Day 5 Polygenic Prediction

Jian Zeng, j.zeng@uq.edu.au

1. **Brief overview of lectures + Q&A** (~15 min)
2. **Practical intro** (~5 min)
3. **Practical part A-C2 in break-out rooms** (~50 min)
4. **Main room discussion + Q&A** (~10 min)
5. **10 min break**
6. **Practical part C-D in break-out rooms** (~50 min)
7. **Main room discussion + Q&A** (~20 min)

Session A Tutors
Aysu Okbay
Margot vd Weijer
Joelle Pasman
Wonu Akingbuwa
Penghao Xia
Md Moksedul Momin
Yuliangzi Sun
Xuemin Wang
Tara Henechowicz

Session B Tutors
Florence Smith
Xuemin Wang
Md Moksedul Momin
Tian Lin
Yuliangzi Sun
Dinka Smajlagic

Overview of lectures

1. **Fundamentals**: understanding, limitations, applications, challenges
2. **Evaluation**: statistics for quantitative and binary traits, visualization, pitfalls
3. **Conventional methods**: clumping & P-value thresholding (C+PT), best linear unbiased prediction (BLUP)
4. **Bayesian methods**: Bayes theorem, priors, posterior inference, sumstats-based methodology
5. **MCMC sampling (optional)**: spike-and-slab model, algorithm, technical details
6. **SBayesRC**: incorporating functional annotations, low-rank modelling, application

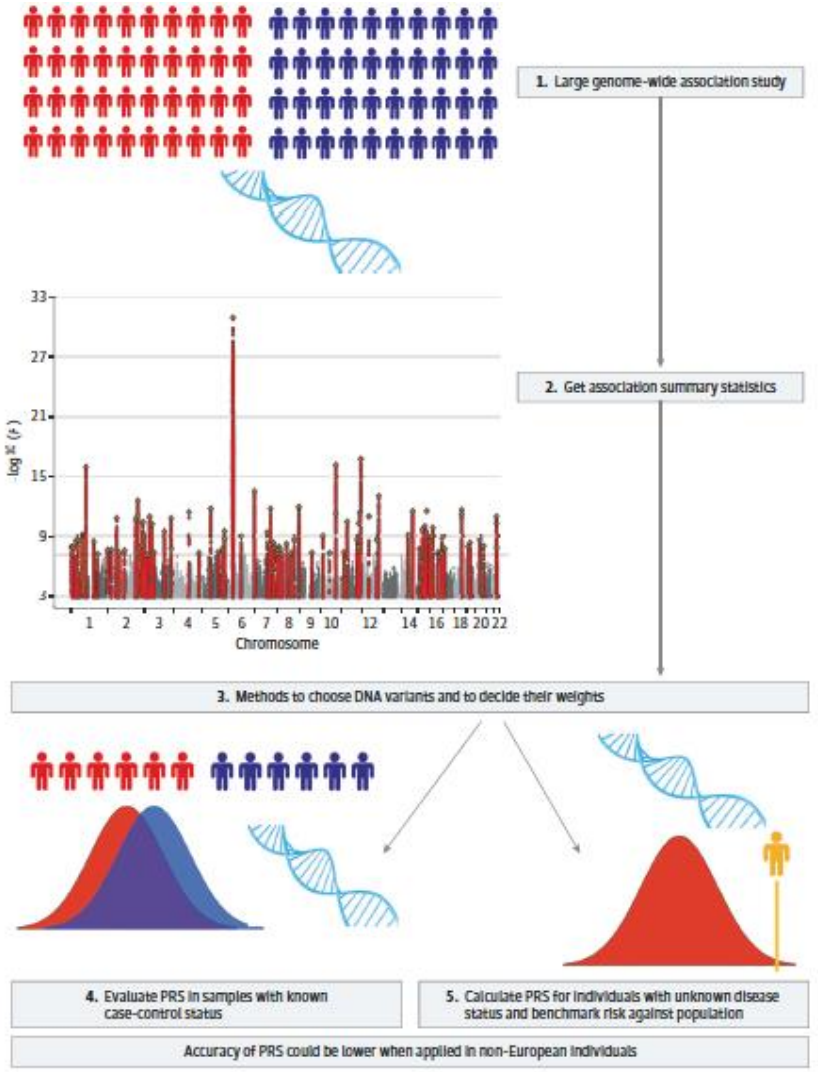
Polygenic scores

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

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



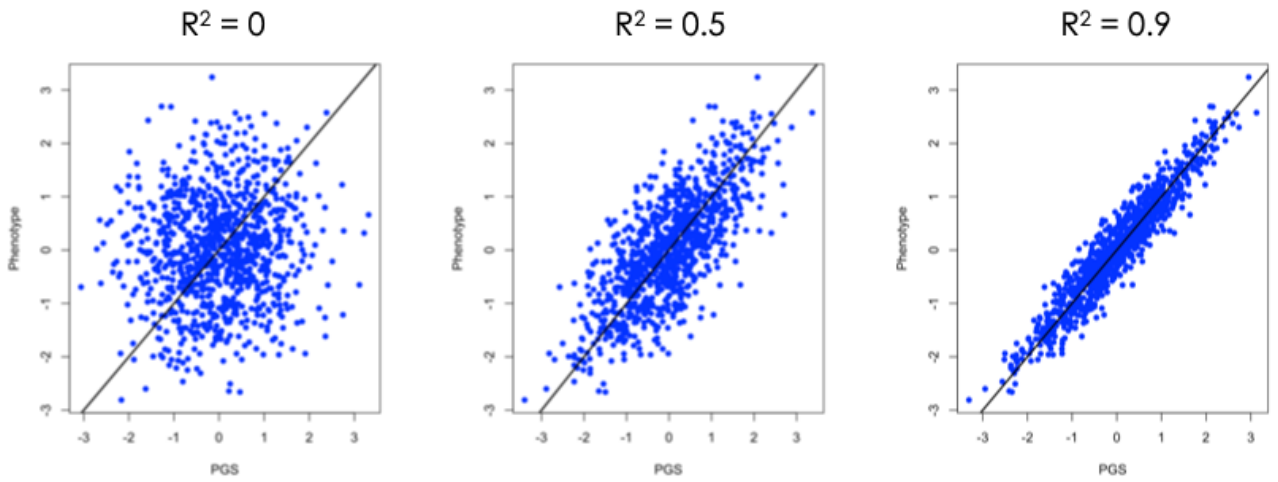
PGS evaluation in quantitative traits

Prediction accuracy

The proportion of phenotypic variance explained by PGS (prediction R^2)

It's common to adjust for covariates (sex, age, top 10 PCs, etc)

