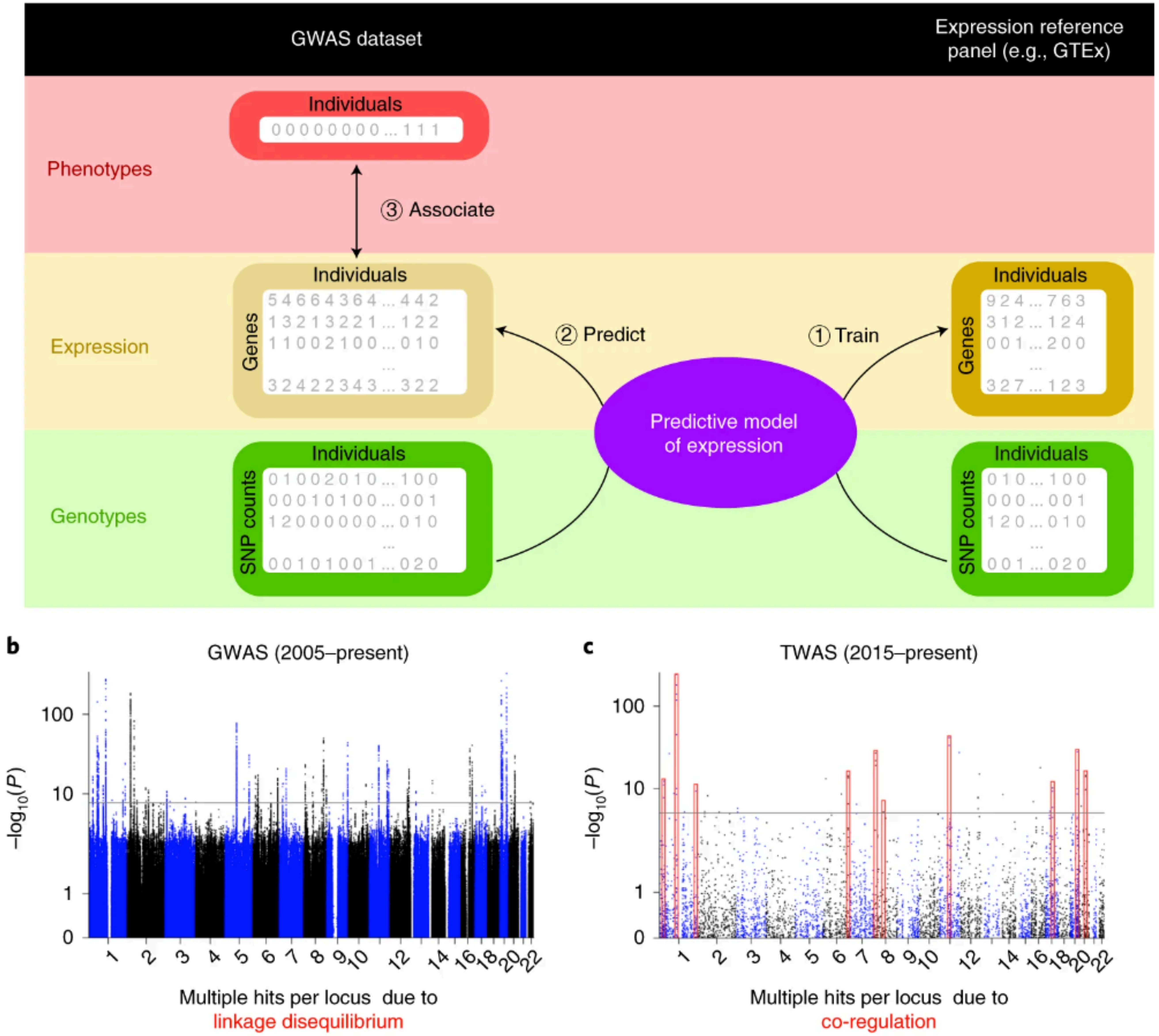
# OTTERS: a powerful TWAS framework leveraging summary-level reference data

Dai, Q., Zhou, G., Zhao, H. et al. OTTERS: a powerful TWAS framework leveraging summary-level reference data. Nat Commun **14**, 1271 (2023)

0601 Teresa Lin
Knowles Lab Journal Club

# GWAS & TWAS



GWAS: Find association between **genetic markers** and phenotype

How does these variants affect downstream genes?

