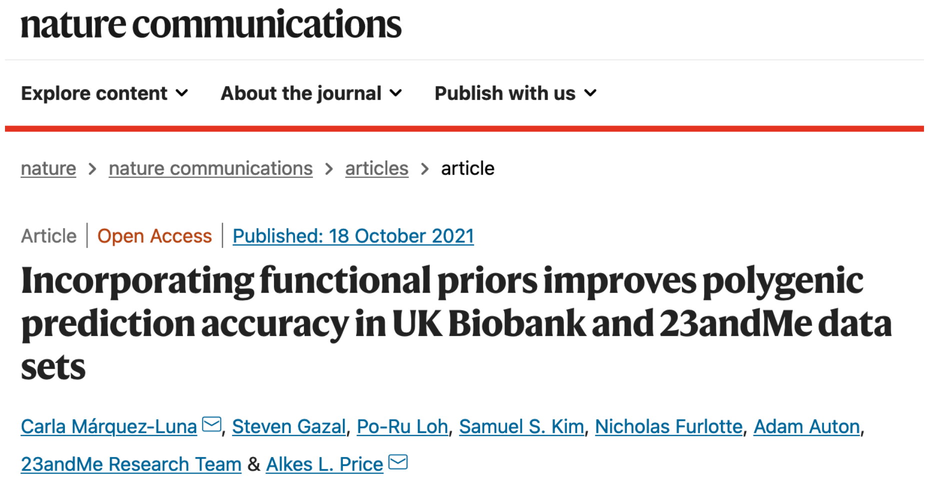
When causal variants are not observed, SNP markers can tag the causal variant by LD but may not tag by annotation.



It's best to model all SNPs simultaneously with their annotations!

Literature

The University of Queensland
AUSTRALIA



nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

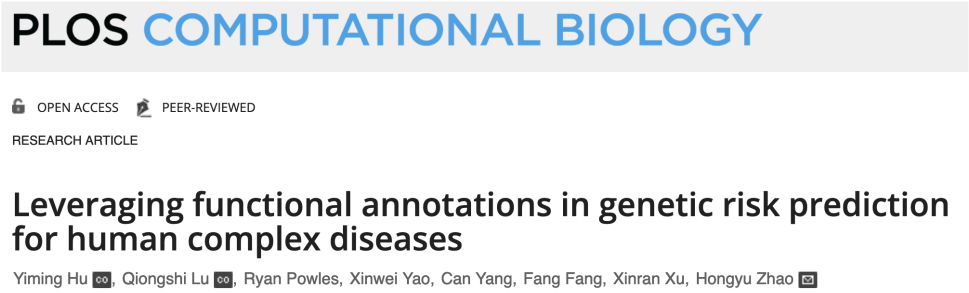
nature > nature communications > articles > article

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

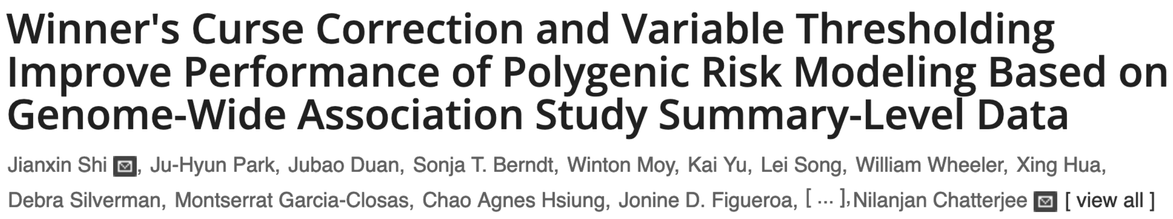
OPEN ACCESS PEER-REVIEWED

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 ✉

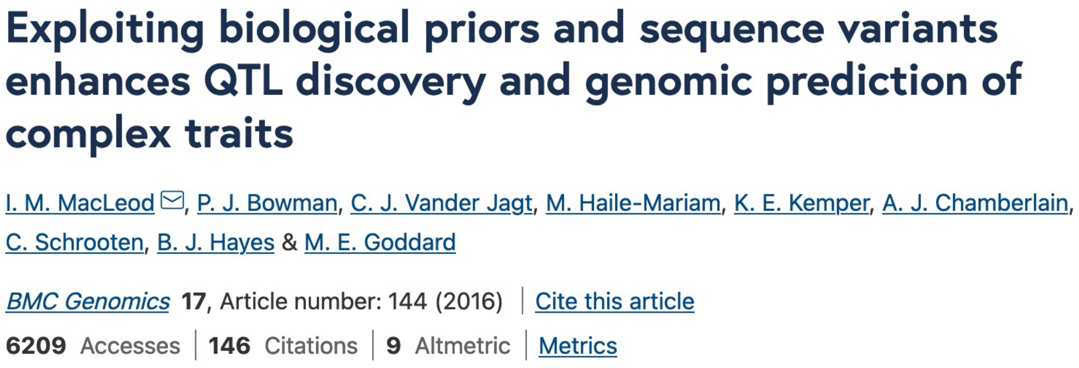
AnnoPred



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



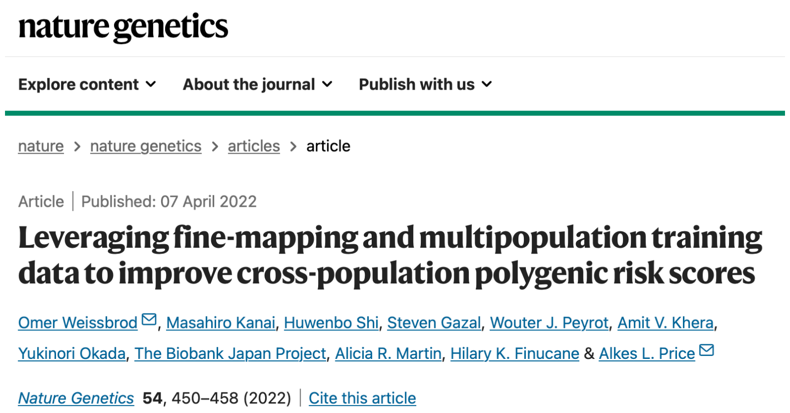
Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits