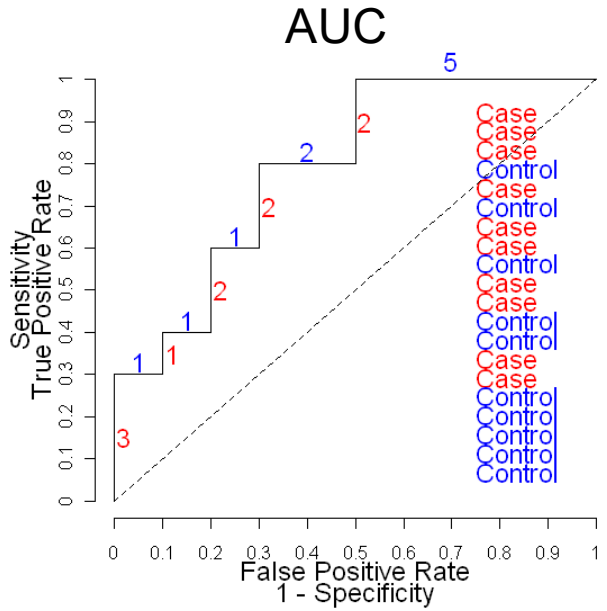
- Null model: y = covariates + e
- Full model: y = covariates + PGS + e
- Incremental R^2: $R^2_{Full} - R^2_{Null}$



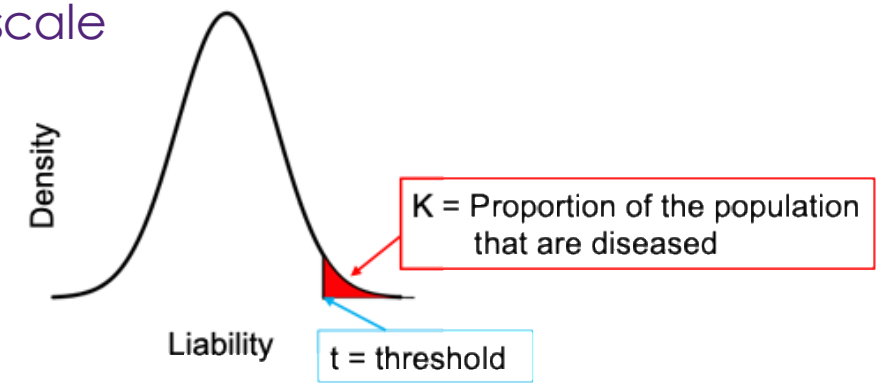
PGS evaluation in diseases (binary traits)

Statistics to measure prediction accuracy

- Pseudo R^2 from logistic regression
- AUC (area under the ROC curve)
- Variance explained on liability scale
- Risk stratification
- Decile odds ratio (OR)

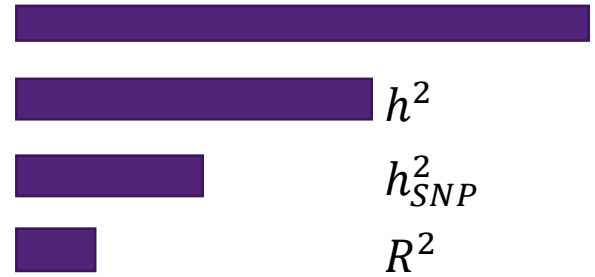


Prediction R^2 on liability scale



Limitations in prediction accuracy

- PGS have a **theoretical** upper limit dependent on the **heritability of the trait**.
- PGS have a **technical** upper limit associated with the proportion of **variance tagged** by the DNA variants measured.
- PGS have a **practical** upper limit dependent on the **sample size of the discovery sample** used to estimate effect sizes of risk alleles, and the **quality** of the discovery sample.
- PGS can be pushed closer to the technical upper limit by the **statistical methodology** used to generate the optimal weighting given to the risk alleles, and new methods integrate new biological data.



