Testing for Associations Using Latent Factor Mixed Models

Kevin Caye[1], Olivier Michel[2], Olivier Francois[1]

[1] TIMC-IMAG, [2] GIPSA-lab

2017-03-01

Outline

Genotypic Data

- single nucleotide polymorphism (SNP): single nucleotide variation occurring commonly within a population.
- Data are matrix of size L loci for n individuals ($L \sim 10^6$ and $n \sim 10^3$)

	chr: 1 pos: 657	chr: 1 pos: 3102	chr: 1 pos: 4268
02B6	0	1	1
DralV 1-16	1	1	1
UllA 2	1	0	1
Zdr-1	1	1	1

Table 1: Sample of a Arabidopsis Thaliana SNP matrix of size $n = 1095$ individual and $L = 214\ 051$ loci.

Environmental Association Study

Data

- G the genotype matrix of size n individuals and L loci.
- X the covariable: environmental information (temperature, location, ...)

We want to detect signs of natural selection and loci involved in local adaptation. These loci should be correlated with X.

Example: Arabidopsis Thaliana Dataset

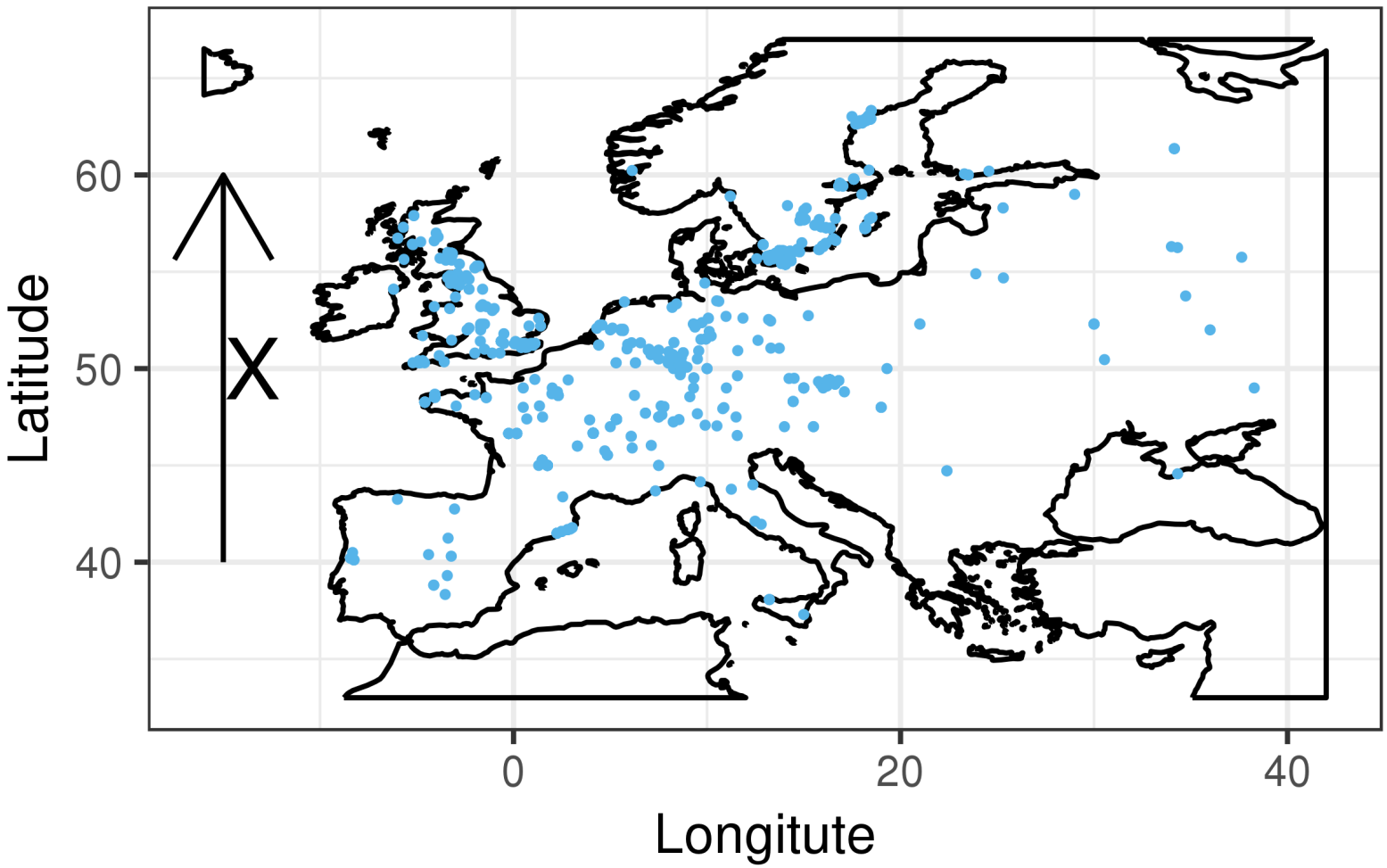


Figure 1: Spatial coordinates for each Arabidopsis Thaliana individual.

Multiple Hypothesis Testing: Simple Linear Regression

Linear model for each locus j

$$G_{.j} = XB_j + E_j$$

where

- B_j is the unknown regression coefficient
- E_j is the residual error

We want to detect loci where we can reject the hypothesis

$$H_0 : B_j = 0$$

.

We compute z-score for each locus j:

$$z_j = \frac{\hat{B}_j}{\hat{\sigma}_j}$$

where $\hat{B}_j$ is an estimation of B_j and $\hat{\sigma}_j$ the estimation its standard deviation.