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

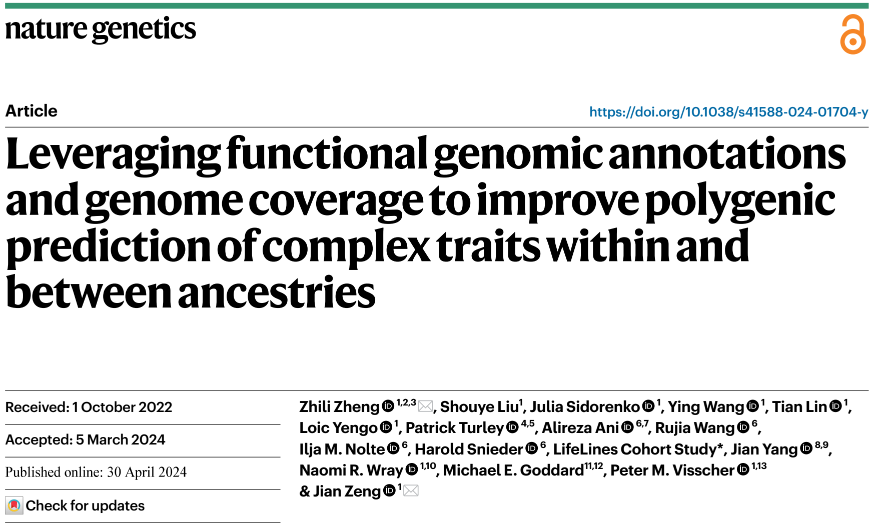
PolyPred

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

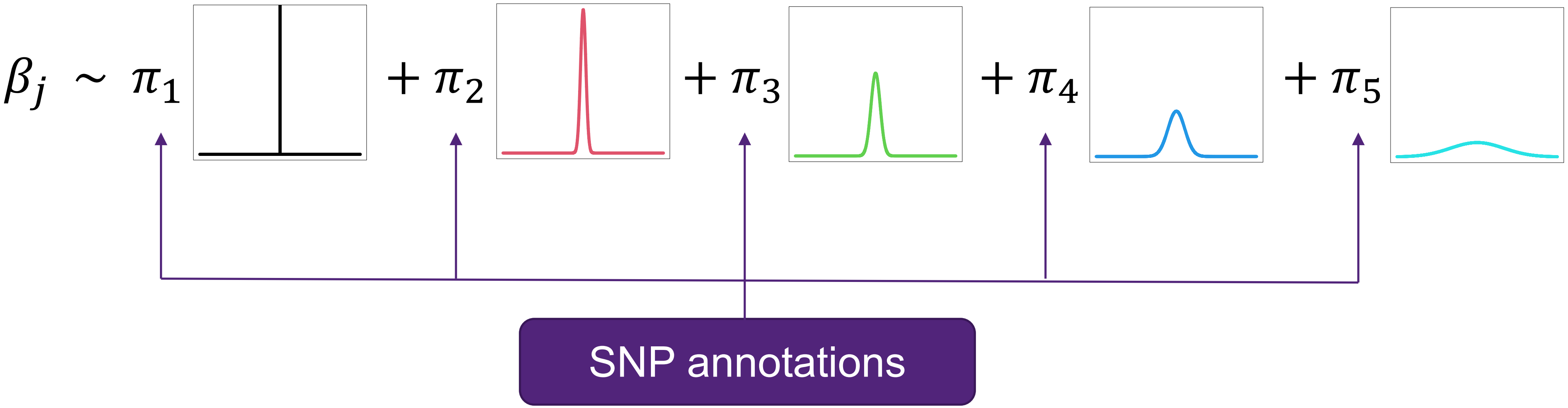
Published online: 30 April 2024

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] ✉

SBayesRC

Incorporate functional annotations through a hierarchical prior:



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

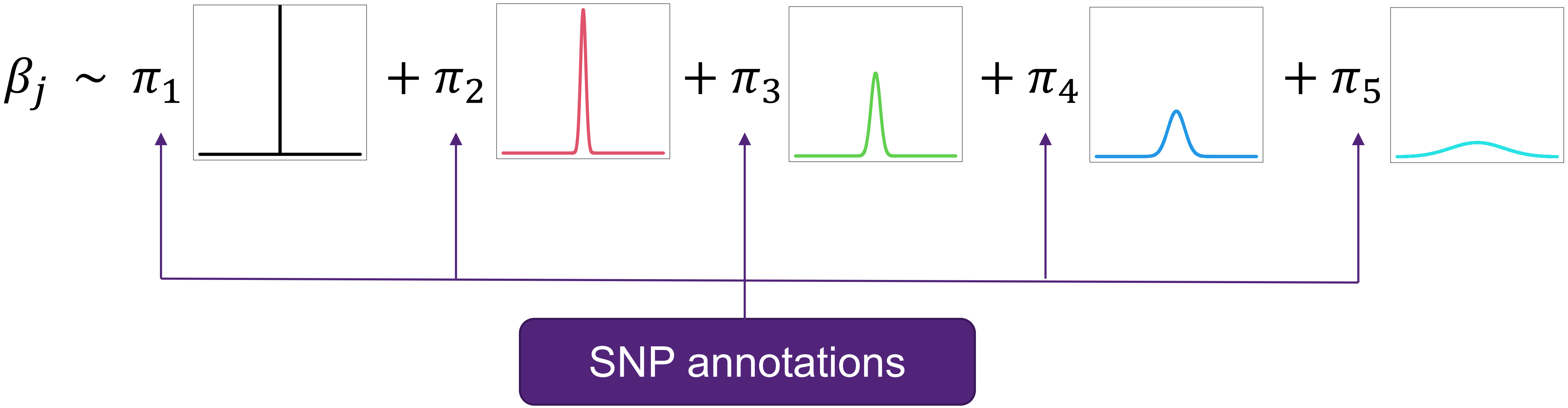
SNP annotations

$$f\left(\pi_{jk}\right) = \sum \text{SNP annotation} \times \text{annotation effect}$$

- The annotation effects are estimated from the data.
- A positive annotation effect increases the probability of the SNP belong to that distribution.