Schizophrenia

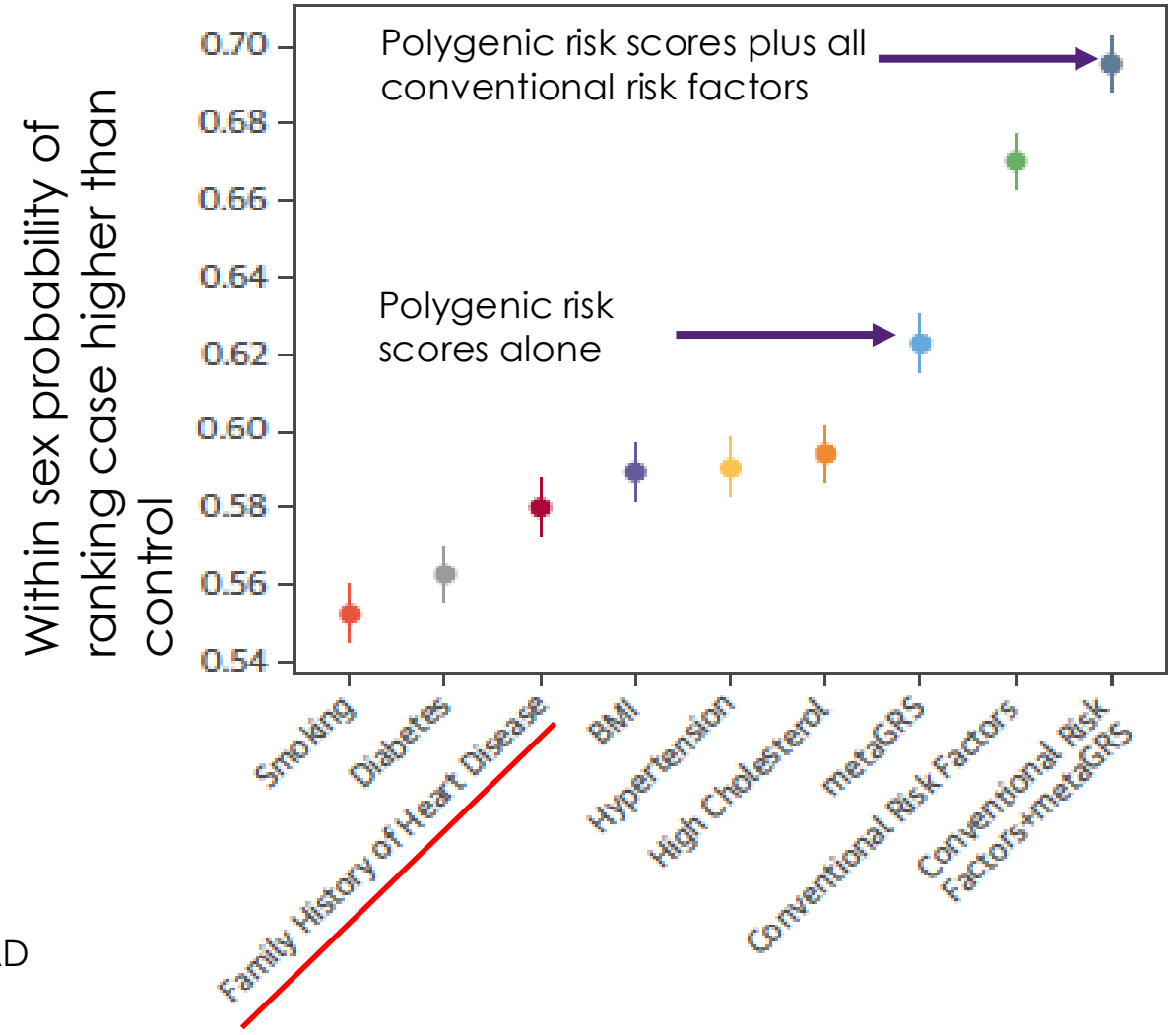
Max:
25% Liability
AUC 0.84

Current:
11% Liability
AUC 0.74

Polygenic scores cannot be highly accurate predictors of phenotypes

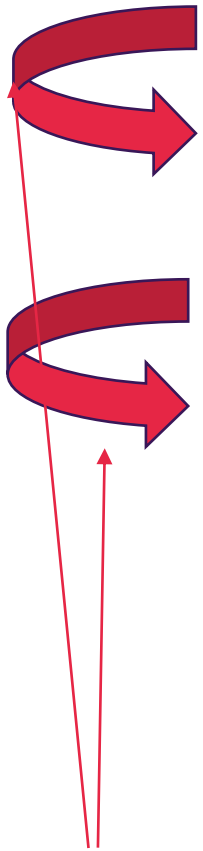
Combine PRS with conventional risk predictors

Coronary Artery Disease



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

Pitfalls of polygenic prediction



- **<u>Discovery</u>/Training/Derivation**
  - Estimate the effect sizes ($\hat{b}$) of SNPs on a trait (y) – GWAS
- **<u>Tunning</u>/Validation**
  - Further estimate some parameters (depends on methods; not all methods require it)
- **<u>Target</u>/Testing/Validation**
  - Build a polygenetic risk score (PRS) ($\hat{y}$):
  - Evaluate the prediction performance/accuracy

Should be independent; no overlap; out-of-sample prediction

GENOME-WIDE ASSOCIATION STUDIES — OPINION

Pitfalls of predicting complex traits from SNPs

Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard and Peter M. Visscher

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)

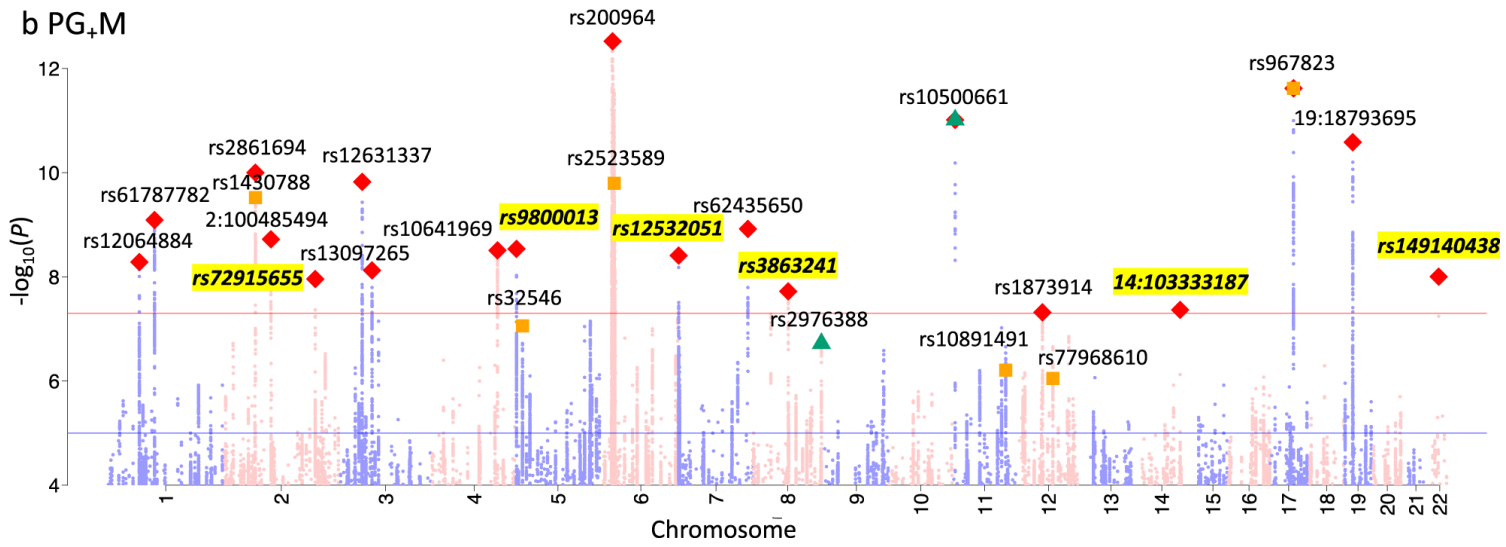
Include only the most strongly associated SNP from each LD block (Purcell et al., 2009)

Whole-genome regression approaches

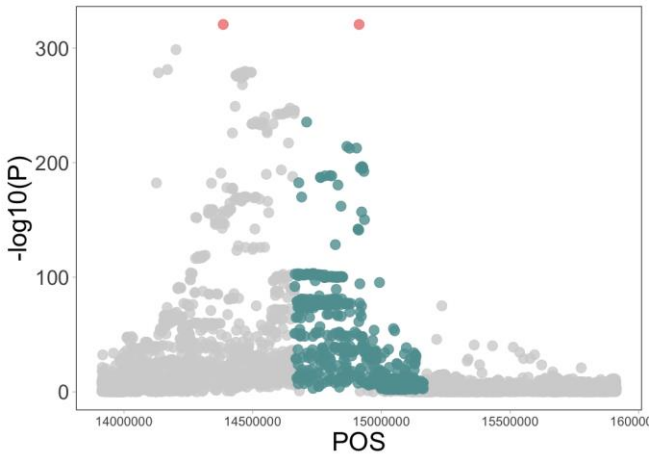
Include all SNPs but adjust the effect sizes for LD.

