FastGxC: Context-Specific eQTLs Detecting and Comparison of Correction Methods

Jiachen Ai
Department of Biostatistics, UCLA

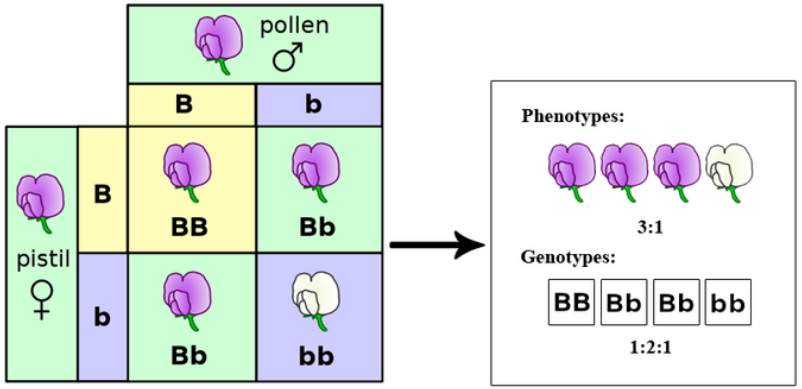
Content

- **Introduction**
  - GWAS
  - Context-specific GWAS
  - Research goal
- **FastGxC**
  - Methods
  - Simulation
  - Results
- **Comparison of correction methods**
  - Methods
  - Simulation
  - Results
- **Conclusion**
- **Discussion**

GWAS

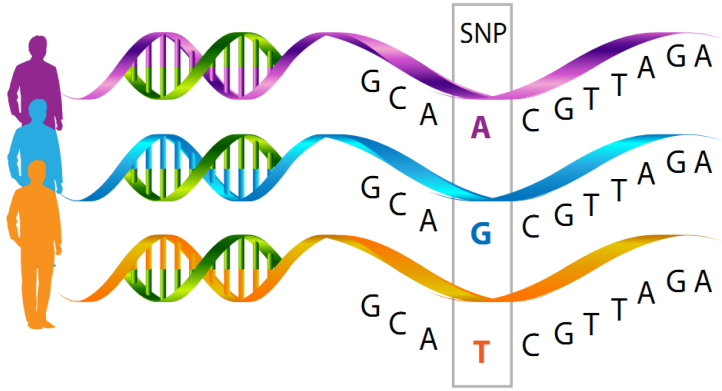
Genome-wide association studies (GWAS) is a research method used to identify **association** between a particular disease (phenotype) or genotype and genetic variations (SNPs), leading to a better understanding of the genetic basis of diseases and paving the way for new diagnostic and therapeutic approaches.

Genotype is the specific combination of alleles an individual possesses for a particular gene or set of genes (eg. BB, Bb, bb)



Y

Single Nucleotide Polymorphism (**SNP**) is a DNA sequence variation that occurs when a single nucleotide in the genome differs between individuals.



X

GWAS

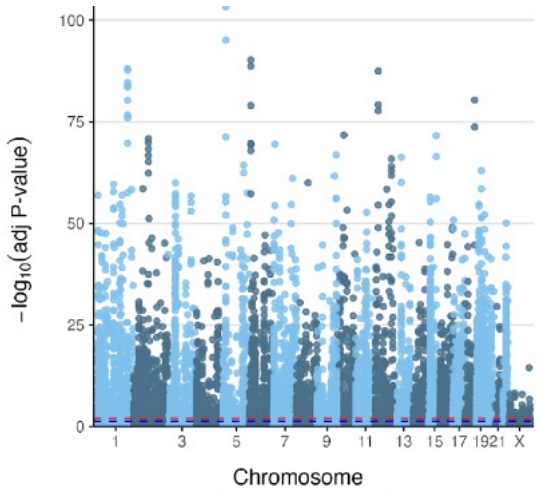
Genome-wide association studies (GWAS) is a research method used to identify **association** between a particular disease (phenotype) or genotype and genetic variations (SNPs), leading to a better understanding of the genetic basis of diseases and paving the way for new diagnostic and therapeutic approaches.

Linear Mixed Model for GWAS

$$\mathbf{y} = W\boldsymbol{\nu} + X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{e},$$
$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \quad \text{where } \mathbf{G} = 2K\sigma_a^2,$$
$$\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad \text{where } \mathbf{R} = \sigma_e^2\mathbf{I}.$$

Where:

- Y: response variable (SNPs' genotype, phenotype, etc.)
- W: covariates (population structure, etc.)
- X: fixed effects design matrix
- Z: random effects design matrix (individuals, etc.)
- $\boldsymbol{\nu}$: covariates coefficients
- $\boldsymbol{\beta}$: fixed effects coefficients
- $\mathbf{u}$: random effects coefficients(e.g., individual-specific effects)
- $\boldsymbol{e}$: residual error



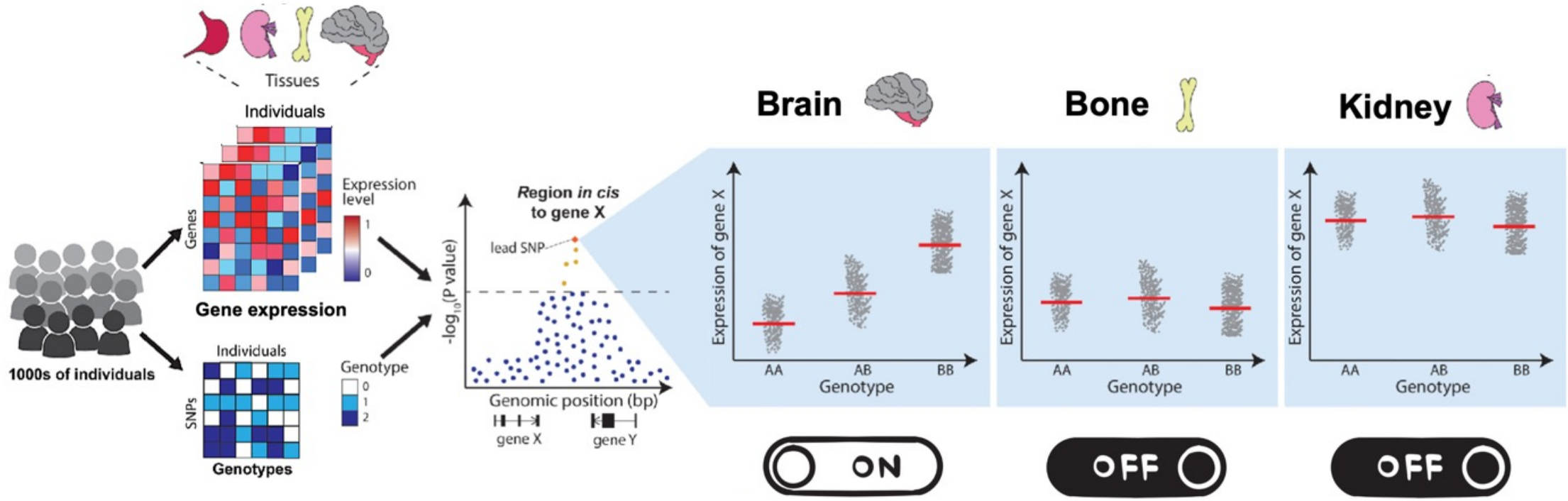
eQTL (Expression quantitative trait loci) are genetic loci (SNP) that explain variation in gene expression -> **biomarker**

Challenges:

1. LMM is **computationally expensive**
2. Millions of SNP sites burden the computational task

Context-Specific GWAS

Interesting to know eQTLs/biomarkers for a specific context, including tissues, cell types, cell stages, etc.



Modified from Cano-Gamez and Trynka Frontiers in Genetics 2020

Research Goal

To develop a computationally efficient pipeline for context-specific genome-wide association studies (GWAS) by integrating existing published tools, providing an alternative to traditional linear mixed model (LMM)-based approaches.