SBayesRC

Incorporate functional annotations through a hierarchical prior:



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

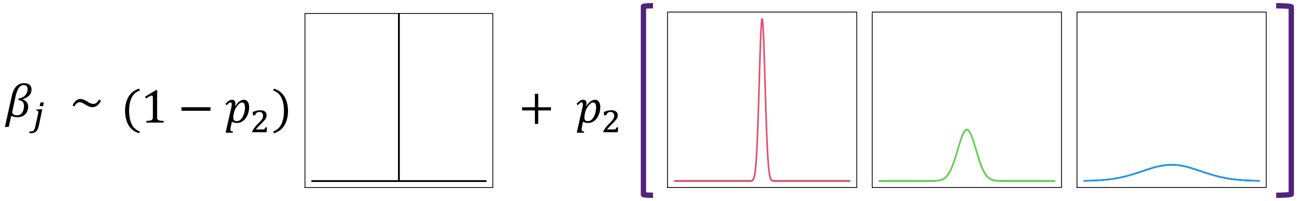
SNP annotations

$$f\left(\pi_{jk}\right) = \sum \text{SNP annotation} \times \text{annotation effect}$$

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

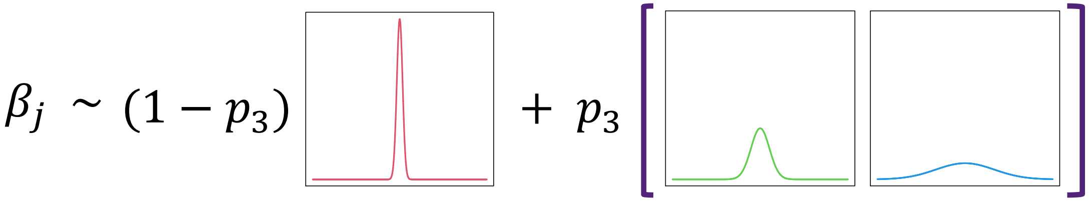
Model annotation effects (suppose 4 components for simplicity)

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



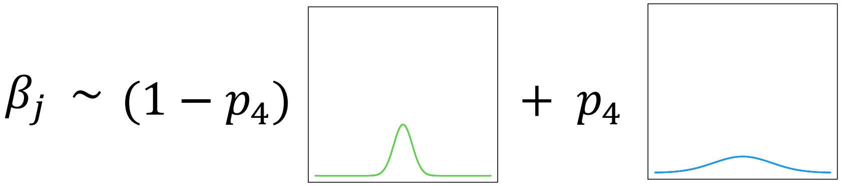
$\beta_j \sim (1-p_2)$ $+ \; p_2$

- For SNPs with nonzero effects



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

- For SNPs with at least medium effects



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

Model 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.

Toy example

	Genome	Region 1	Region 2	Region 3
SNP 1	1	**1**	0	0
SNP 2	1	0	**1**	0
SNP 3	1	**1**	**1**	0
SNP 4	1	0	0	**1**
SNP 5	1	**1**	0	0

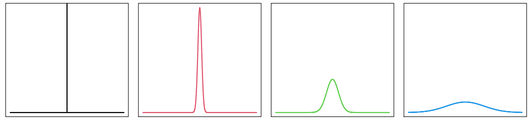
Input data

X

Anno Effect Matrix

Estimate from the data

$\boldsymbol{p}$

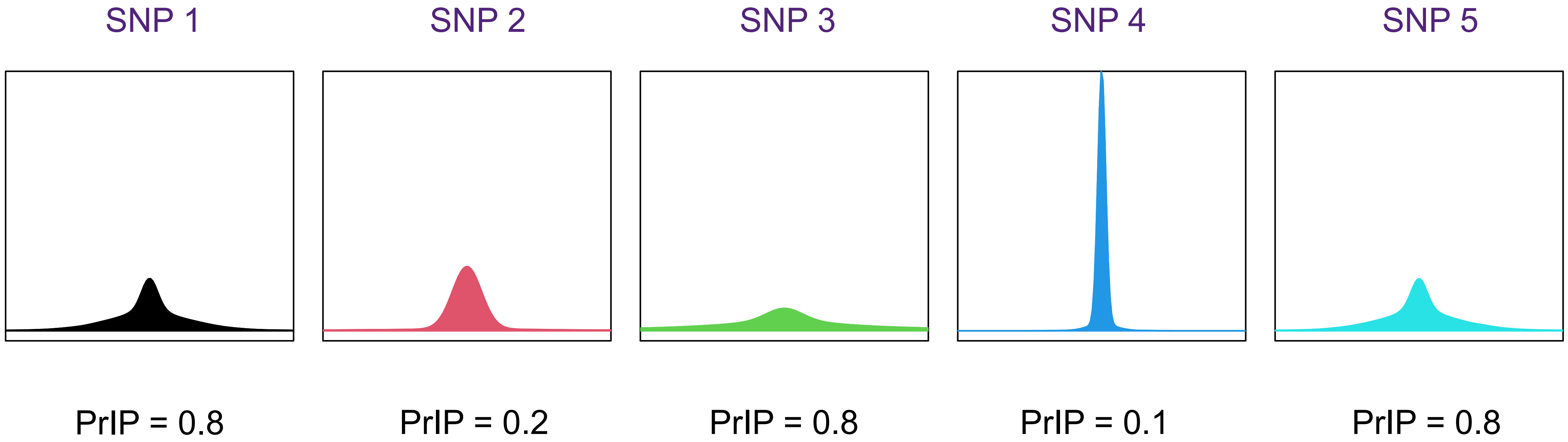


	π_1	π_2	π_3	π_4
SNP 1	0.2	0.1	0.6	0.1
SNP 2	0.8	0.02	0.02	0.16
SNP 3	0.2	0.0	0.2	**0.6**
SNP 4	0.9	0.08	0.01	0.01
SNP 5	0.2	0.1	0.6	0.1

sum is PrIP
(prior inclusion probability)

Toy example

Prior distribution of SNP effect is annotation dependent.