↓

TWAS: Find association between **gene expression** and phenotype

Integrating eQTL and GWAS result

- TWAS are not causal-gene tests
  Genetically predicted expression ≠ total expression

**Total expression**

Genetic
Environmental
Technical

←

**Common cis eQTLs**
Rare cis eQTLs
trans eQTLs

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N. *et al.* *Nat Genet* **51**, 592–599 (2019).

# Traditional TWAS Analysis

**Stage 1:**

Genetically regulated expression

- Using **individual-level data** from tissues of interest to create a **GReX** imputation model.
- Training tools like PrediXcan, **FUSION**, and TIGAR can be used after model configuration.

$$\mathbf{e}_\mathbf{g} = \mathbf{X}_\mathbf{g}\mathbf{w} + \boldsymbol{\epsilon}_\mathbf{g}, \quad \boldsymbol{\epsilon}_\mathbf{g} \sim \mathrm{N}(0, \sigma_\epsilon^2\mathbf{I})$$

$\mathbf{e}_\mathbf{g}$: gene expression levels of gene **g**
$\mathbf{X}_\mathbf{g}$: genotype data of SNP predictors proximal within gene **g**
**w**: genetic effect sizes

**Stage 2:**

- Uses the trained eQTL effect sizes (**w**) to impute gene expressing (using GReX) in an independent GWAS.
- Test for association between GReX and phenotype.

$$phenotype \ \ ? \ \ G\hat{R}eX = X_{new}\hat{w}$$

Equivalent to a gene-based association test which takes **eQTL effect sizes** as corresponding test **SNP weights**
eQTL summary data are analogous to GWAS summary data where gene expression represents the phenotype

# OTTERS TWAS Variation

## Stage 1: Estimate cis-eQTL effect size

- Adapt PRS methods for TWAS
- Using **summary-level** reference data from the following single variant regression models.
- Using **marginal least squared effect size estimates** and **p values** from eQTL sum stats to estimate effect size

(Assuming summary-level data provide information between a single variant j and expression of gene g)

$$\mathbf{e}_g = \mathbf{x}_j w_j + \boldsymbol{\epsilon}_j, \boldsymbol{\epsilon}_j \sim N\left(0, \sigma^2_{\epsilon_j}\mathbf{I}\right), j = 1, \ldots, m.$$