Example: Arabidopsis Thaliana Dataset

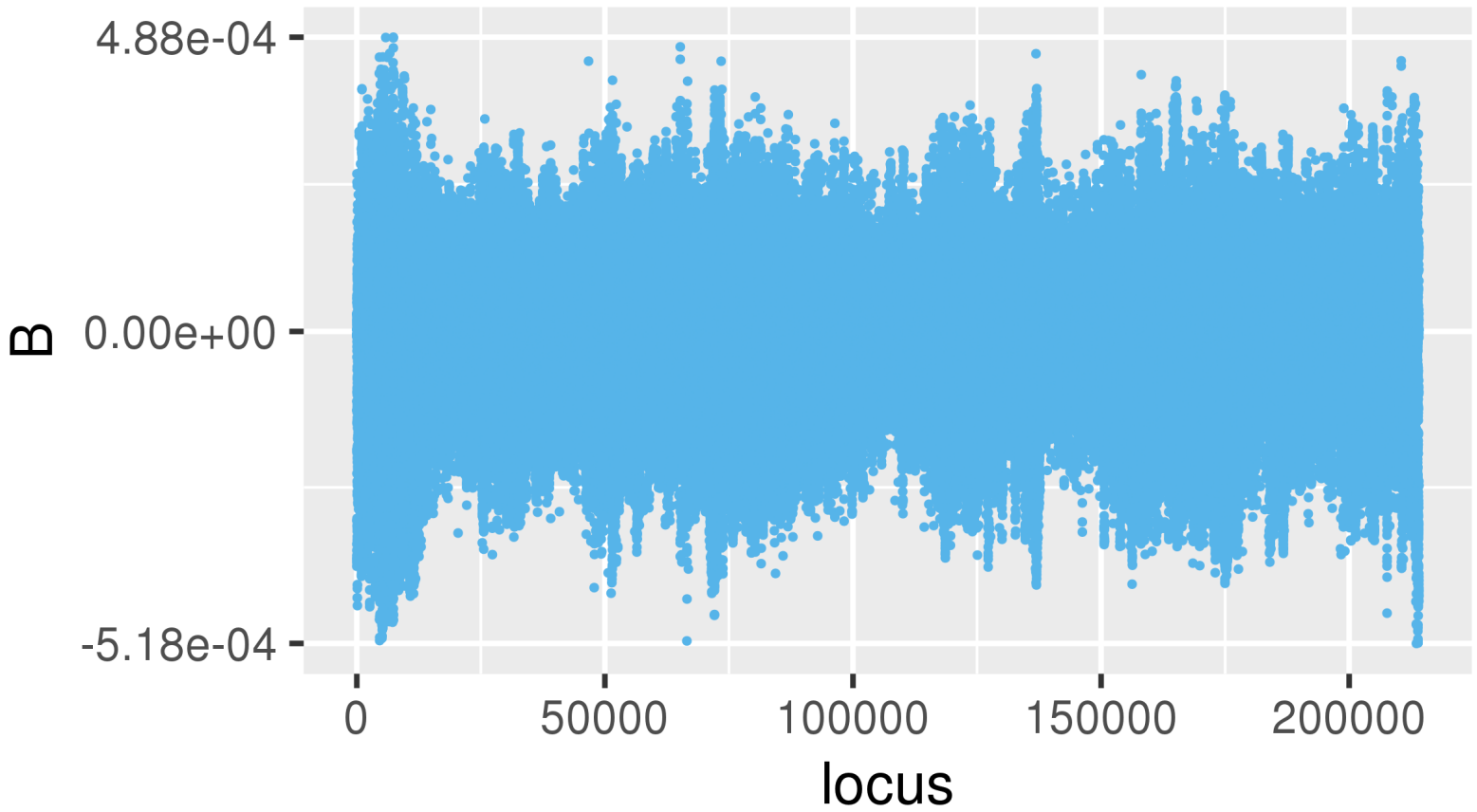


Figure 2: B regression coefficient computed with a simple linear model $G_{.j} \sim X$.

False Discovery Rate Control

We want to identify a small percentage of interesting cases that deserve further investigation.

False Discovery Rate Control:

We search a list $\Gamma = \{1, .., J\}$ such that

$$p(B_j = 0 | j \in \Gamma) = T$$

where T is the expected FDR threshold.

Such list can be computed using pvalue and the Benjamini-Hochberg procedure (**?**)

Example: Arabidopsis Thaliana Dataset

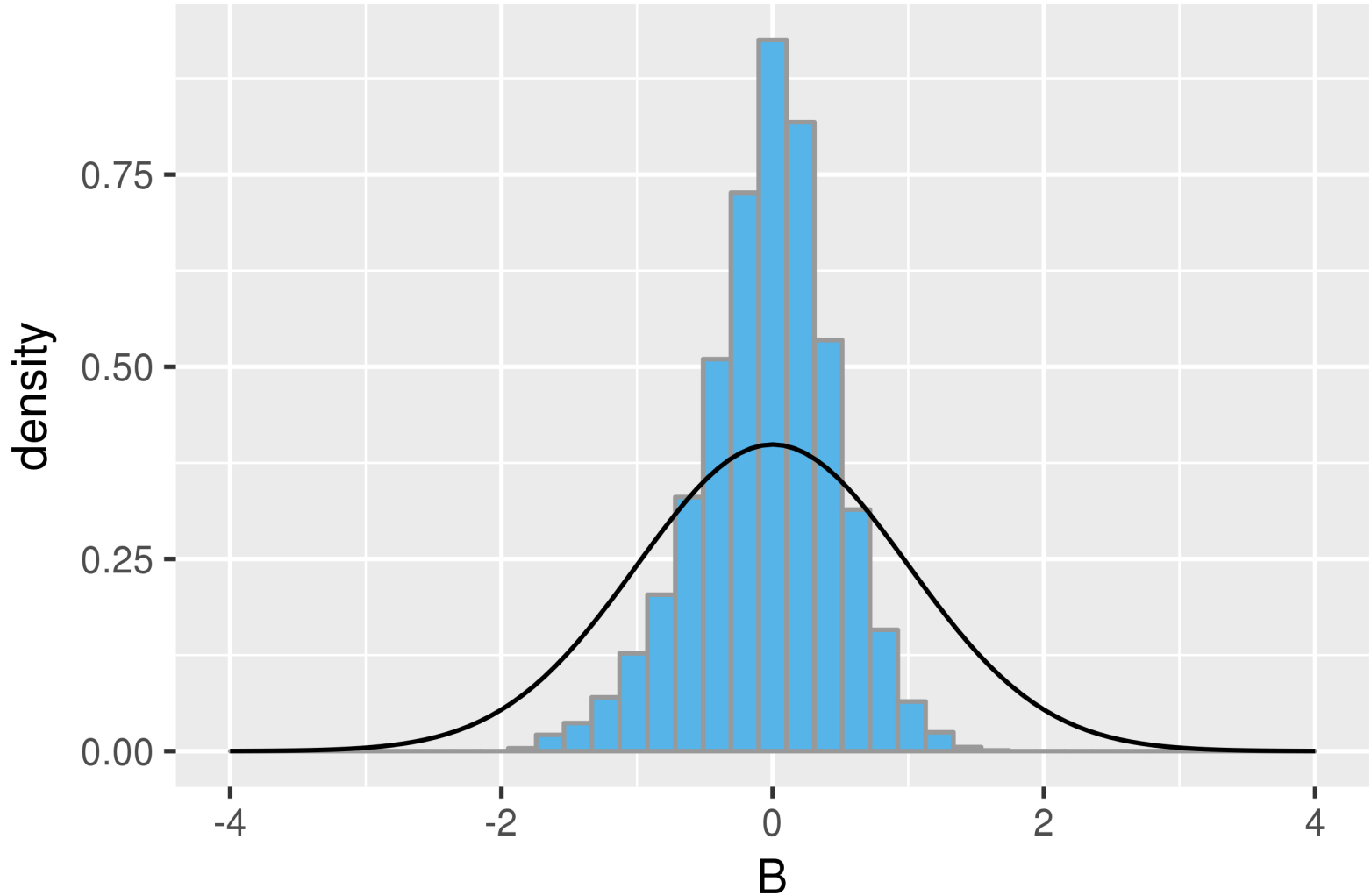


Figure 3: Zscore computed with a simple linear model $G_{.j} \sim X$ and its theorical H_0 distribution function $N(0, 1)$.