PrIP: Prior Inclusion Probability $= \pi_2 + \pi_3 + \pi_4 = 1 - \pi_1$

Real data analysis

GWAS datasets

biobank uk

FINNGEN

lifelines

PAGE

GIANT CONSORTIUM

Multiple ancestries

- European (EUR)
- East Asian (EAS)
- South Asian (SAS)
- African (AFR)

SNP panels (MAF>0.01)

- 1M HM3 SNPs
- 7M imputed SNPs

Annotation data

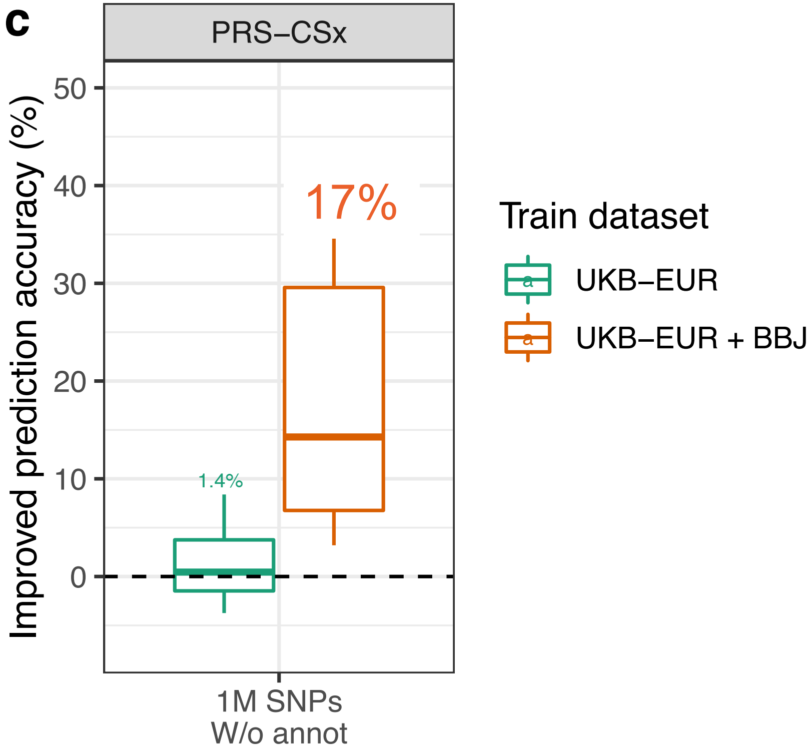
- BaselineLDv2.2 (Gazal et al 2017 NG)
- 96 genomic annotations

Methods compared

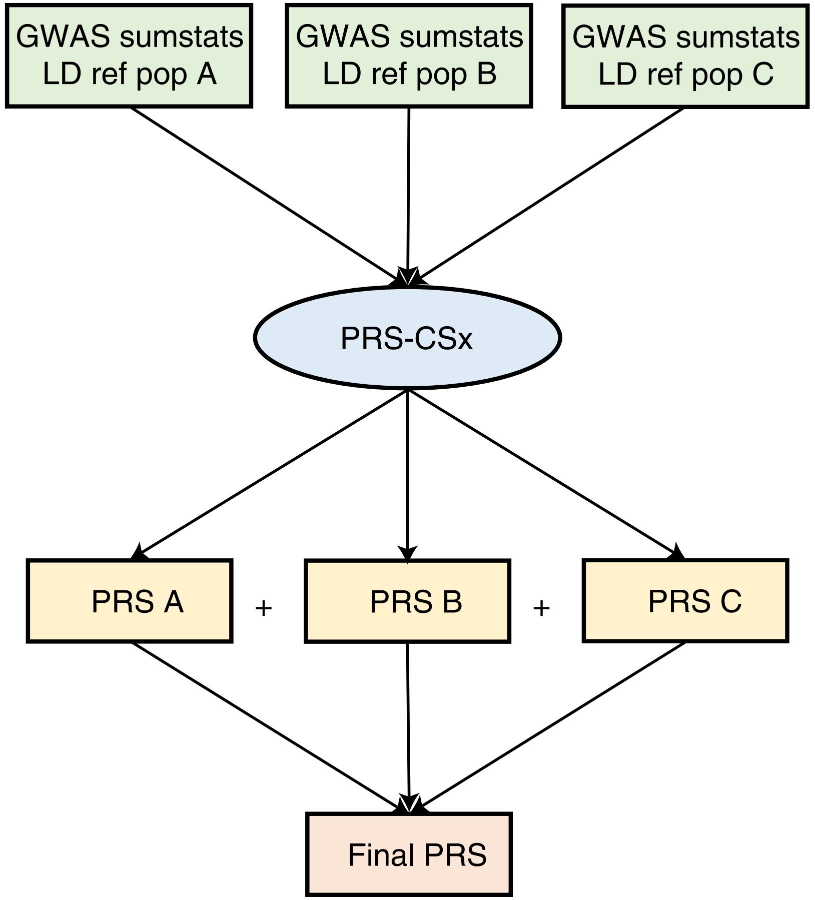
- SBayesR
- LDpred2
- LDpred-funct
- MegaPRS
- PolyPred-S
- PRS-CSx