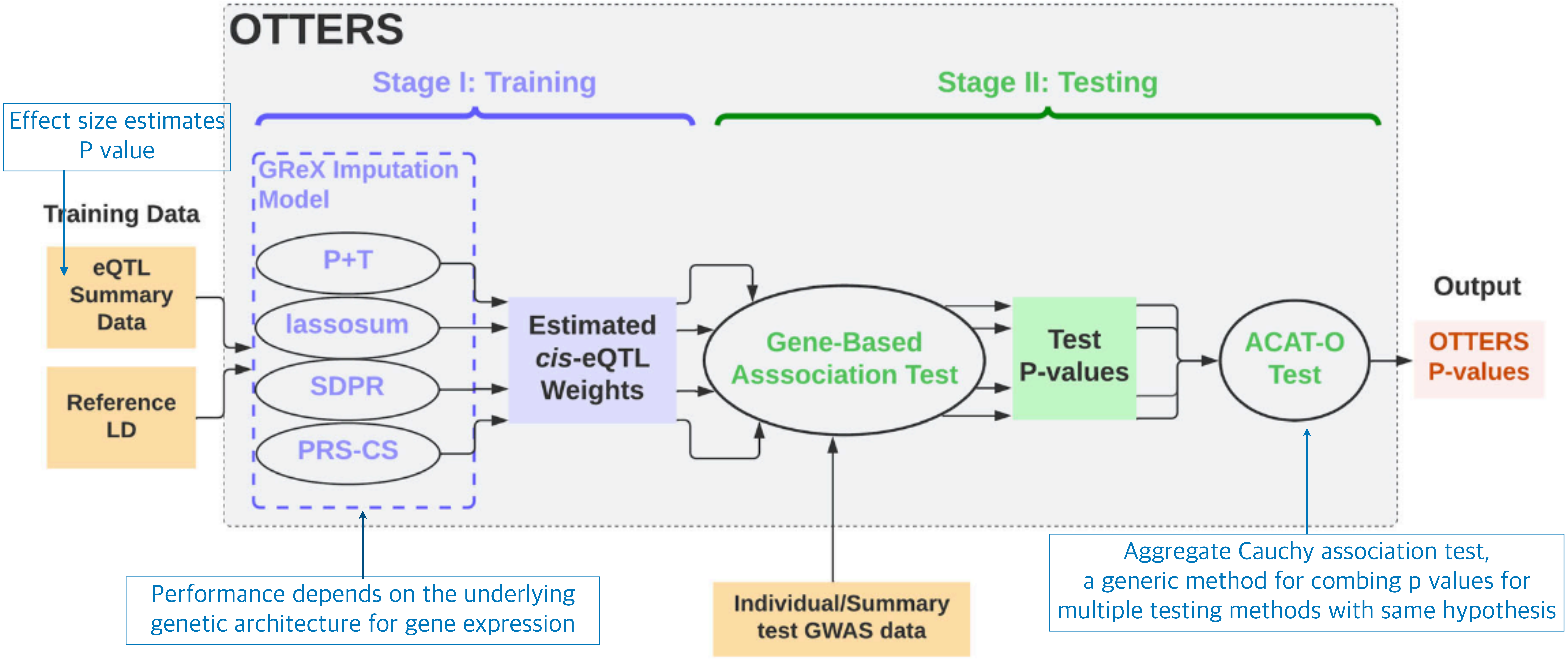
$$\widetilde{\mathrm{w}}_j \approx Z_j / \sqrt{\mathrm{median}(n_{g,j})}$$

**$\mathbf{e}_g$**: gene expression levels of gene **g**
**$\mathbf{X}_j$**: genotype data for generic variant **j**
**$\mathbf{w}_j$**: effect size estimates
**$\mathbf{Z}_j$** : corresponding eQTL statistic value by single variant test
**median($\mathbf{n}_{g,j}$)**: median sample size of cis-eQTLs for target gene **g**

## Stage 2:

- Uses the trained eQTL effect sizes (**w**) to impute gene expressing (using GReX) in an independent GWAS.
- Test for association between GReX and phenotype.