BLUP, Bayesian methods

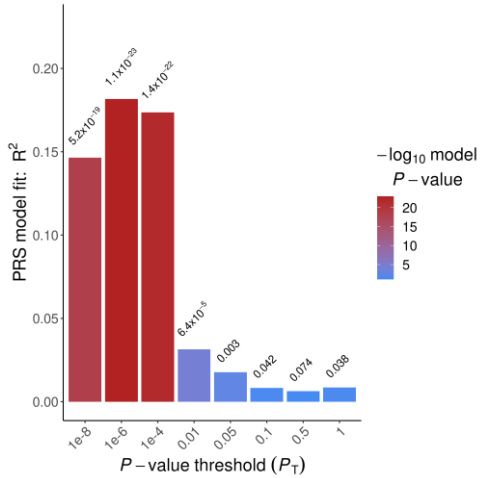
Clumping & P-value thresholding (C+PT, or P+T, C+T)



Step 1. Select most associated SNP in tower (LD-based clumping)



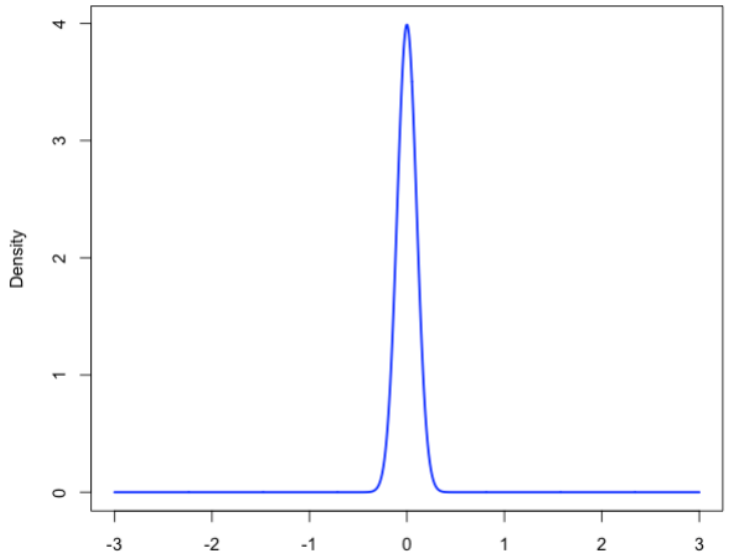
Step 2. Select on a p-value threshold in an independent tuning sample



Best Linear Unbiased Prediction (BLUP)

- Based on a linear mixed model

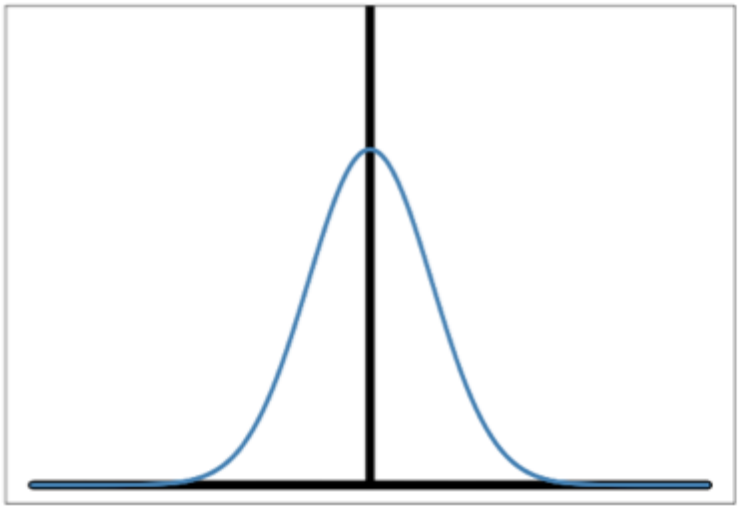
- Assumes SNPs effects are (*infinitesimal* model):
  - **all non-zero**
  - **all very small**
  - normally distributed



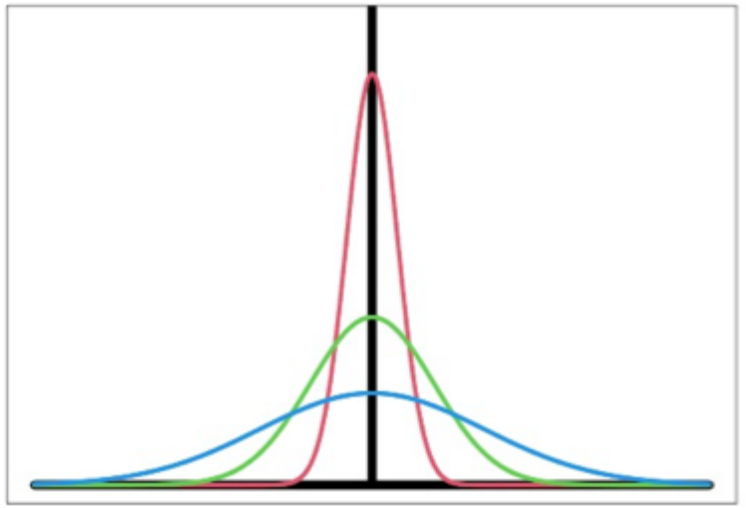
Bayesian methods

- Bayesian methods can estimate all parameters including SNP effects simultaneously
- Allow alternative assumptions regarding the distribution of SNP effects