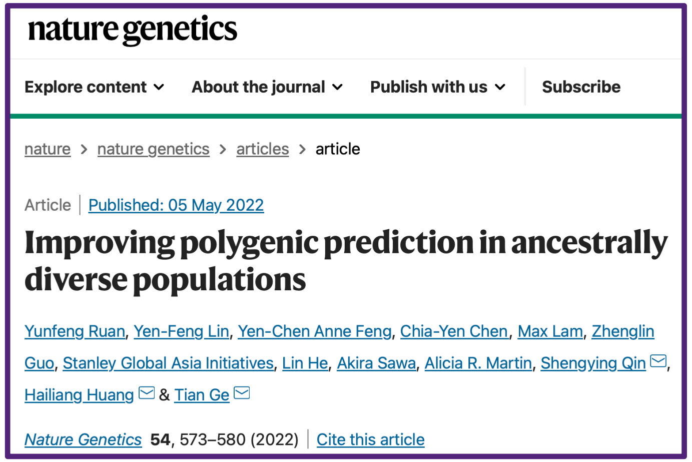
Trans-ancestry prediction

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



PRS-CSx





nature genetics

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

nature > nature genetics > articles > article

Article | Published: 05 May 2022

Improving polygenic prediction in ancestrally diverse populations

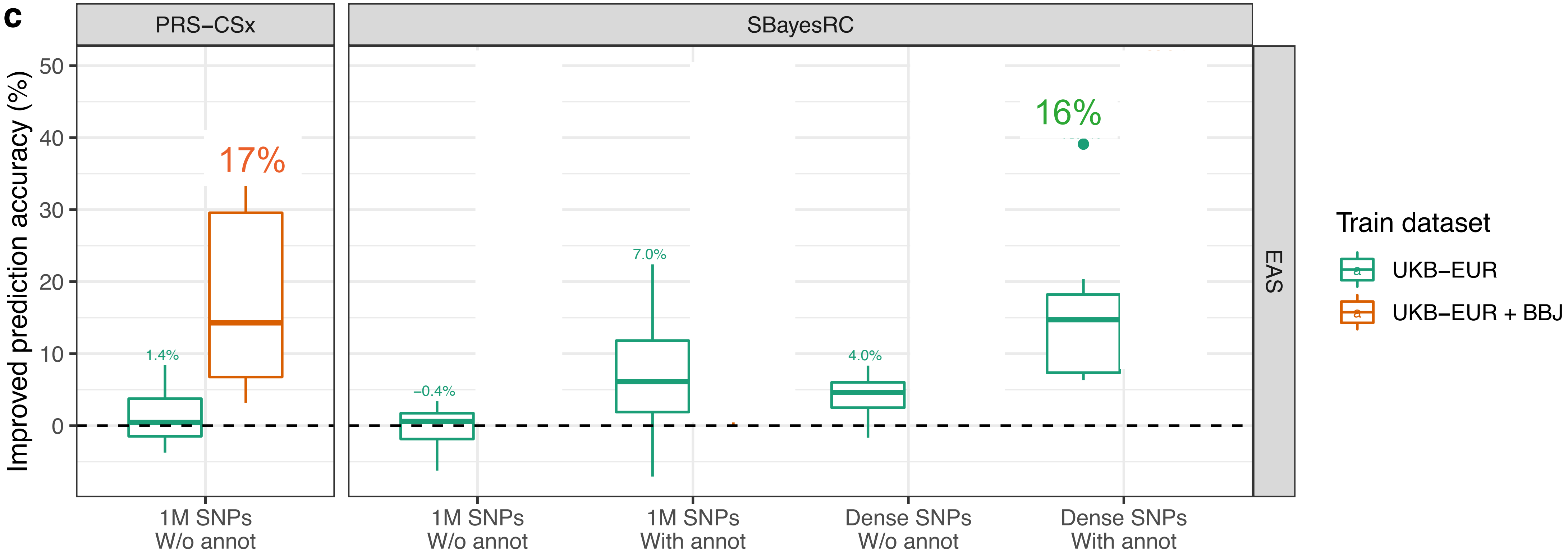
Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Stanley Global Asia Initiatives, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin ✉, Hailiang Huang ✉ & Tian Ge ✉