# Framework:
# Omnibus Transcriptome Test using Expression Reference Summary Data

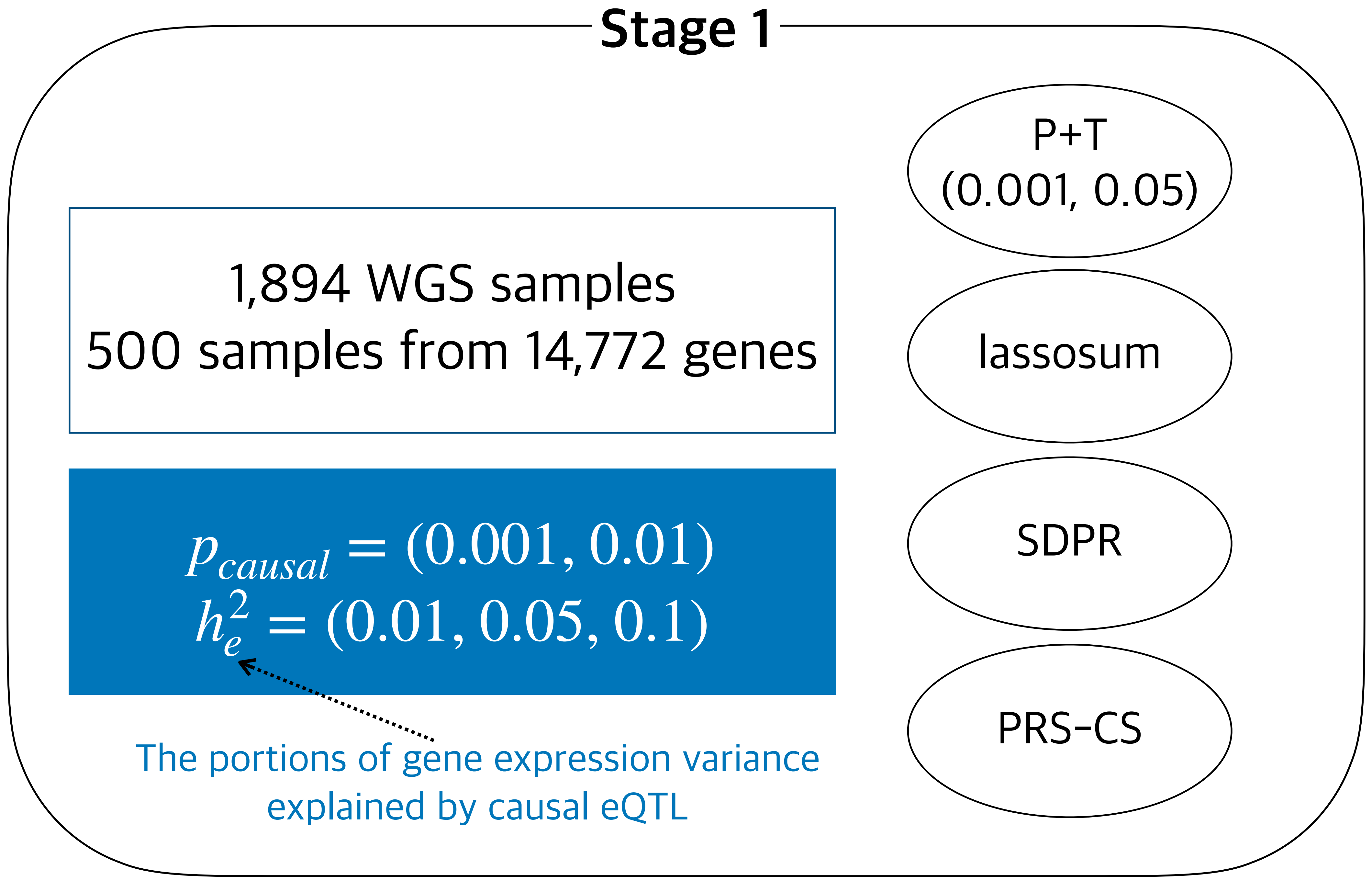


Effect size estimates
P value

Performance depends on the underlying genetic architecture for gene expression

Aggregate Cauchy association test,
a generic method for combing p values for
multiple testing methods with same hypothesis