BayesC



BayesR



Posterior inference on SNP effects

Using Bayes theorem, the posterior distribution of SNP effects

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

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

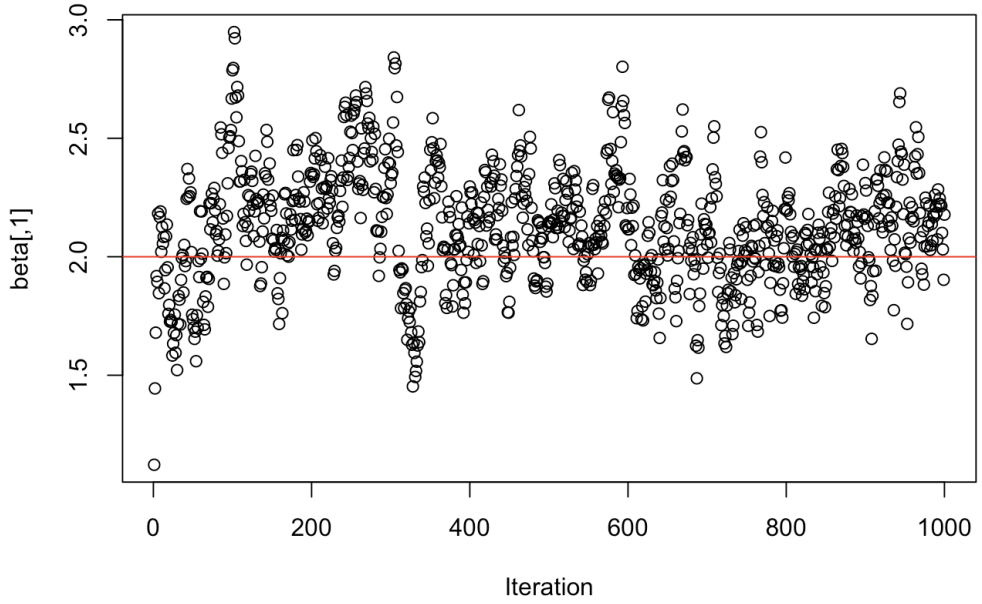
MCMC sampling

- 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 $\boldsymbol{\beta}$

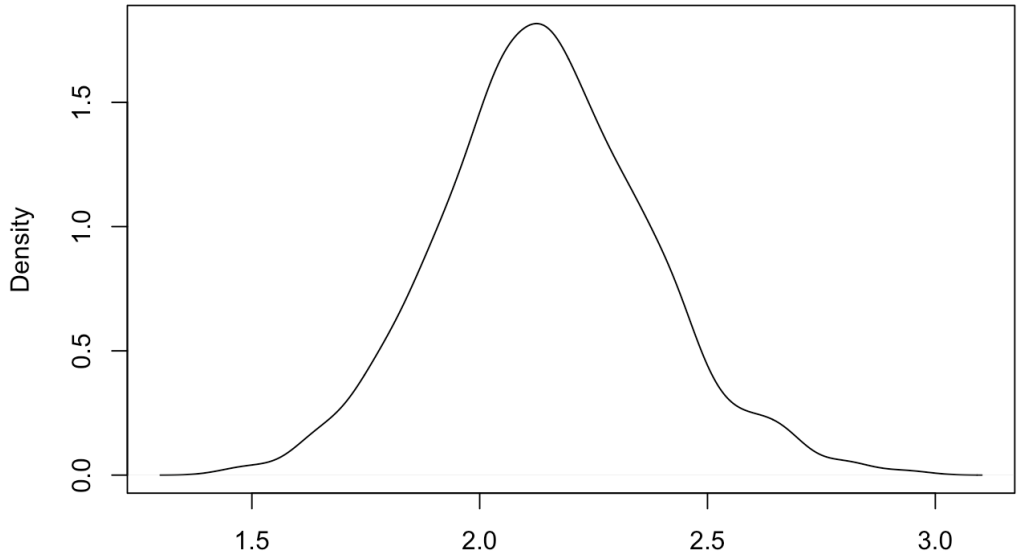
Samples reconstruct posterior distributions of parameters