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

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

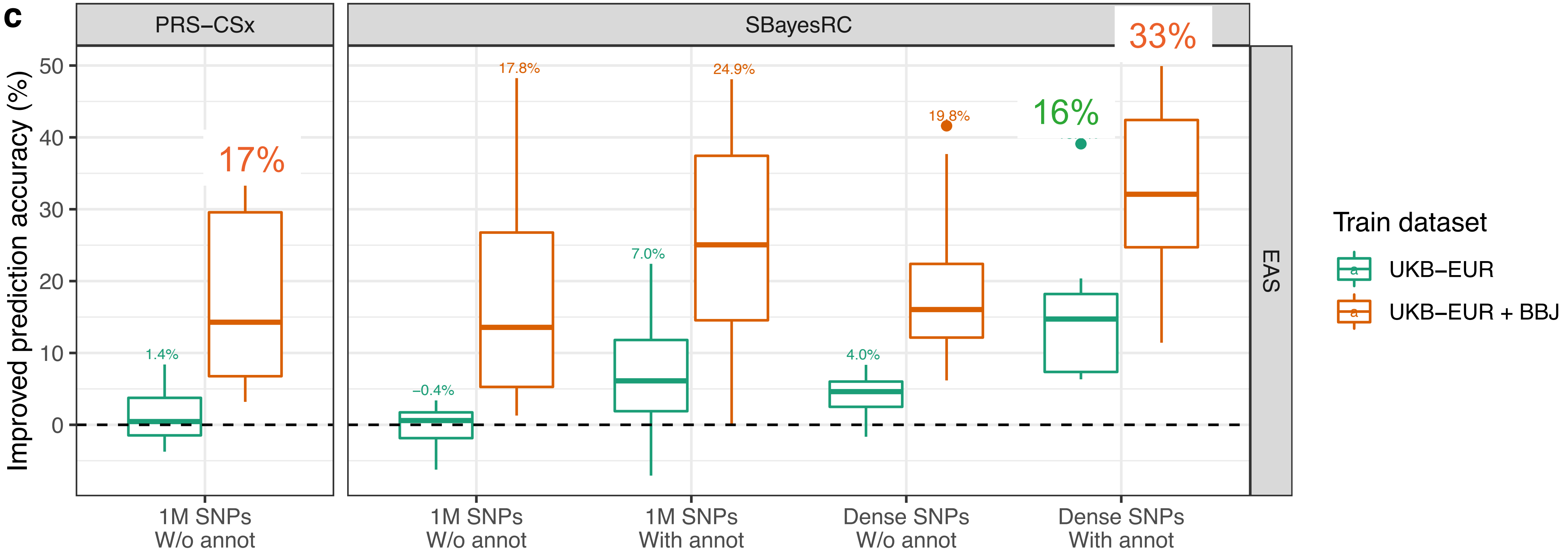
Trans-ancestry prediction

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



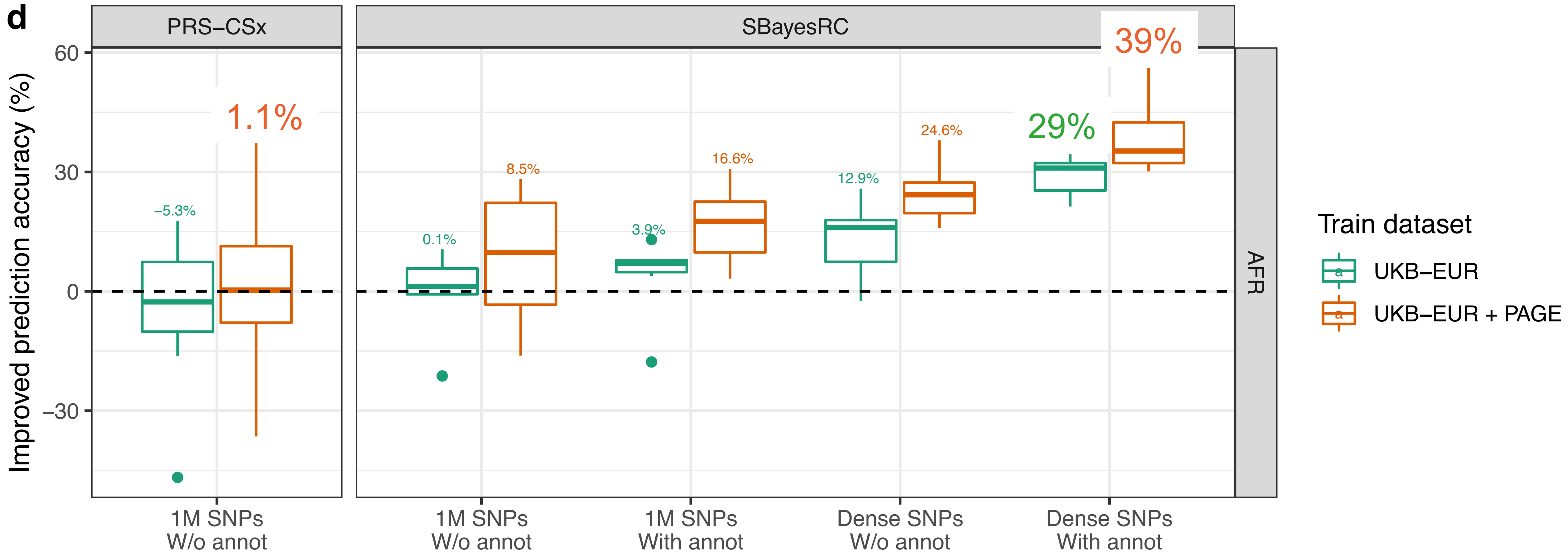
Trans-ancestry prediction

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

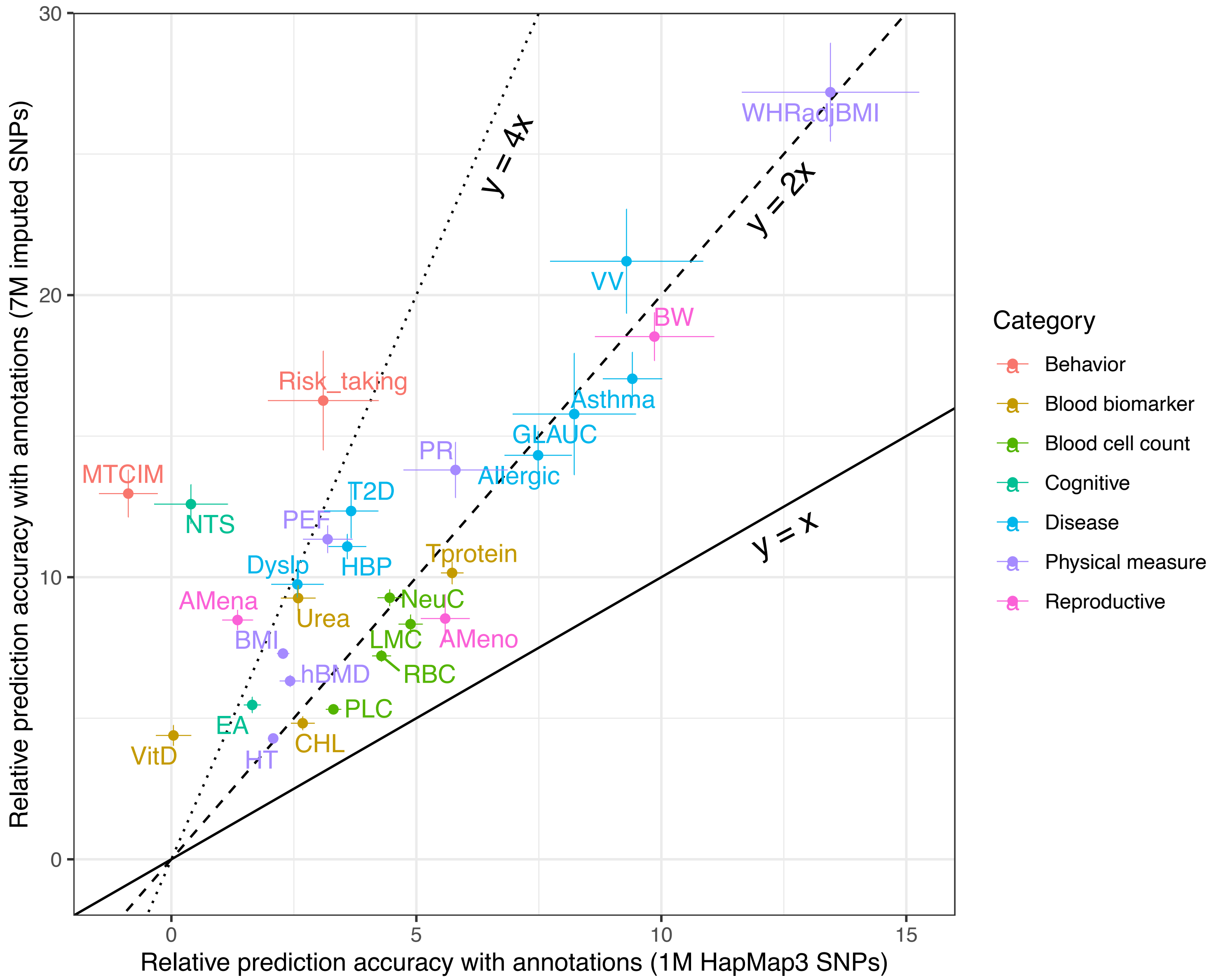