# Simulation Study



## Stage 1

1,894 WGS samples
500 samples from 14,772 genes

$$p_{causal} = (0.001, 0.01)$$
$$h_e^2 = (0.01, 0.05, 0.1)$$

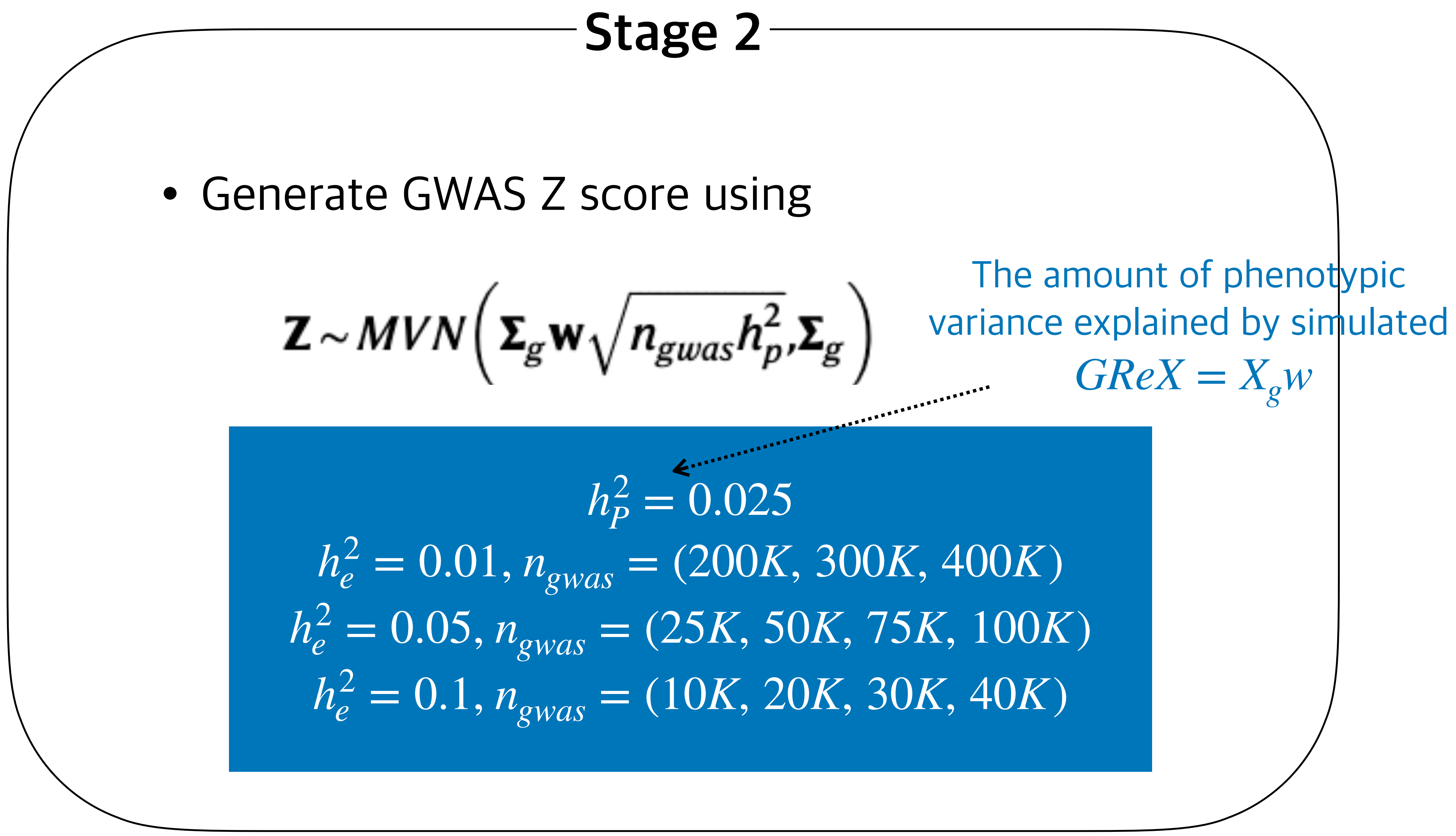
The portions of gene expression variance explained by causal eQTL

P+T
(0.001, 0.05)

lassosum

SDPR

PRS-CS



## Stage 2

- Generate GWAS Z score using

$$\mathbf{Z} \sim MVN\left(\mathbf{\Sigma}_g \mathbf{w} \sqrt{n_{gwas} h_p^2}, \mathbf{\Sigma}_g\right)$$