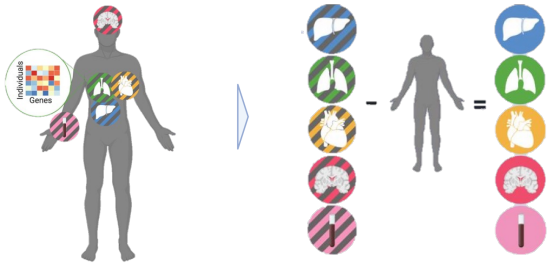
Question 1: Can we set up a pipeline where we can decompose the bulk/single-cell gene expression data into the context-shared and context-specific gene expression and model them separately by SLR?

1 Decomposition

- i: individual, where $i = 1, \ldots, I$
- c: context, where $c = 1, \ldots, C$
- E_{ic}: observed expression of a gene for individual i in context c

We decompose gene expression E_{ic} as:

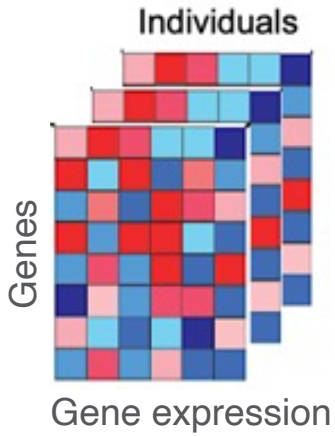
$$E_{ic} = \bar{E}_{..} + \underbrace{(E_{i.} - \bar{E}_{..})}_{E_i^{sh}} + \underbrace{(E_{ic} - E_{i.})}_{E_{ic}^{sp}}$$

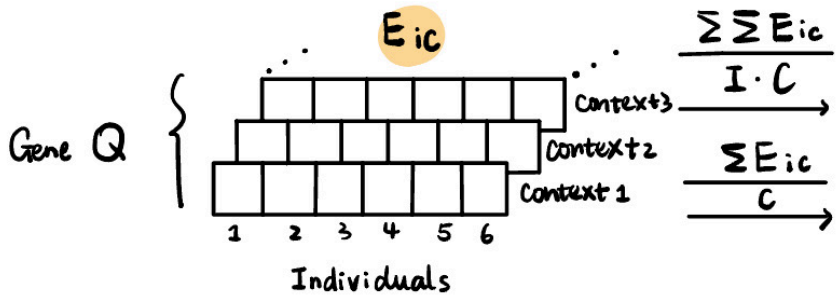


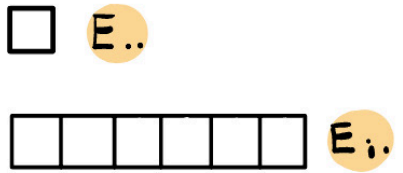
Where:

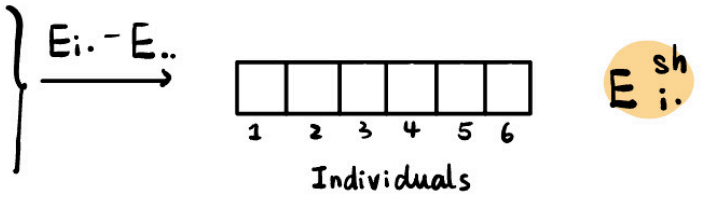
- $\bar{E}_{..} = \frac{\sum_{i=1}^{I}\sum_{c=1}^{C} E_{ic}}{I \times C}$: average expression of the gene across all individuals and contexts (population mean).
- $E_{i.} = \frac{\sum_{c=1}^{C} E_{ic}}{C}$: average expression of the gene for individual i across all contexts (individual mean)
- $E_i^{sh} = E_{i.} - \bar{E}_{..}$: context-shared expression for individual i.
- $E_{ic}^{sp} = E_{ic} - E_{i.}$: context-specific expression for individual i in context c.

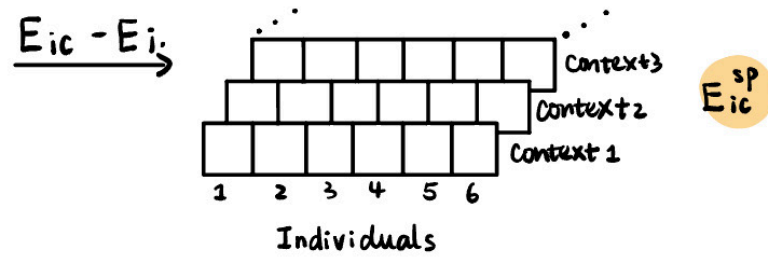
1 Decomposition

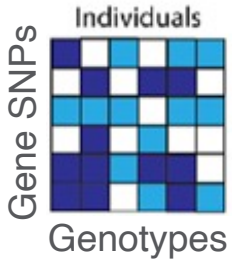


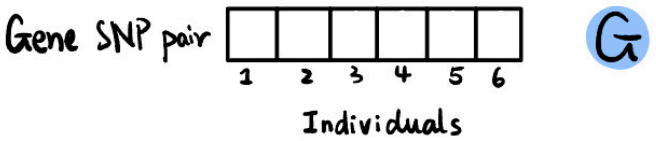






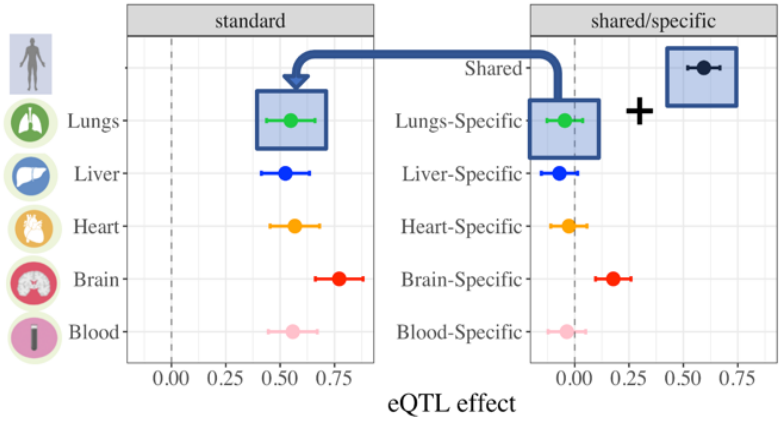






2 eQTL mapping

SLR between shared and context-specific gene expression with it cis-SNPs



$$E^{sh} = \alpha^{sh} + \beta^{sh} G + \varepsilon^{sh}$$

Shared eQTL (**sh-eQTL**)
(~ average eQTL effect across contexts)

$$E_1^{sp} = \alpha_1^{sp} + \beta_1^{sp} G + \varepsilon_1^{sp}$$

$$E_2^{sp} = \alpha_2^{sp} + \beta_2^{sp} G + \varepsilon_2^{sp}$$

$$\vdots$$

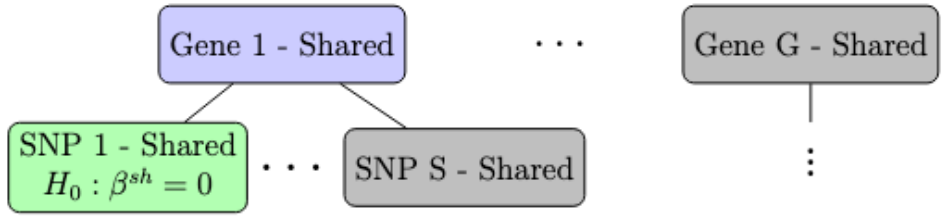
$$E_C^{sp} = \alpha_C^{sp} + \beta_C^{sp} G + \varepsilon_C^{sp}$$

Context-specific eQTL (**sp-eQTL**)
(~ deviation of eQTL in each context from shared eQTL effect)

Implemented by **MatrixEQTL** (Shabalin et al., 2012)

3 Multiple Testing Correction

Hierarchical FDR Correction by **TreeQTL** (Peterson et al., 2016)



Shared Effect

Model:

$$E_i^{sh} = \alpha_i^{sh} + \beta_{is}^{sh} G_{is} + \varepsilon_i^{sh},$$
$$\text{for } i = 1, \ldots, I \text{ (Genes)}, s = 1, \ldots, S \text{ (SNPs)}$$

Hypothesis:

- **Level 1**: Is there any shared cis-eQTL for the gene?

$$H_0^{(i)} : \beta_{is}^{sh} = 0 \quad \text{for all cis-SNPs } s \text{ of gene } i$$
$$H_A^{(i)} : \exists\ s \text{ such that } \beta_{is}^{sh} \neq 0$$

If gene i passes Level 1, proceed to Level 2

- **Level 2**: Which SNPs show shared effect?