Trans-ancestry prediction

Use GWAS data from UKB EUR and PAGE (mixed) AFR to predict UKB AFR



Interaction between SNP density and annotation information



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

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

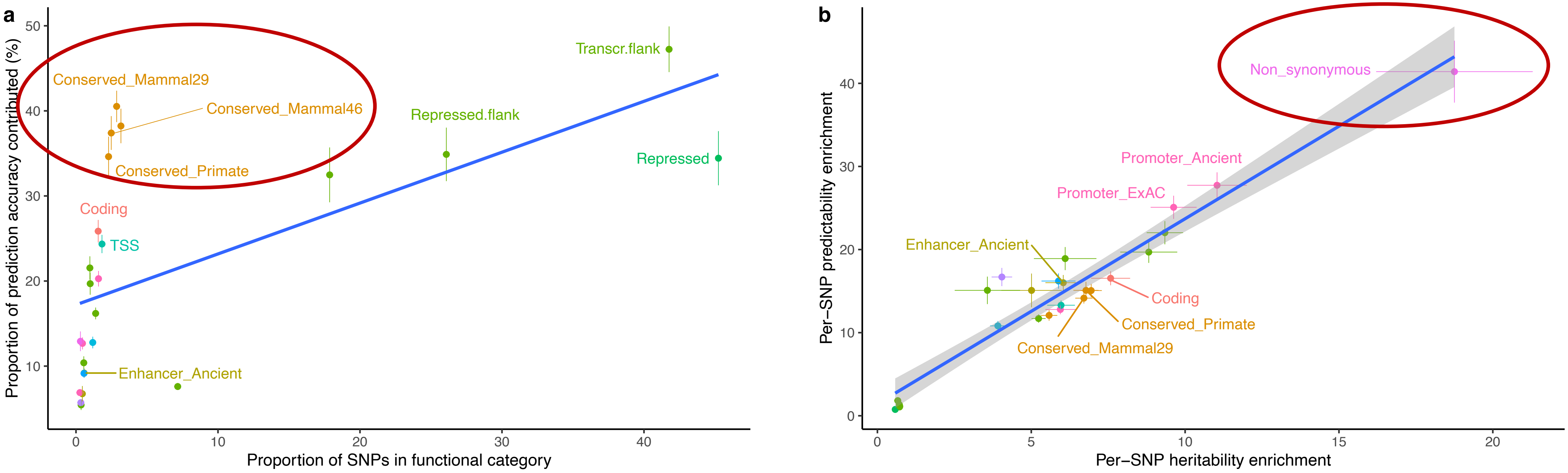
using 7M imputed SNPs (y-axis) or 1M HapMap3 SNPs (x-axis).