Multiple Hypothesis Testing Adjustment

Problems

- Scores are not statistically independent
- model misspecification (individual are not independent)

Literature

One can change hypothesis testing:

- genomic inflation factor (**?**)
- empirical null estimation (**?**)
- empirical-Bayes adjustments (**??**)
- ...

Example: Arabidopsis Thaliana Dataset

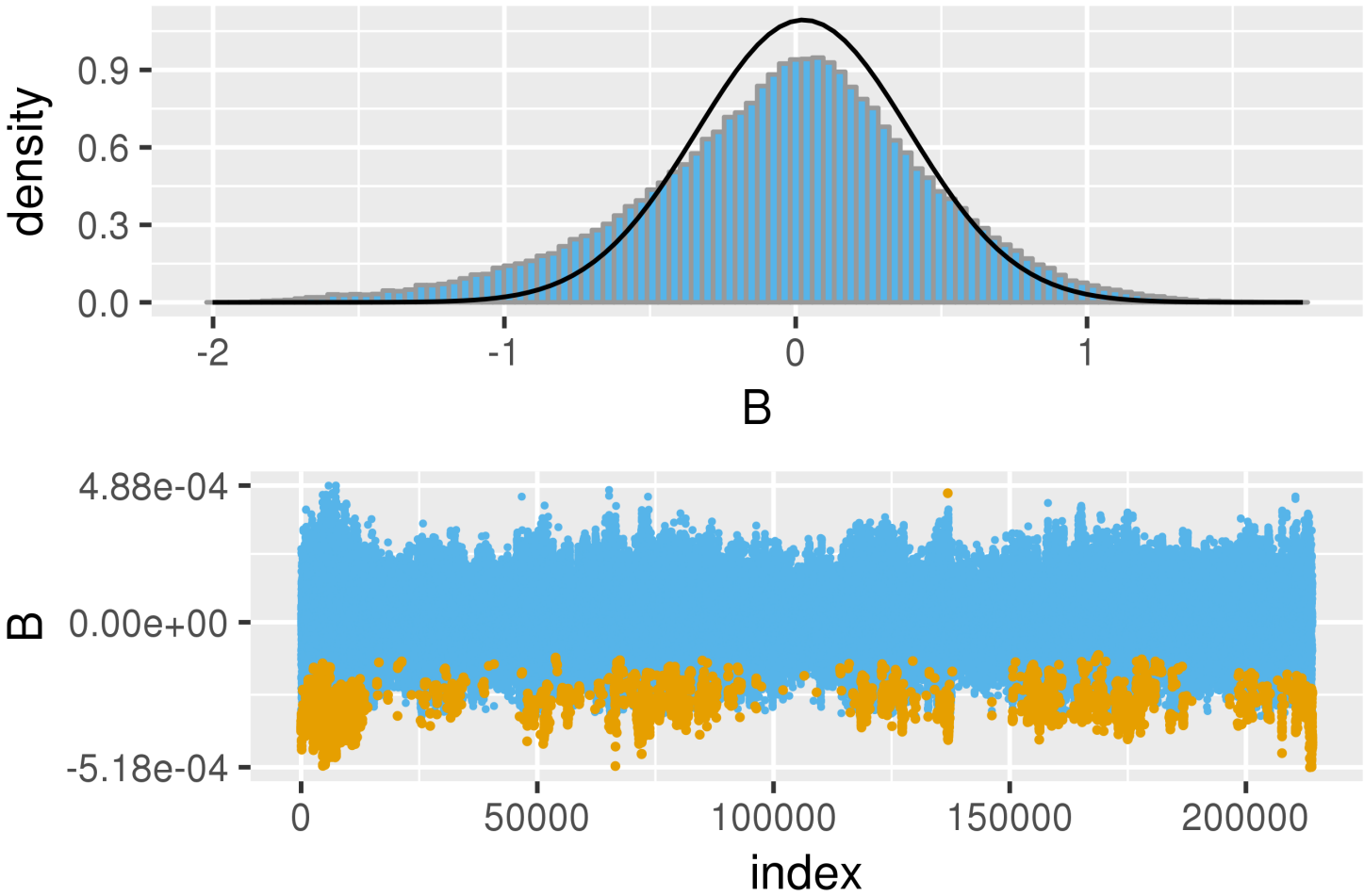


Figure 4: Simple linear model zscore and the empirical null distribution (locfdr). In yellow, significant correlated loci with a FDR threshold of 0.05.

Latent Confounding Factors

Problem

- there are latent variables correlated with X which induce spurious discovery.

Literature

One can add latent variables to the model:

- principal component analysis
- Surrogate Variable Analysis (?)
- Latent Variable Mixed Model (?)
- ...

Latent Confounding Factors

Latent variables correlated with X can induce spurious association.