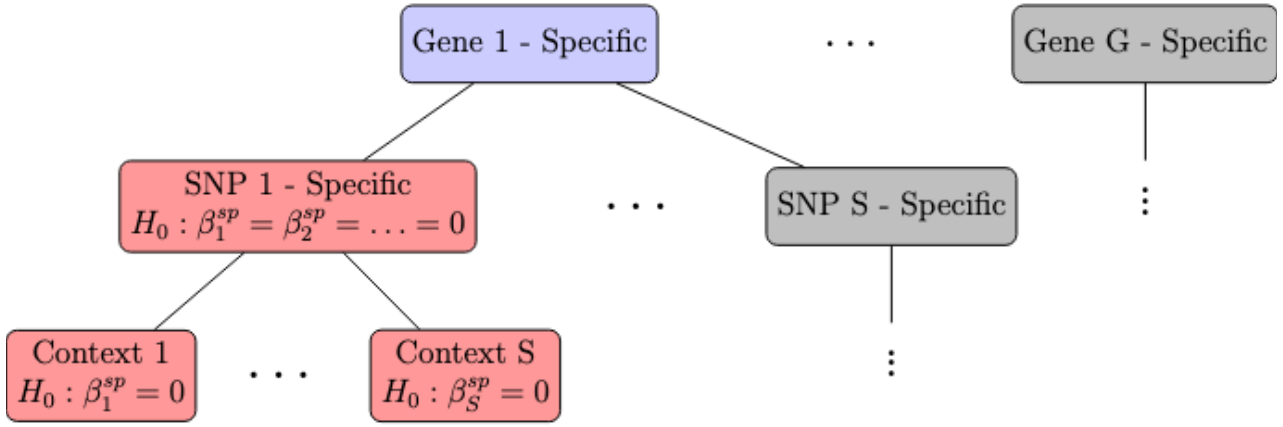
$$H_0^{(is)} : \beta_{is}^{sh} = 0 \quad \text{for each cis-SNP } s$$

Decision logic:

- Apply FDR correction on each level
- If $H_0^{(is)}$ is rejected, declare SNP s is a **shared eQTL** for gene i.

3 Multiple Testing Correction

Hierarchical FDR Correction by **TreeQTL** (Peterson et al., 2016)



Context-specific Effect

Model:

$$E_c^{sp} = \alpha_c^{sp} + \beta_c^{sp} G + \varepsilon_c^{sp},$$
$$\text{for } i = 1, \ldots, I \text{ (Genes)}, s = 1, \ldots, S \text{ (SNPs)}, c = 1, \ldots, C \text{ (Contexts)}$$

Hypothesis:

- **Level 1**: Is there any context-specific cis-eQTL for the gene?

$$H_0^{(i)} : \beta_{sc}^{sp} = 0 \quad \text{for all cis-SNPs } s \text{ and all contexts } c \text{ of gene } i$$
$$H_A^{(i)} : \exists\, s, c \text{ such that } \beta_{sc}^{sp} \neq 0$$

If gene i passes Level 1, proceed to Level 2

- **Level 2**: Is there any SNP that shows context-specific effect in any context?

$$H_0^{(is)} : \beta_{sc}^{sp} = 0 \quad \text{for all contexts } c \text{ of SNP } s \text{ in gene } i$$
$$H_A^{(is)} : \exists\, c \text{ such that } \beta_{sc}^{sp} \neq 0$$

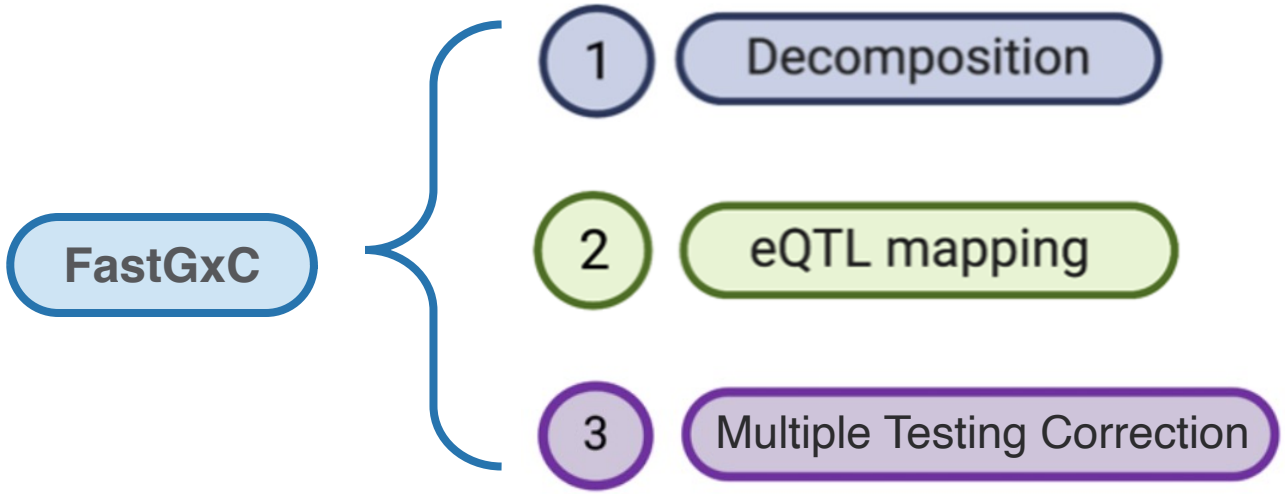
If SNP s passes Level 2, proceed to Level 3

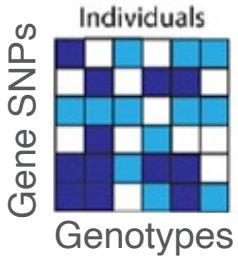
- **Level 3**: Which contexts show the effect?

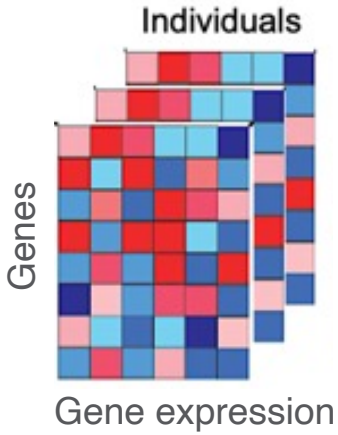
$$H_0^{(isc)} : \beta_{sc}^{sp} = 0$$
$$H_A^{(isc)} : \beta_{sc}^{sp} \neq 0$$

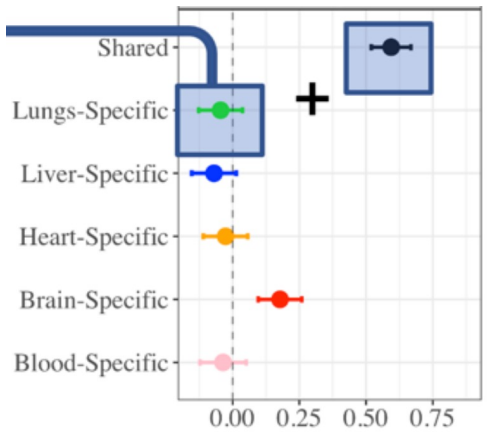
Decision logic:

- Apply FDR correction on each level.
- If $H_0^{(is)}$ is rejected, declare SNP s a **specific eQTL** for gene i.
- If $H_0^{(isc)}$ is rejected, declare SNP s a **context-specific eQTL** for gene i in context c.