MCMC sampling

Trace plot



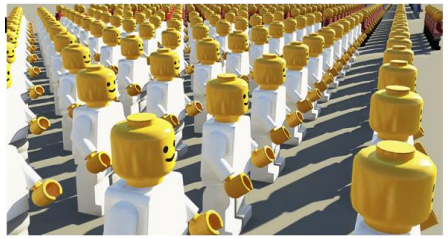
Posterior distribution

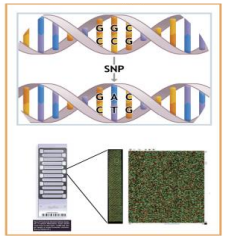


Posterior mean is used as the point estimate of the SNP effect

Individual-level model

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



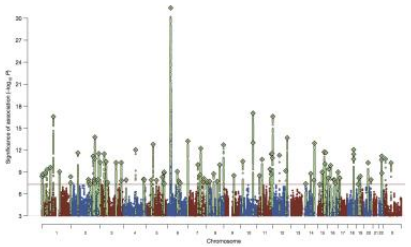


BLUP

Bayes

Summary-level model

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

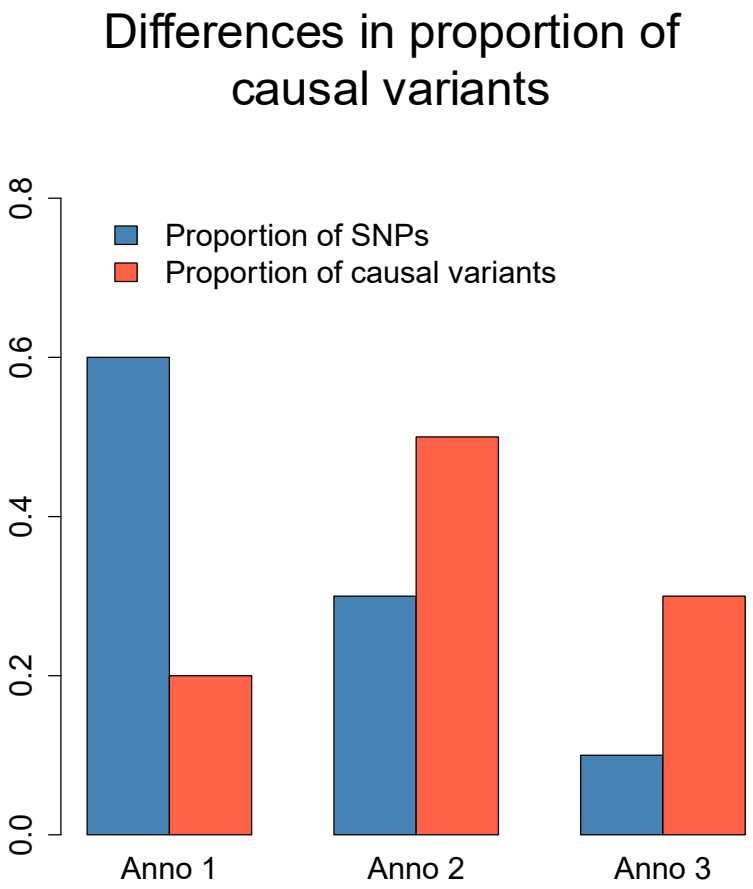


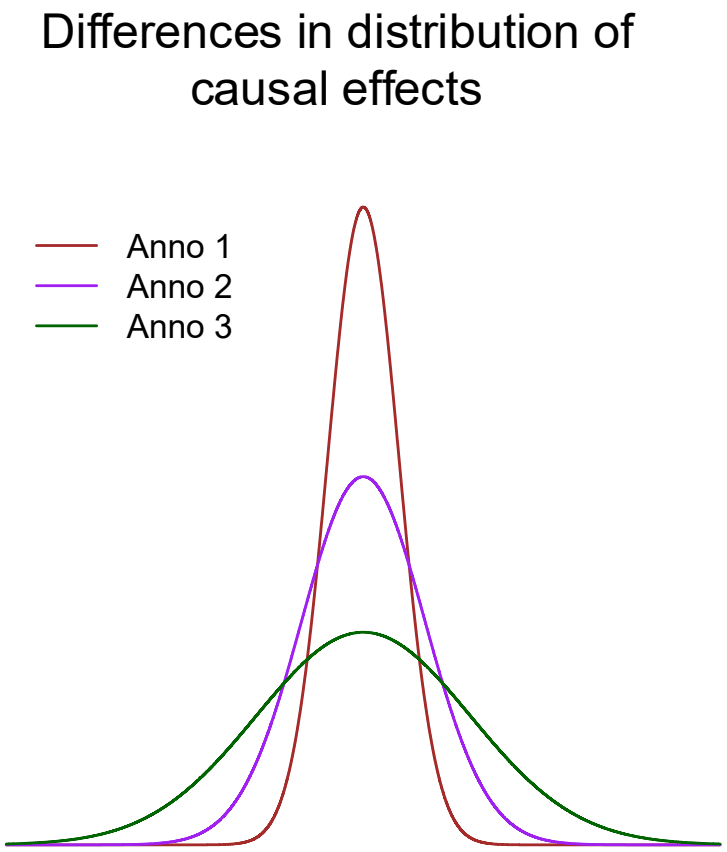
SBLUP

SBayes

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

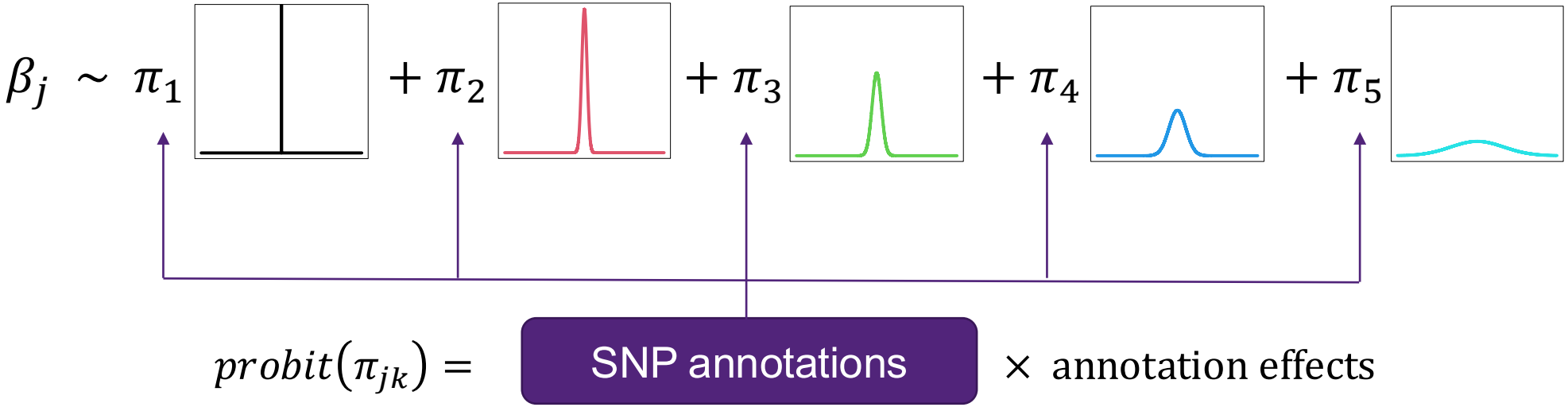
Functional annotation information

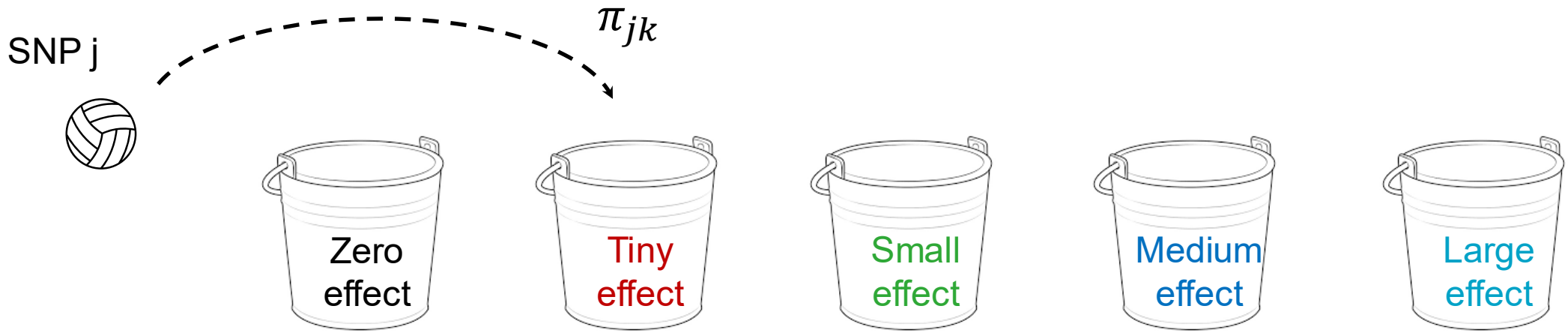




SBayesRC

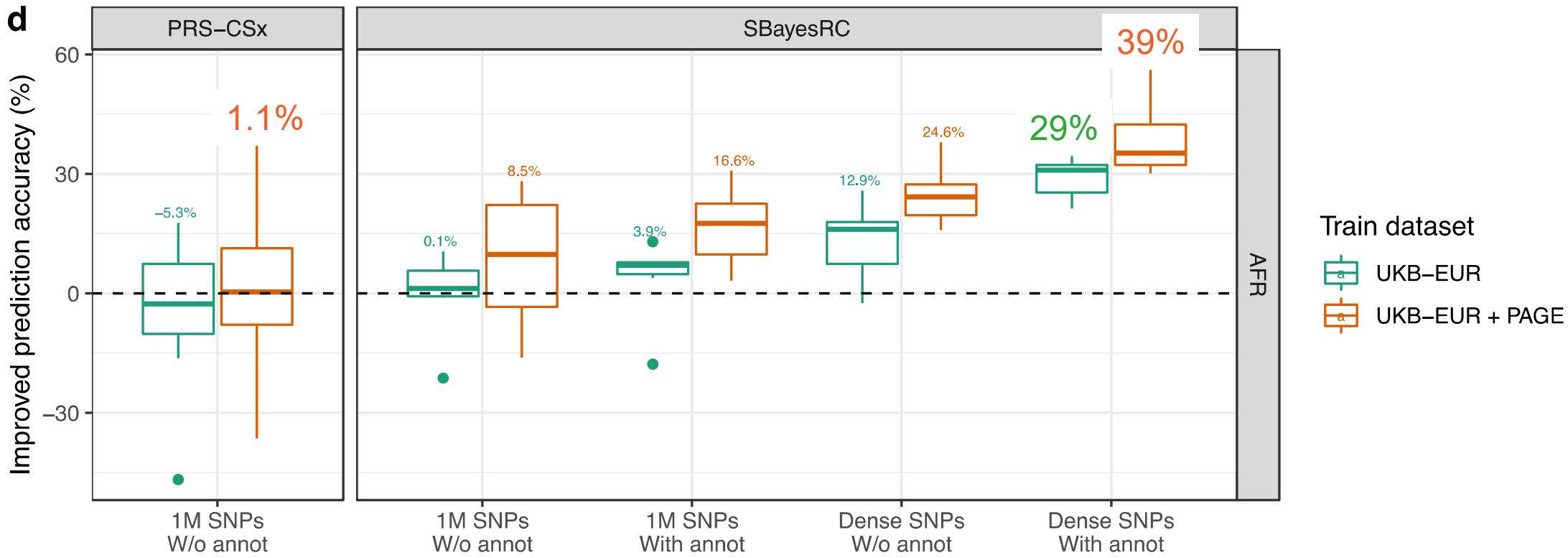
Incorporate functional annotations through a hierarchical prior:





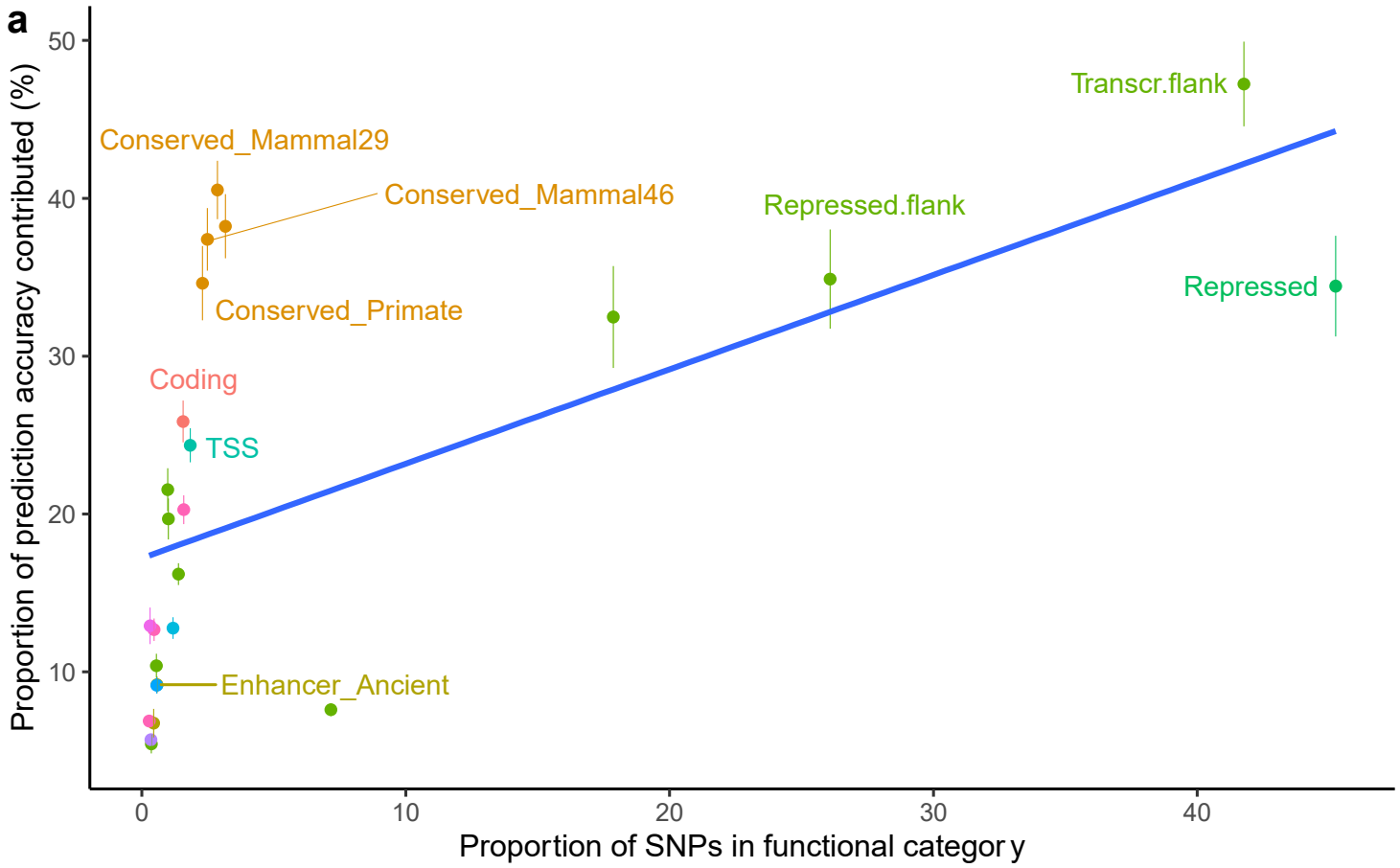
Trans-ancestry prediction

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



Contributions of functional categories to prediction accuracy

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



Summary

- Polygenic scores are imperfect but useful genetic predictors, with the prediction accuracy limited by heritability, SNP set, and sample size.
- Pitfalls in the prediction analysis – non-independence of discovery & target samples.
- C+PT (clumping and P-value thresholding) is a conventional and commonly used method to calculate PGS.
- BLUP and Bayesian methods are more morden approaches that utilise all SNPs and can incorporate additional information by priors.

Any questions?

Where is the practical?

Prediction

Overview

Summary

From fundamentals to cutting-edge methods, this session explores how polygenic scores are constructed, evaluated, and improved using conventional and Bayesian approaches, including MCMC and functional genomic annotations.

- Originally presented: June 8, 2026
- Lead: Jian Zeng
- Topics: prediction theory, polygenic risk scores, SBayesRC

Lectures

Polygenic prediction (110 minutes required, 22 minutes optional)

Practicals

The Qualtrics page for the practical.