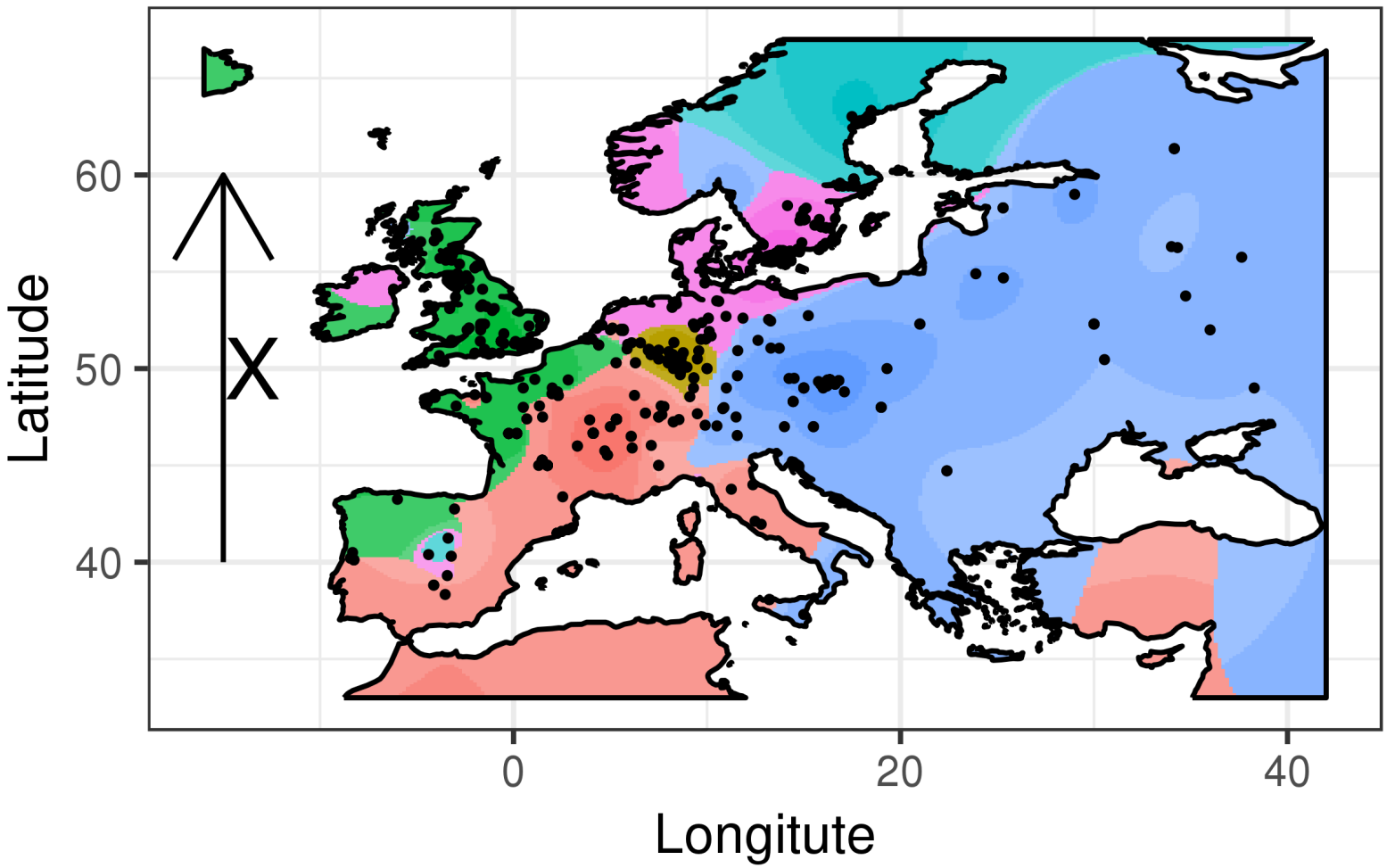


Figure 5: Each color stands for a population.

Example: Arabidopsis Thaliana Dataset

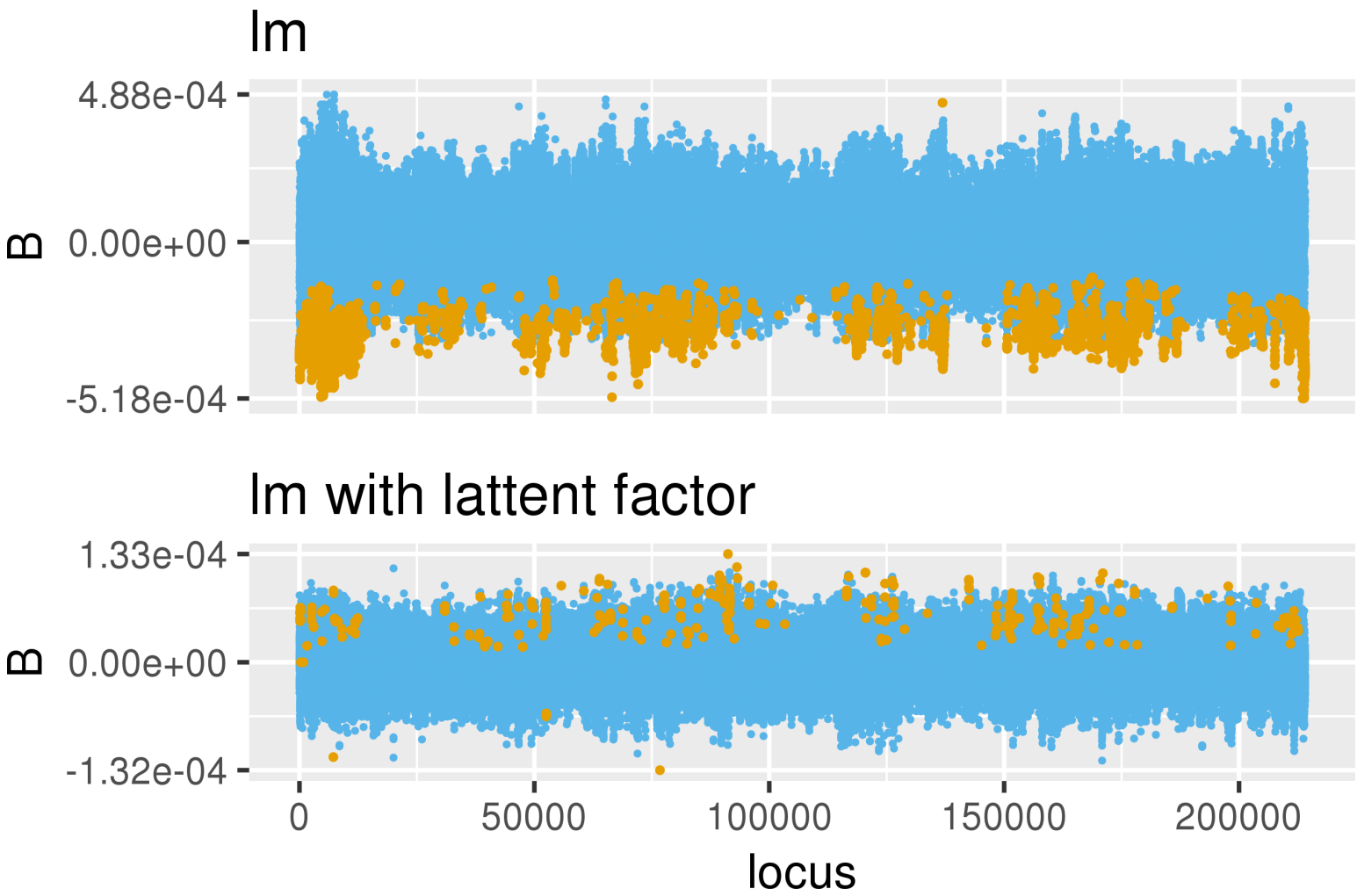


Figure 6: In yellow, significant correlated loci with a fdr threshold of 0.05.

Our Method: Latent Factor Mixed Model (LFMM) (?)