FastGxC
1
Decomposition
2
eQTL mapping
3
Multiple Testing Correction







Genotype

1. Each SNP genotype is drawn independently from a binomial distribution:

$$G_{js} \sim \text{Binomial}(2, 0.1), \text{for individual } j \text{ and SNP } s$$

Gene expression

2. Gene expression depends linearly on genotype

$$Y_{jic} = \mu_c + \beta_{ic} G_{ji} + \varepsilon_{jc}$$

where:

- Y_{jic} is the expression of gene i in context c for individual j,
- μ_c is the context-specific baseline expression,
- β_{ic} is the effect size,
- G_{ji} is the genotype at the causal SNP,
- ε_{jc} is the residual noise.

3. The effect size is determined by the assumed heritability h^2_{ic}:

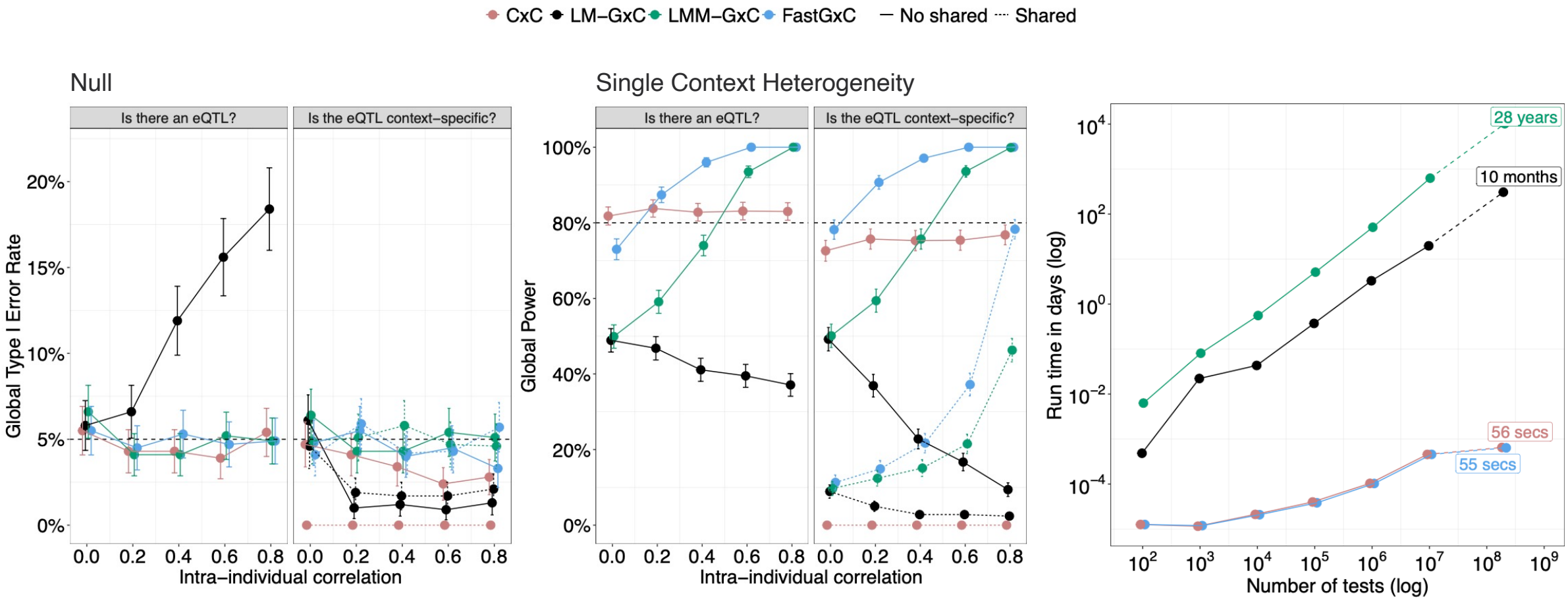
$$\beta_{ic} = \sqrt{\frac{h^2_{ic} v_e}{(1 - h^2_{ic}) \cdot \text{Var}(G_{ji})}}$$

- Under the **null** scenario, $h^2_{ic} = 0$ for all contexts. So, $Y_{jic} = \mu_c + \varepsilon_{jc}$.
- Under the **single context heterogeneity** scenario, each gene is assumed to be dominant in only one context $c^\star_i$ with larger effect size, i.e. $h^2_{ic} = \begin{cases} 0.2 & \text{if } c = c^\star_i \\ 0.1 & \text{otherwise} \end{cases}$. So, $Y_{jic} = \mu_c + \beta_{ic} G_{ji} + \varepsilon_{jc}$

4. Error follows a multivariate normal distribution with intra-individual correlation w_{corr} among contexts:

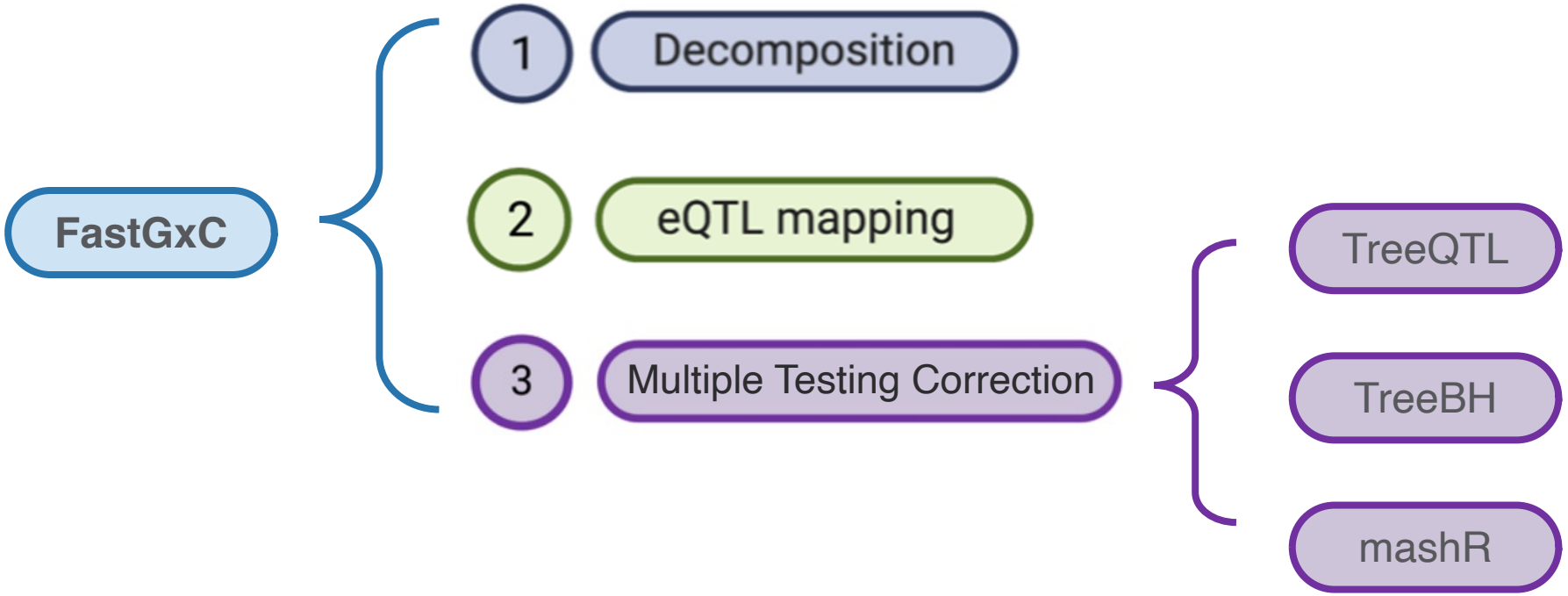
$$\boldsymbol{\varepsilon}_j \sim \mathcal{N}_C(\mathbf{0}, \Sigma), \quad \Sigma = (1 - w_{\text{corr}}) I_C + w_{\text{corr}} \mathbf{1}_C \mathbf{1}_C^T$$

5. Only one SNP per gene is causal and affects expression in all contexts. eg. For each gene i, only the first SNP $(s = 1)$ is causal.