Online Lectures and Practicals

2026 International Statistical Genetics Workshop

Syllabus

- Day 0, Before the start of the course
- Day 1, June 1, Background
- Day 2, June 2, Working with data
- Day 3, June 3, Modeling genetic and environmental components
- Day 4, June 4, LMM GWAS
- Day 5, June 8, Prediction
- Day 6, June 9, Multivariate Concepts
- Day 7, June 10, Mendelian randomization
- Day 8, June 11, Biological Interpretation

2025 International Statistical Genetics Workshop

ISG International Scholar and Cultural Exchange Program

Welcome to the 2026 International Statistical Genetics workshop - Polygenic score (PGS) practical

We recommend using Chrome browser to open this practical. (OnDemand)

This note is written so you can follow it in your breakout group without needing extra oral explanation. If something is unclear, use the hints in the questionnaire, ask in chat, or use the **Ask for help** button.

Scope of this document:

- Part A is C+PT with PRSice.
- Part B is threshold overfitting and a tuning / held-out split.
- Part C is SBayesR with GCTB on the same data as Part A.
- Part D is SBayesRC with simulated annotations, plus an optional bonus task on SNP mismatch.

Rough pacing (~2 h total, flexible):

- Part A (C+PT) ~ 20 min
- Part B ~ 20 min
- Part C (SBayesR) ~ 40 min
- Part D (SBayesRC) ~ 30 min.

Setup

Computing environment

This practical uses the ISG workshop server. Open your browser and go to:

https://workshop.colorado.edu/

You will need a **Terminal** for bash and PLINK commands, and a **RStudio** for plotting. You can either use the separate terminal application, or the Terminal tab inside RStudio. Both work the same way.

Please request 4 cores (default 2 cores) for terminal!

Using 2 cores is also fine. Would just take 2-3 more munites to run.

Notes: in this practical, scripts run in a Terminal are shown in black background and R scripts shown in gray background.

Data

First, copy the data files to your own working directory. In the Terminal, run:

TERMINAL

```
mkdir PGS
cd PGS
cp -r /faculty/jian/2026/PGS_practical/* .
```

Q2: Locate the row with the largest `R2`. Which `Threshold` is it? How many `Num_SNP` are in that score? Then discuss why R^2 across P-value thresholds is often non-monotone, and what factors might shape that pattern in real data.

	Optimal P-value threashold	Number of SNPs
Results		

Write down the notes from your discussion:

←

→

C2. Impute summary statistics, then fit SBayesR

Eigen LD requires a GWAS row for every SNP in the LD reference. If any SNPs in the LD reference are missing in your summary statistics file, GCTB can impute them using LD and the SNPs that do match. The `--out` flag sets the output prefix.

TERMINAL

```
gctb --ldm-eigen ldm_eigen \
  --gwas-summary gwas.ma \
  --impute-summary \
  --thread 4 \
  --out gwas
```

You should obtain `gwas.imputed.ma`. During imputation, GCTB also does basic QC (allele matching and dropping SNPs whose per-SNP *N* differs from the median by more than three standard deviations).

On an HPC cluster: you can impute one block per job with `--block $i`.

~5 min run

Time for a break!

SNP results are also provided (sbayesr.snpRes)

When imputation is done, run SBayesR:

```
gctb --ldm-eigen ldm_eigen \
  --gwas-summary gwas.imputed.ma \
  --sbayes R \
  --thread 4 \
  --out sbayesr
```

Day 5 Polygenic Prediction

Jian Zeng, j.zeng@uq.edu.au

1. **Brief overview of lectures + Q&A** (~15 min)
2. **Practical intro** (~5 min)
3. **Practical part A-C2 in break-out rooms** (~50 min)
4. **Main room discussion + Q&A** (~10 min)
5. **10 min break**
6. **Practical part C-D in break-out rooms** (~50 min)
7. **Main room discussion + Q&A** (~20 min)

Session A Tutors
Aysu Okbay
Margot vd Weijer
Joelle Pasman
Wonu Akingbuwa
Penghao Xia
Md Moksedul Momin
Yuliangzi Sun
Xuemin Wang
Tara Henechowicz

Session B Tutors
Florence Smith
Xuemin Wang
Md Moksedul Momin
Tian Lin
Yuliangzi Sun
Dinka Smajlagic

D1. Run SBayesRC with genomic annotations

Use the same imputed GWAS file as SBayesR: `gwas.imputed.ma`. The annotation file must include all the SNPs in the LD reference (could have more SNPs). `annotations.txt` has 13 columns: SNP ID, a column of ones for all SNPs (intercept), then 11 annotation indicators (0/1).

Set `--sbayes RC` and pass the annotation file. The run usually takes longer than SBayesR. For the interest of time, we set a shorter chain length for this practical data, which is sufficient giving the highly informative annotation.

~5 min run

Use a shorter chain if it takes too long

SNP results are also provided (sbayesrc.snpRes)

```
gctb --ldm-eigen ldm_eigen \
    --gwas-summary gwas.imputed.ma \
    --annot annotations.txt \
    --sbayes RC \
    --chain-length 1000 \
    --burn-in 400 \
    --thread 4 \
    --out sbayesrc
```

Discussion: while SBayesRC is running, discuss what would be the potential benefit of incorporating these annotations into the prediction model?

Hopefully we have achieved these goals:

By the end of this practical, participants should be able to:

- Explain in simple terms what a polygenic score (PGS) is and how to calculate it for an individual.
- Run PRSice for clumping and P-value thresholding (C+PT) on provided data and understand the main output.
- Describe why choosing the “best” threshold in the same sample you evaluate can be misleading, and how to address.
- Run SBayesR in GCTB using GWAS summary statistics and an LD reference, and compare target-sample prediction accuracy to C+PT.
- Run SBayesRC with a simulated annotation file and compare results to SBayesR.
- (Optional) Use GCTB to map predictor effects onto SNPs present in the target when many SNPs in the predictor list are missing from the target genotypes, then calculate PGS.

Genetics & Genomics Winter School

July 6 - 10, 2026 | Brisbane, Australia

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Statistical and Computational Methods

Statistical Genomics 1 Genetic Mapping

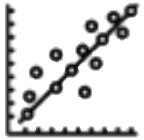
Dr Kathryn Kemper

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

Statistical Genomics 2 Heritability Estimation

Prof Loic Yengo

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



Statistical Genomics 3 Polygenic Prediction

Dr Jian Zeng

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

Cellular Transcriptomics

A/Prof Quan Nguyen

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

Genetic Epidemiology

Dr Daniel Hwang

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

Systems Genomics & Pharmacogenomics

A/Prof Sonia Shah

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

On-site lectures + hands-on practical exercises

Construct your own week-long course

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

Each class size limited to 60 participants

Special Seminar, Social Events, HPC & Lab tour

Scholarships available for undergraduate students

Registration opens on 7th April