We assume that

$$G = UV^T + XB^T + E$$

where

- U is a $n \times K$ matrix of latent factor scores
- V is $L \times K$ matrix of latent factor loadings
- B is the unknown regression coefficient
- E is the error matrix

We want to find j such that $B_j \neq 0$

LFMM Estimation

optimization problem

$$min_{rk(UV^T)\leq K} \frac{1}{2}||G - UV^T - XB^T||_F^2 + \lambda||B||_2^2$$

Analytic solutions

we compute

- $P = Id - (X^T X + \lambda Id)^{-1} X^T X$
- $P = \sqrt{P}^2$

then

- $\hat{U}\hat{V}^T = \sqrt{P}^{-1} svd_K(\sqrt{P}G)$
- $\hat{B} = (X^T X + \lambda Id)^{-1} X^T (G - \hat{U}\hat{V}^T)$

LFMM Algorithm With Missing Values

When there is missing value in the dataset we use an alternated algorithm (inspired from EM algorithm for PCA **?**)

Algorithm steps

1. impute at random missing values in G
2. compute $\hat{U}$, $\hat{V}$ and $\hat{B}$
3. impute missing value with $\hat{U}\hat{V}^T + X\hat{B}^T$
4. go to set 2 if no convergence

Hypothesis Testing

Latent factor scores U can be used as covariable of regression model:

- linear model
- generalized linear model
- ...

If the hypothesis testing is not well calibrated we can use empirical bayes FDR estimation method to adjuste for finite sample, model mispicification or score correlation.
In practice, we use U in a simple linear regression model.

Model Selection

There are two parameters for LFMM method, the number of latent factor K and the regularization parameter λ.

K number of latent factor

- PCA to visualize latent variable structure
- population structure analysis
- cross-validation with LFMM algorithm with missing values

λ regularization parameter

- cross-validation with LFMM algorithm with missing values

LFMM method run are very fast when there are no missing value. The user can run with several parameters to explore different hypothesis.

Validation On Simulation

Simulated data was generated from of the 1000 genomes dataset ?. Only European individuals and the loci from the chromosome 22 was keep.
A new genotype matrix is genererated as follow :

$$G = UV^T + XB^T + E$$

Where

- U and V are loading and score matrix retrun by PCA with $K = 4$ (5 populations in the dataset)
- X is generated such that $cor(X, U_i) = c_i$
- B is such that $B_j = 0$ for neutral loci and $B_j \sim N(0, sd)$ for selected loci
- E is the residual matrice of the PCA

Comparison With Other Method

Methods similar to LFMM:

- SVA (?)
- FAMT (?)
- LEA-LFMM (?)
- refactor (?)

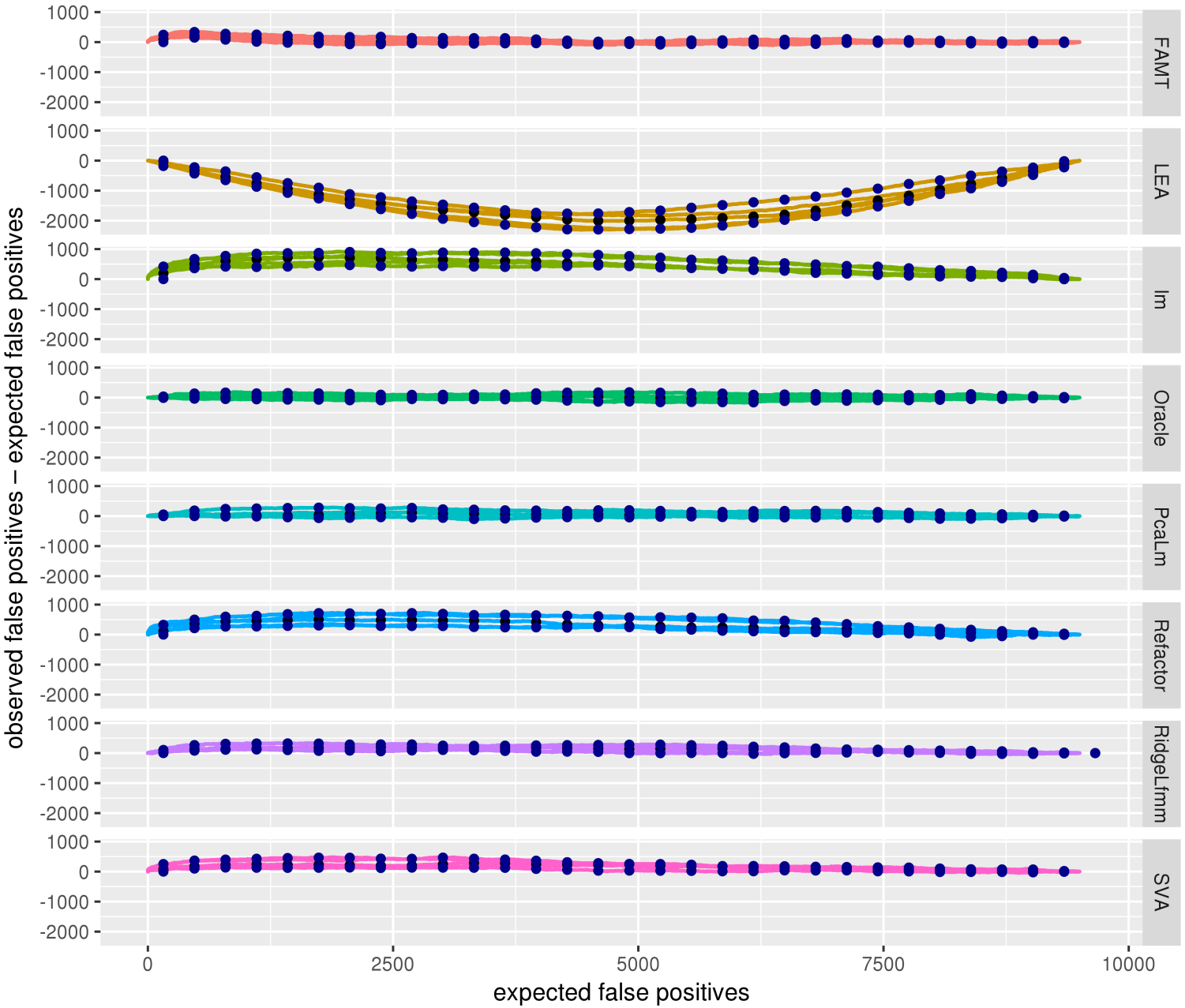
Classic methods:

- simple linear regression
- simple linear regression with PCA score

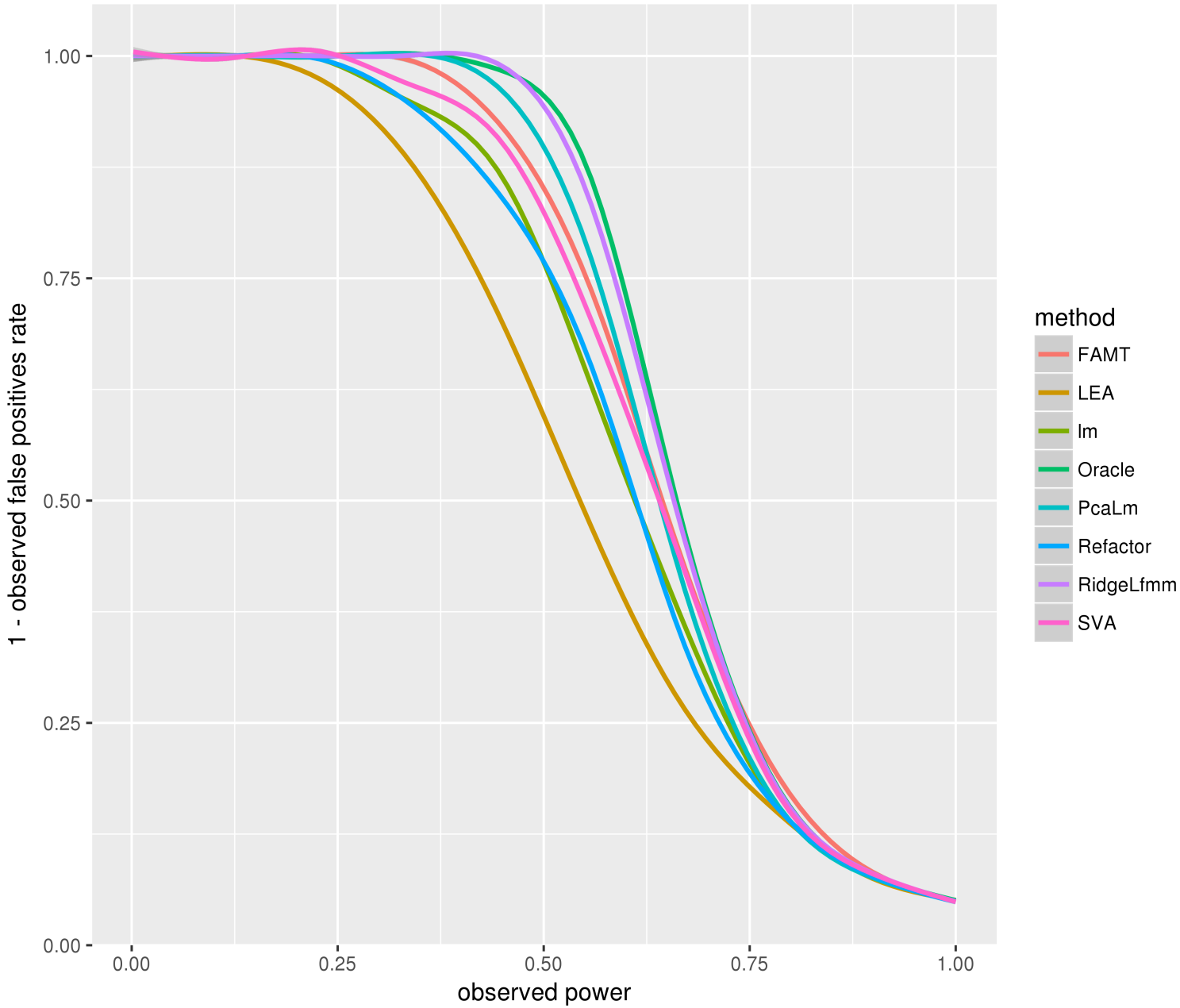
Oracle method

- simple linear regression knowing lattent factor scores

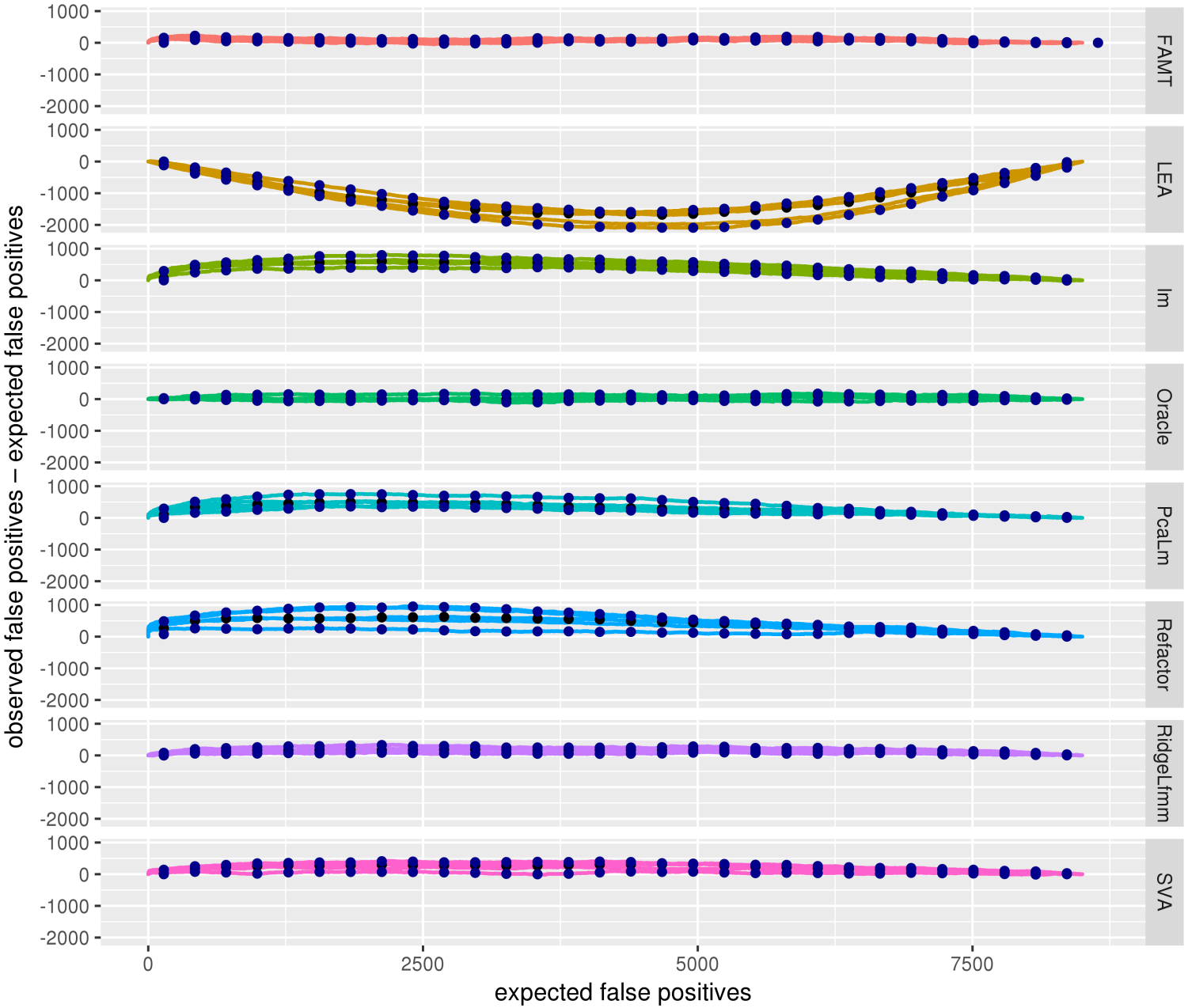
Simulation with 5% of outlier



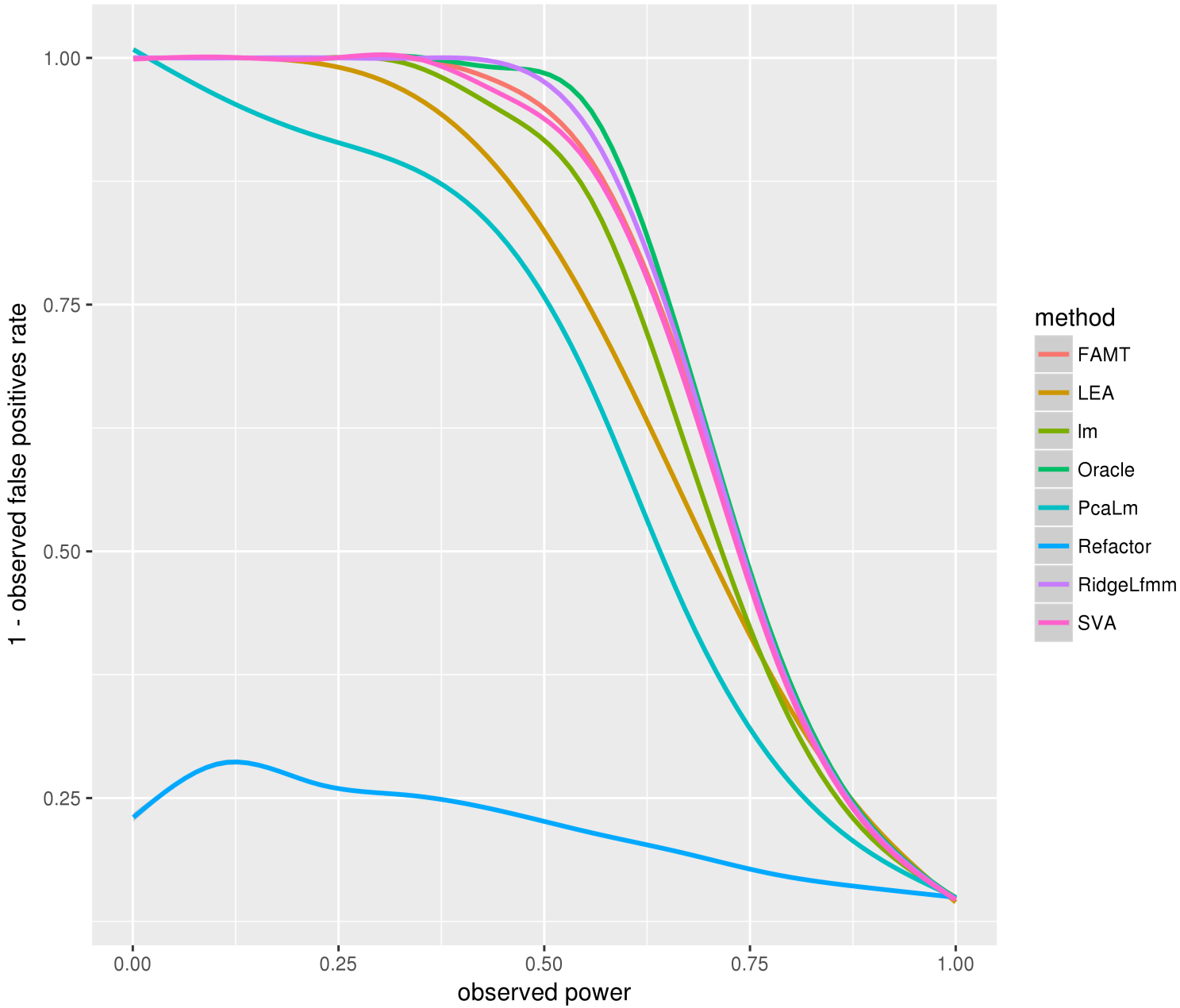
Simulation with 5% of outlier



Simulation with 15% of outlier