FastGxC can control the type I error under null and have relatively high eQTL detecting power under the alternative with less computational expense.

Question 2: If we can set up the pipeline, can we refine the multiple testing correction step by comparing different methods?



TreeQTL and TreeBH are **Hierarchical Correction**

	TreeQTL (Peterson et al., 2016)	**TreeBH** (Bogomolov et al., 2021)
Application	**eQTL-specific** (genes, SNPs, tissues)	**General** hierarchical testing
Structure	Fixed tree	Arbitrary tree
Correction	Simes' method, BH/BY procedures	New error rates and testing strategies

mashR

Observation:

$$\hat{\boldsymbol{b}}_j \mid \boldsymbol{b}_j \sim \mathcal{N}_R(\boldsymbol{b}_j, V_j)$$

Prior:

$$\boldsymbol{b}_j \sim \sum_{k,\ell} \pi_{k\ell} \mathcal{N}_R(\mathbf{0}, \omega_\ell U_k)$$

- U_k: intra- and inter-context covariance
- ω_ℓ: scaling magnitudes
- $\pi_{k\ell}$: mixture weights

Posterior:

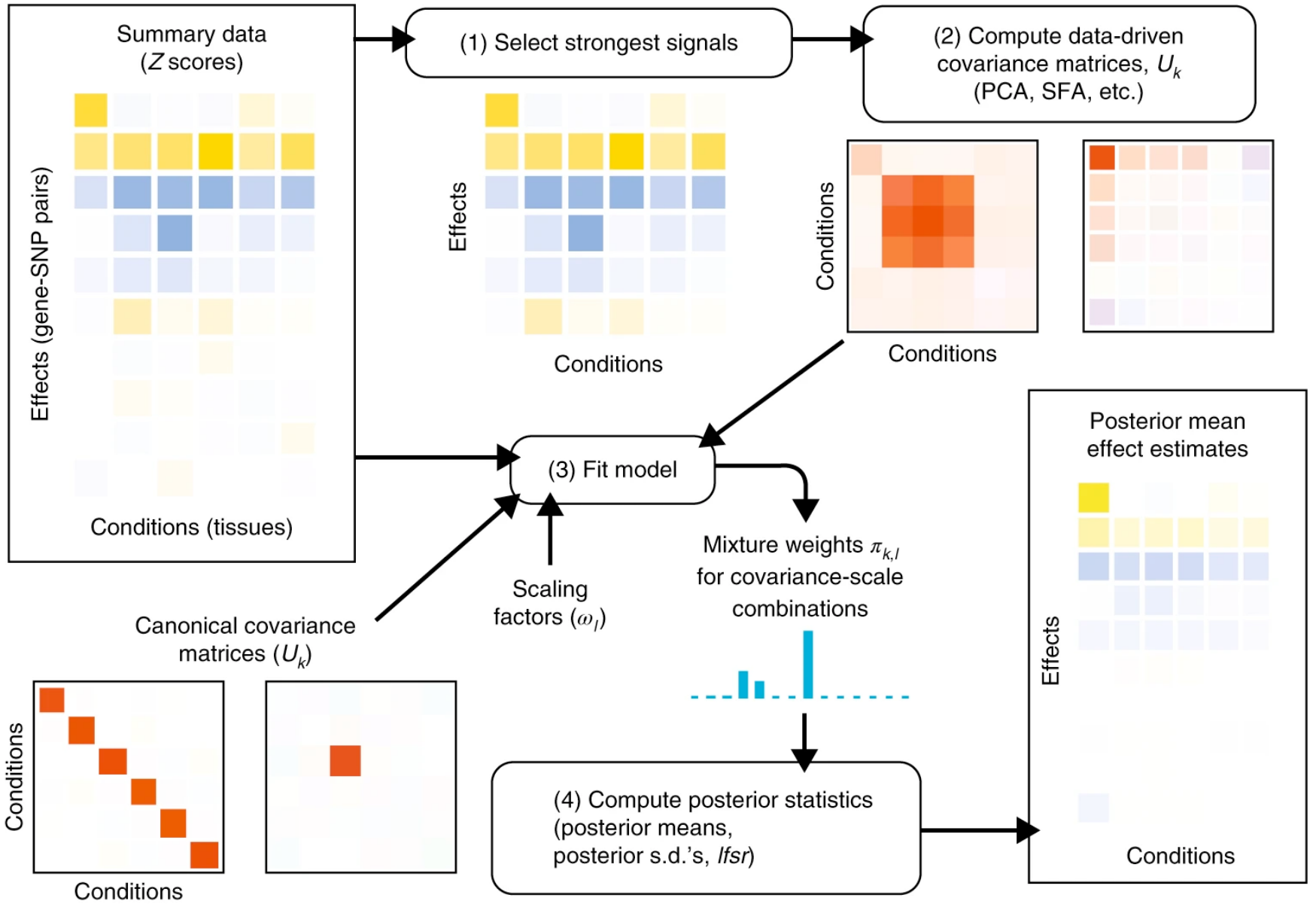
$$\boldsymbol{b}_j \mid \hat{\boldsymbol{b}}_j \sim \sum_{k,\ell} \tilde{\pi}_{j,k\ell} \mathcal{N}_R(\tilde{\boldsymbol{\mu}}_{j,k\ell}, \tilde{U}_{j,k\ell})$$

via the **EM** algorithm

Shrinkage estimator:

$$\mathbb{E}[\boldsymbol{b}_j \mid \hat{\boldsymbol{b}}_j] = \sum_{k,\ell} \tilde{\pi}_{j,k\ell} \tilde{\boldsymbol{\mu}}_{j,k\ell}$$

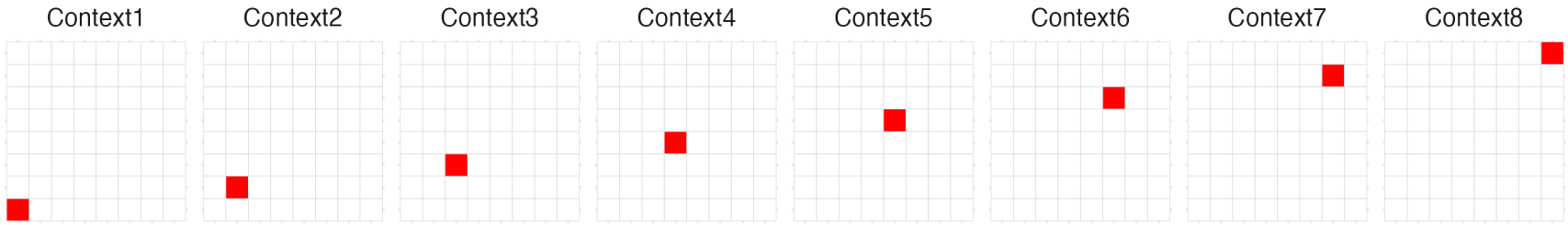
Multivariate Adaptive Shrinkage in R (Urbut et al. 2018)