Annotations help more with more SNPs - **Why?**

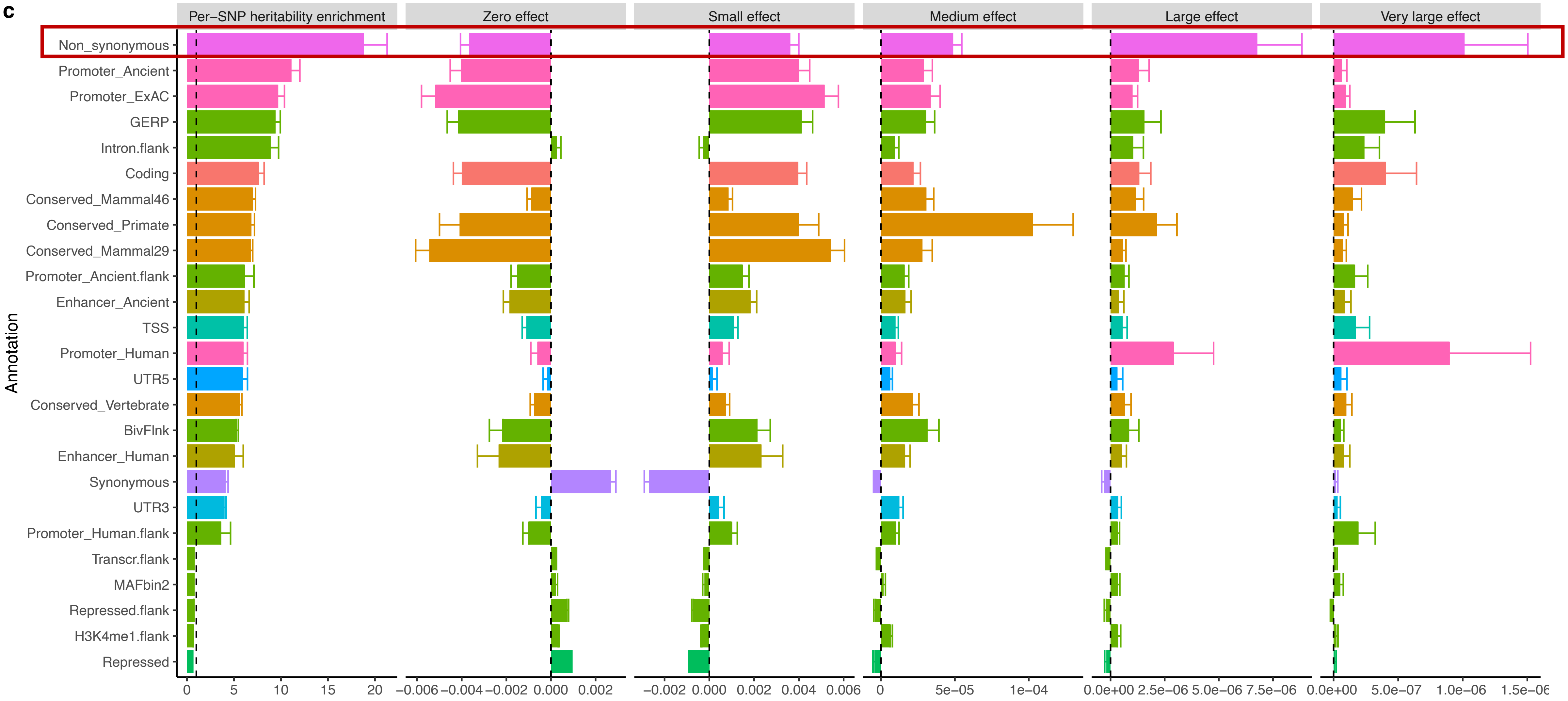
SNP markers can tag the causal variant by LD but may not tag by annotation.

Contributions of functional categories to prediction accuracy

Regions conserved across 29 mammals covers 3% genome but contributed 41% prediction accuracy!



Functional genetic architecture



Summary (3)

Methodology

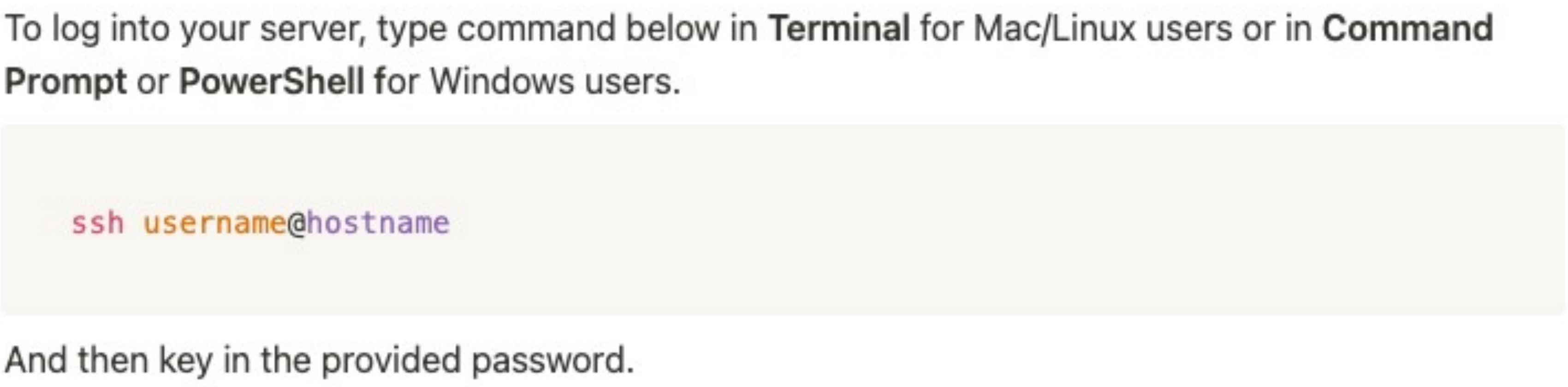
- Develop a low-rank method that fits all SNPs to better model LD (more robust & efficient).
- Incorporate functional annotations to better capture causal effects (improved accuracy).

Science

- For trans-ancestry prediction, functional annotations with genome coverage provide comparable and additive information to the use of additional GWAS dataset of target ancestry.
- Significant interaction between SNP density and annotation information, suggesting whole-genome sequence variants with annotations may further improve prediction.
- Functional partitioning highlights a major contribution of evolutionary constrained regions to prediction accuracy and the largest per-SNP contribution from non-synonymous SNPs.

Practical 5: Polygenic prediction using SBayes

https://cnsgenomics.com/data/teaching/GNGWS24/module5/Practical5_SBayes.html



To log into your server, type command below in **Terminal** for Mac/Linux users or in **Command Prompt** or **PowerShell** for Windows users.

```
ssh username@hostname
```

And then key in the provided password.