Simulation with 15% of outlier

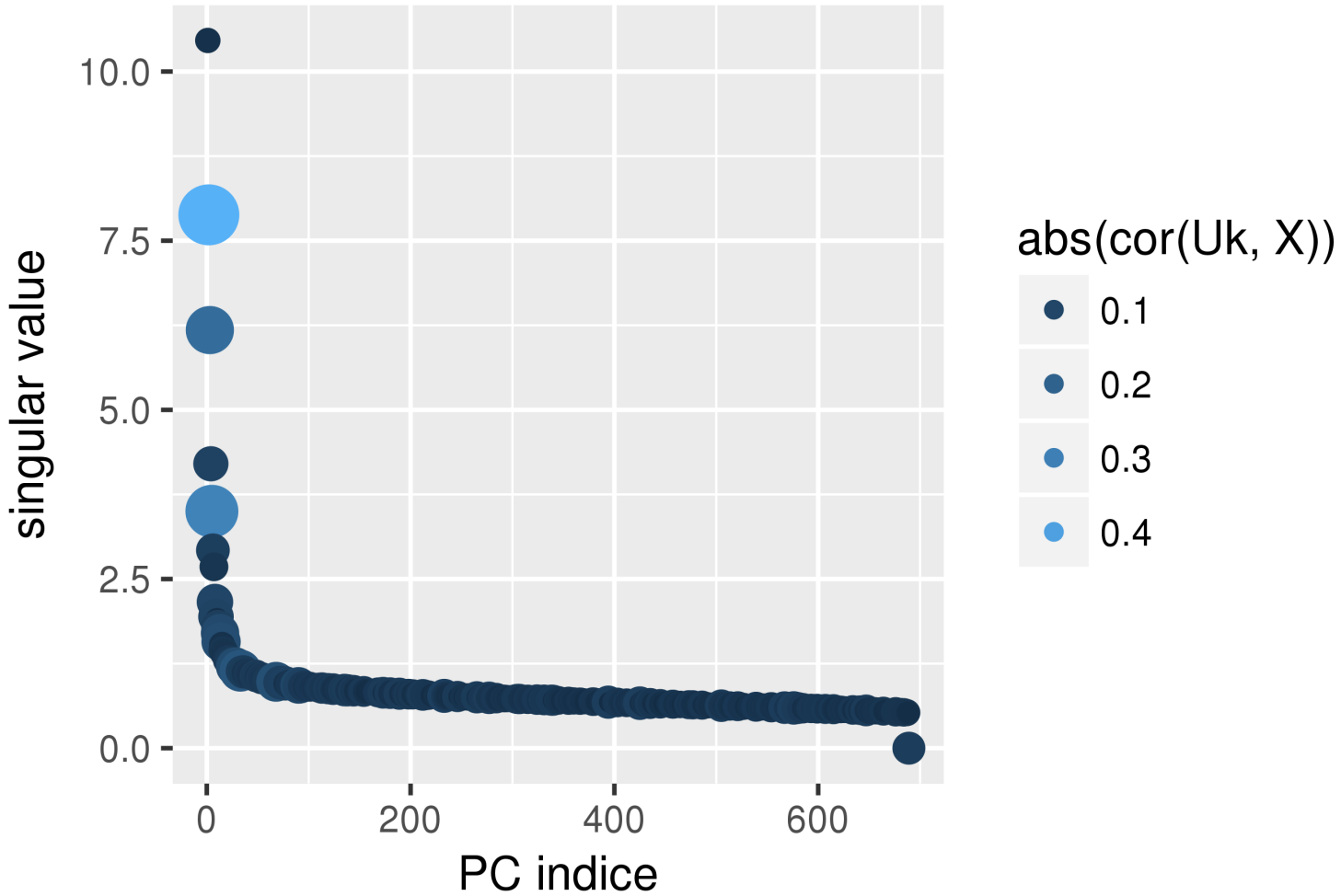


Epigenome-Wide Association Study: GSE42861 Dataset (**?**)

Association study of DNA methylation with rheumatoid arthritis.

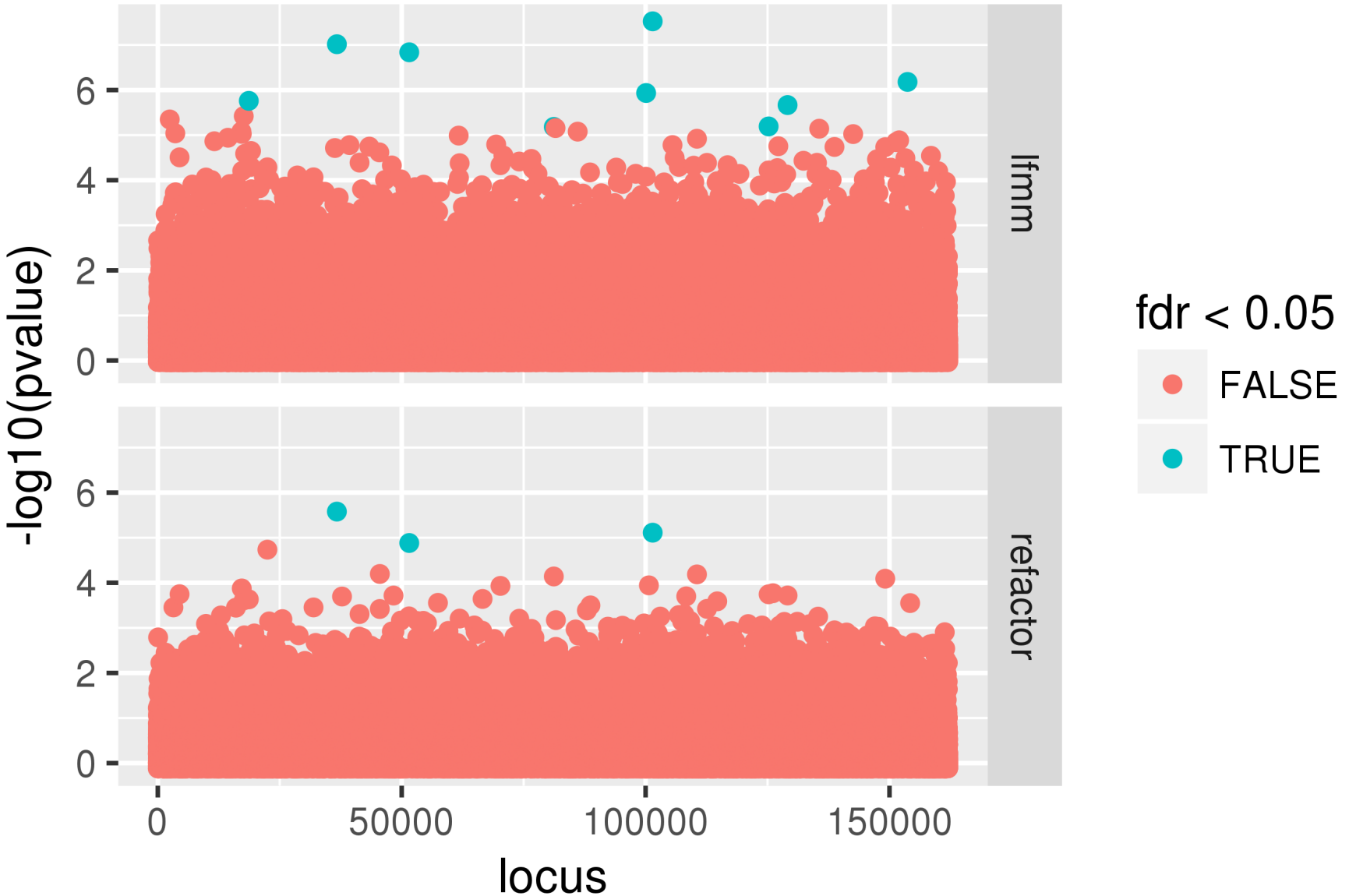
- G: contains observed beta-normalized methylation levels for 686 individuals and 103 638 loci
- X: 354 cases and 332 controls for rheumatoid arthritis

GSE42861 Dataset: Model selection



Cross-validation was too slow. . . (still running)

Result


-log10(pvalue)
6
4
2
0
lfmm
refactor
0
50000
100000
150000
locus
fdr < 0.05
FALSE
TRUE

Work In progress

- Cross validation to slow
- Testing statistic not well calibrated on generative dataset