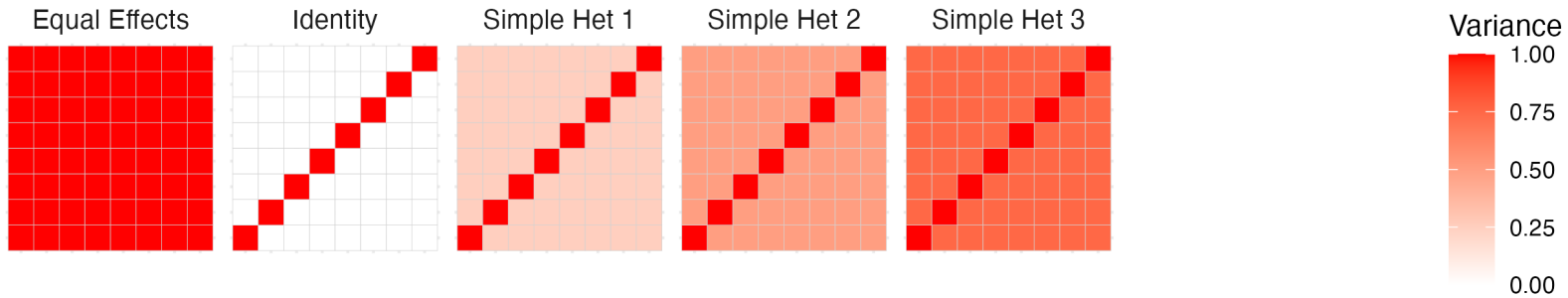
Empirical Bayes Shrinkage across Conditions

mashR

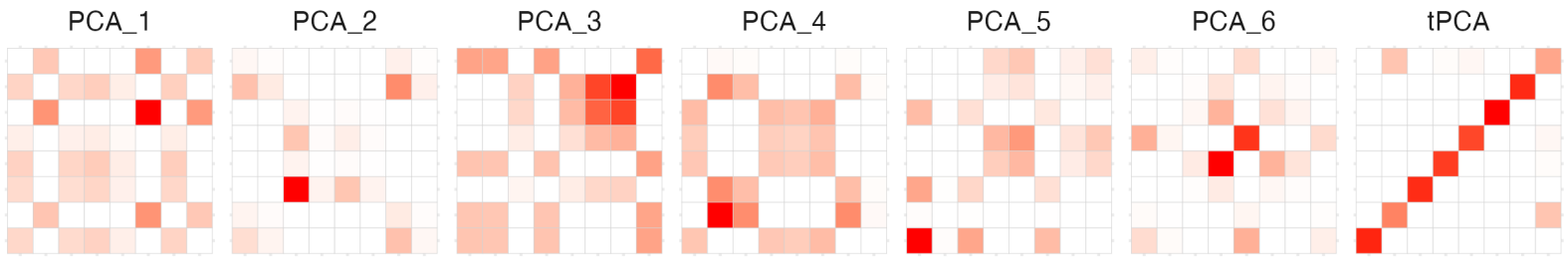
cov_canonical

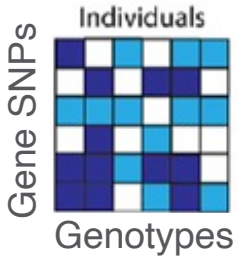


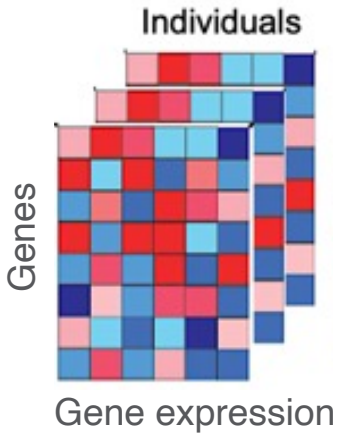
cov_canonical



cov_PCA







Genotype

1. Each SNP genotype is drawn independently from a binomial distribution:

$$G_{js} \sim \text{Binomial}(2, 0.1), \text{for individual } j \text{ and SNP } s$$

Gene expression

2. Gene expression depends linearly on genotype

$$Y_{jic} = \mu_c + \beta_{ic} G_{ji} + \varepsilon_{jc}$$

where:

- Y_{jic} is the expression of gene i in context c for individual j,
- μ_c is the context-specific baseline expression,
- β_{ic} is the effect size,
- G_{ji} is the genotype at the causal SNP,
- ε_{jc} is the residual noise.

3. The effect size is determined by the assumed heritability h^2_{ic}:

$$\beta_{ic} = \sqrt{\frac{h^2_{ic} v_e}{(1 - h^2_{ic}) \cdot \text{Var}(G_{ji})}}$$

- Under the **null** scenario, $h^2_{ic} = 0$ for all contexts. So, $Y_{jic} = \mu_c + \varepsilon_{jc}$.
- Under the **single context heterogeneity** scenario, each gene is assumed to be dominant in only one context $c^\star_i$ with larger effect size, i.e. $h^2_{ic} = \begin{cases} 0.2 & \text{if } c = c^\star_i \\ 0.1 & \text{otherwise} \end{cases}$. So, $Y_{jic} = \mu_c + \beta_{ic} G_{ji} + \varepsilon_{jc}$

4. Error follows a multivariate normal distribution with intra-individual correlation w_{corr} among contexts:

$$\boldsymbol{\varepsilon}_j \sim \mathcal{N}_C(\mathbf{0}, \Sigma), \quad \Sigma = (1 - w_{\text{corr}}) I_C + w_{\text{corr}} \mathbf{1}_C \mathbf{1}_C^T$$

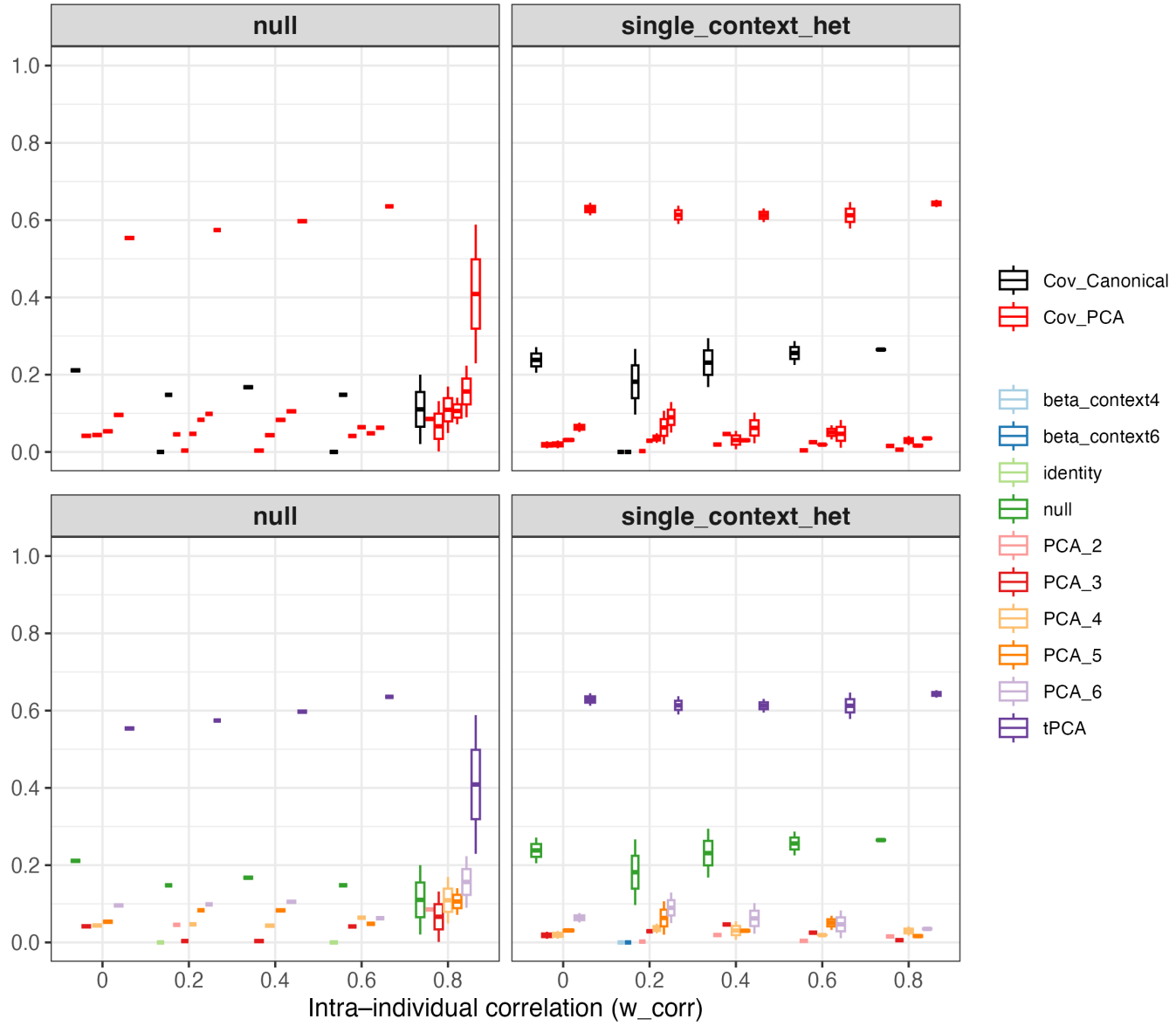
5. Only one SNP per gene is causal and affects expression in all contexts. eg. For each gene i, only the first SNP $(s = 1)$ is causal.

mashR

Posterior:

$$\boldsymbol{b}_j \mid \hat{\boldsymbol{b}}_j \sim \sum_{k,\ell} \tilde{\pi}_{j,k\ell} \mathcal{N}_R(\tilde{\boldsymbol{\mu}}_{j,k\ell}, \tilde{U}_{j,k\ell})$$

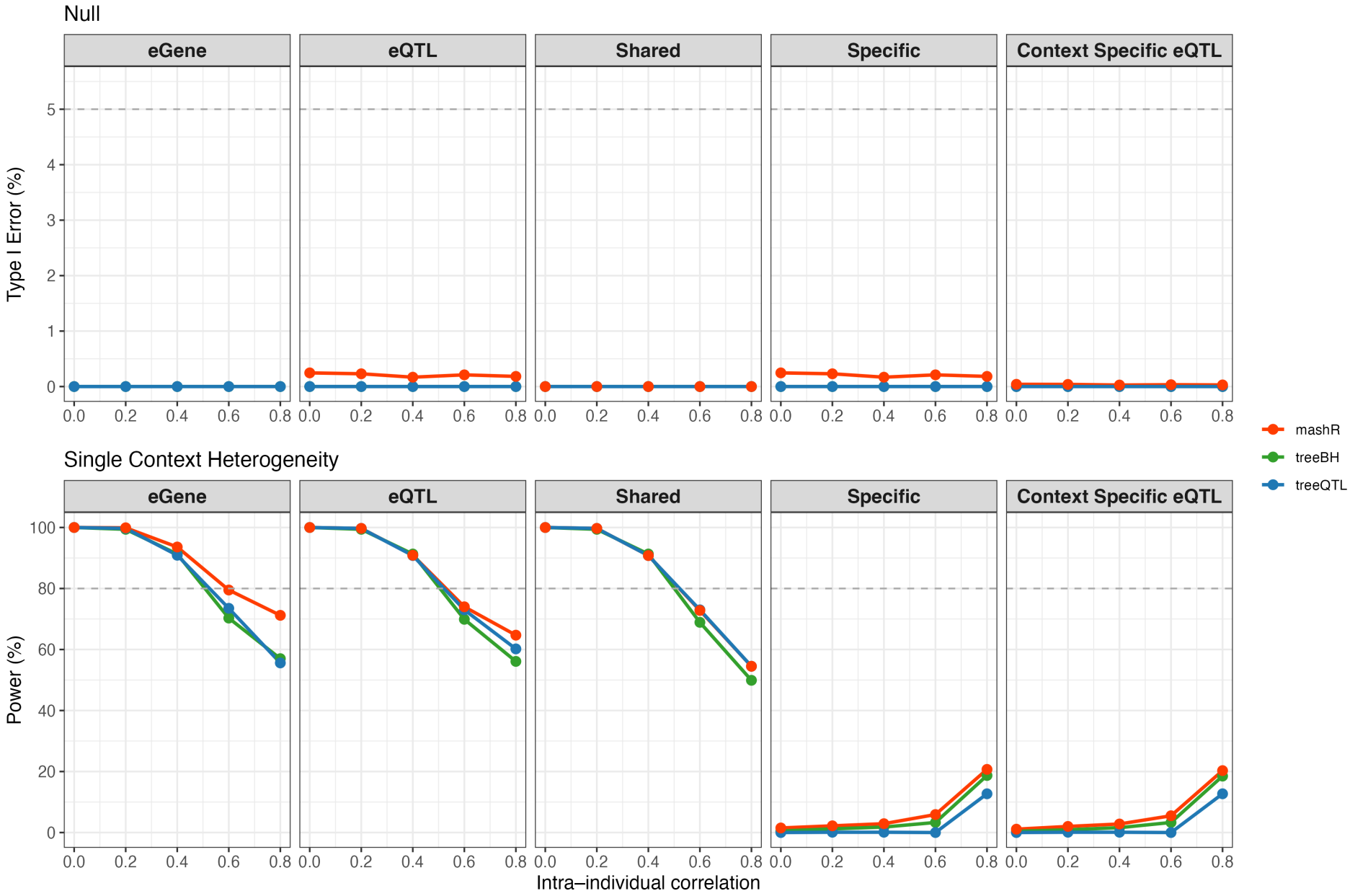
Covariance matrix by PCA contribute more to the posterior than canonical ones under null and alternative



100 Individuals Scenario

All methods are valid and control type I error, even with small sample sizes.

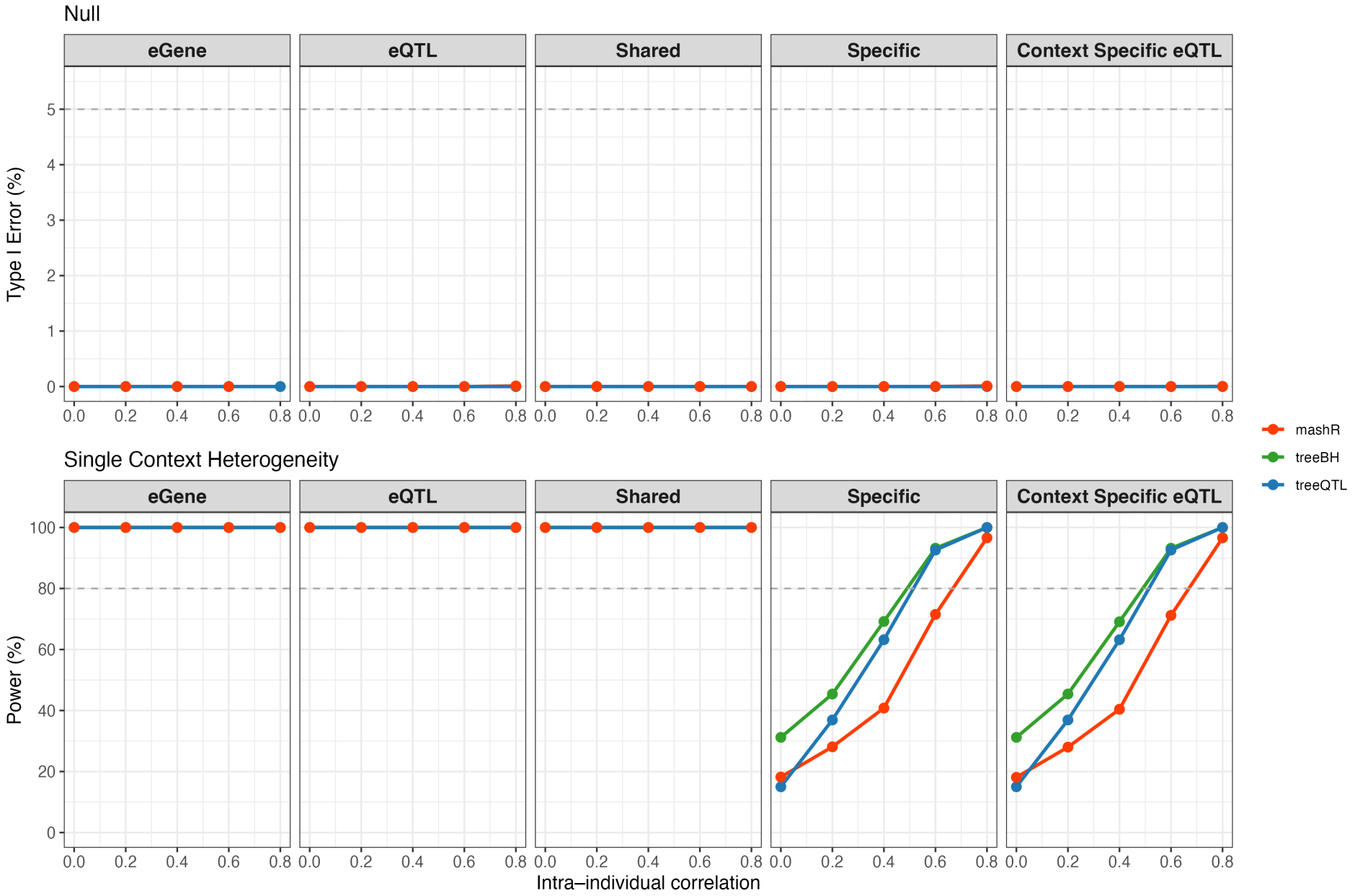
mashR's **adaptive shrinkage gives it an edge** in moderate and high correlation.



500 Individuals Scenario

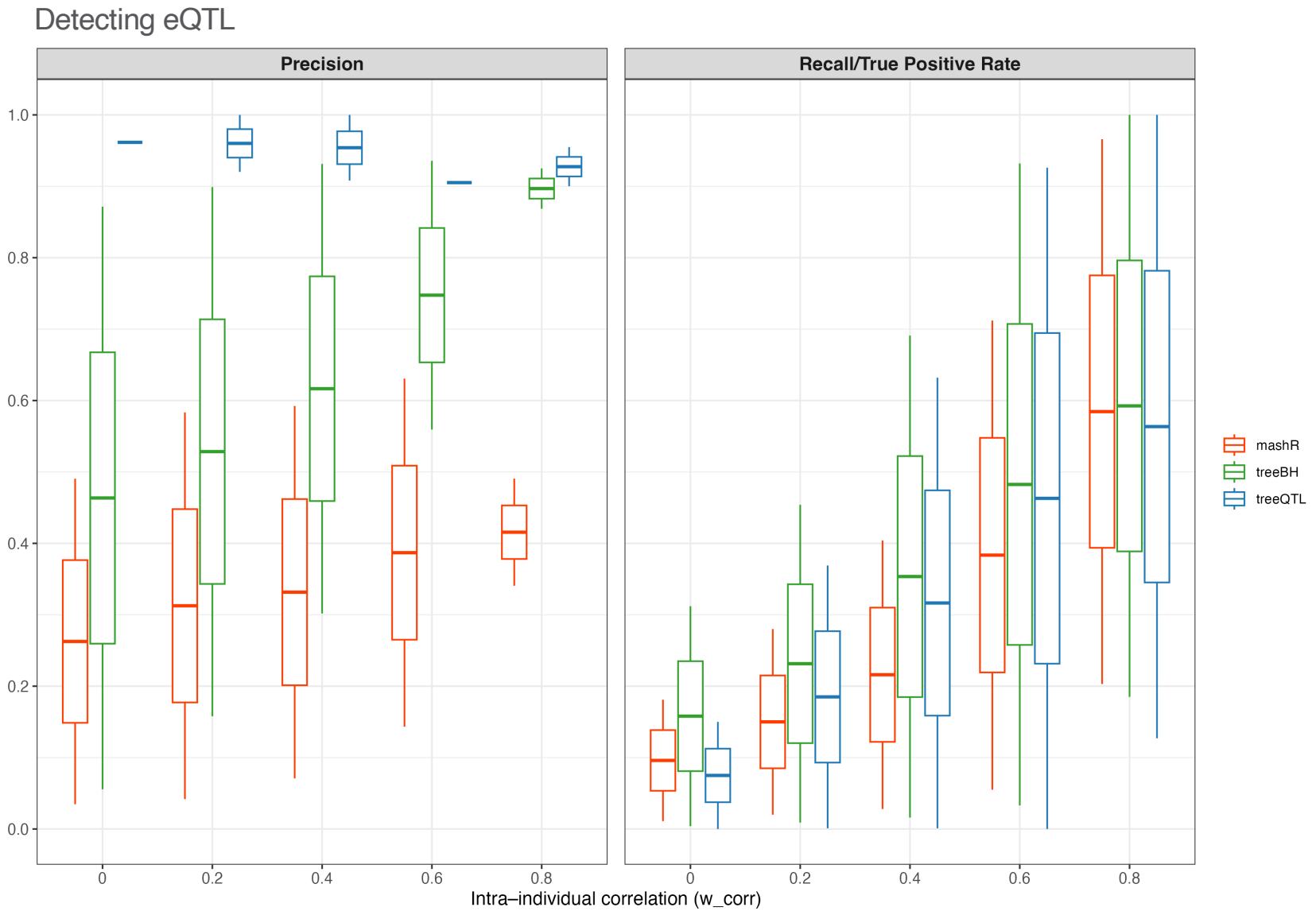
All methods are valid and control type I error

Hierarchical correction methods benefit from structure and outperform mashR in detecting more specific/context-specific effects.



TreeQTL achieves a high and consistent precision across all correlation levels.

Hierarchical correction methods outperform mashR in precision and recall.



1. FastGxC can provide an efficient framework for detecting both context-shared and context-specific eQTLs using bulk or single-cell RNA sequencing data across diverse biological contexts.
2. TreeQTL demonstrates better precision and power in large-sample scenarios
3. mashR's adaptive shrinkage method enables better detecting power in smaller cohorts and moderate-to-high intra-individual correlation settings

Discussion

1. The basic FastGxC pipeline has already been published on GitHub (https://github.com/BalliuLab/FastGxC). We will extend the tool to support multiple package options and provide a detailed comparison of their pros and cons to guide users in selecting the most suitable configuration for their analysis.
2. The main limitation is that the method comparison has not yet been validated on real RNA-seq data, where factors like batch effects and biological variability may impact performance.

Key References

Uffelmann, E., Huang, Q. Q., Munung, N. S., *et al.* (2021). Genome-wide association studies. *Nature Reviews Methods Primers, 1*, 59.

Cronbach, L. J., & Webb, N. M. (1975). Between-class and within-class effects in multilevel analysis. *Sociological Methodology, 6*, 57–101.

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics, 11*, 424.

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics, 28*(10), 1353–1358.

Peterson, C. B., Bogomolov, M., Benjamini, Y., & Sabatti, C. (2016). TreeQTL: hierarchical error control for eQTL findings. *Bioinformatics, 32*(16), 2556–2558.

Flutre, T., Wen, X., Pritchard, J., & Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics, 9*(5), e1003486.

Urbut, S. M., Wang, G., Carbonetto, P., *et al.* (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics, 51*, 187–195.

Thank You