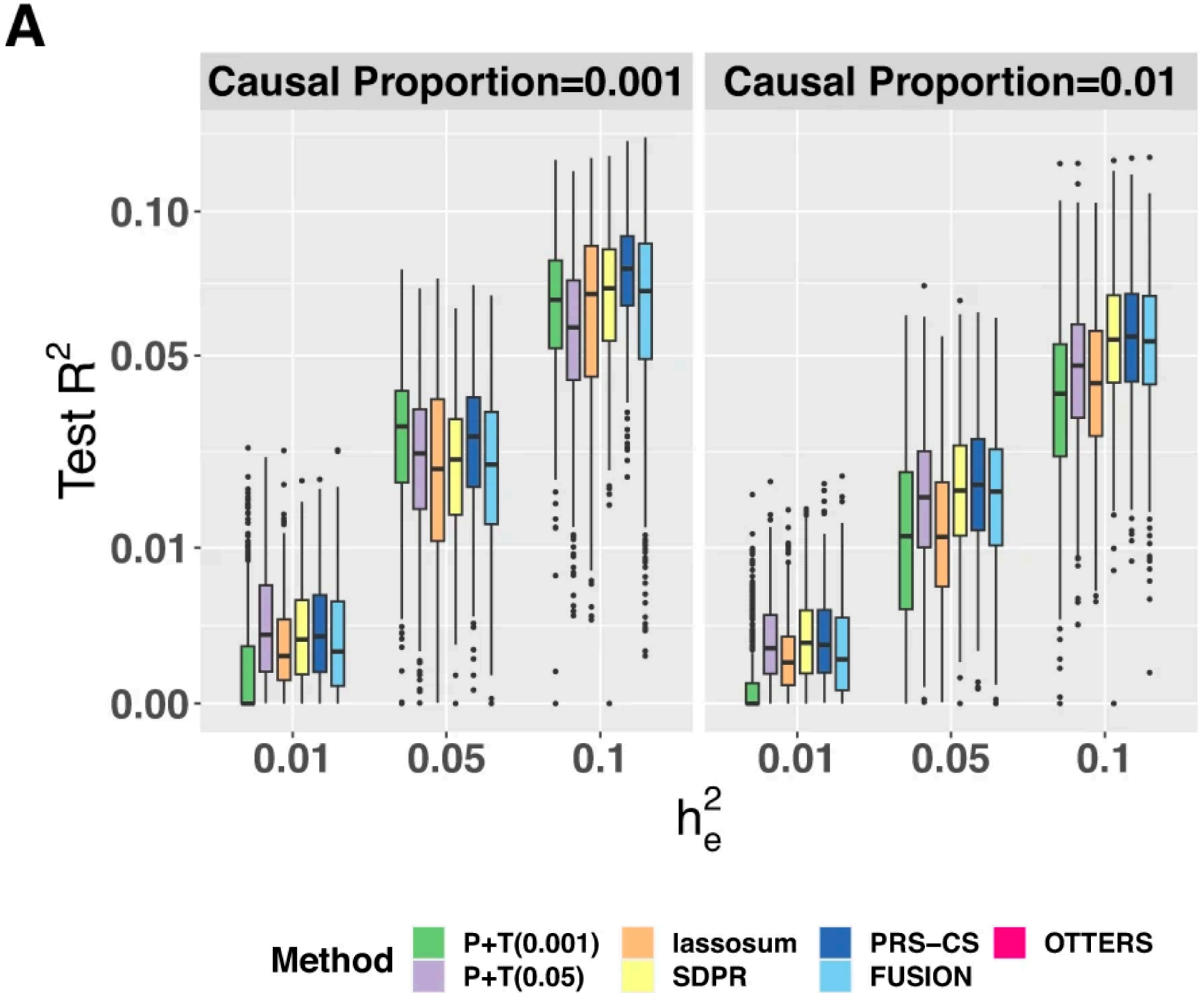
The amount of phenotypic variance explained by simulated $GReX = X_g w$

$$h_P^2 = 0.025$$
$$h_e^2 = 0.01, n_{gwas} = (200K, 300K, 400K)$$
$$h_e^2 = 0.05, n_{gwas} = (25K, 50K, 75K, 100K)$$
$$h_e^2 = 0.1, n_{gwas} = (10K, 20K, 30K, 40K)$$

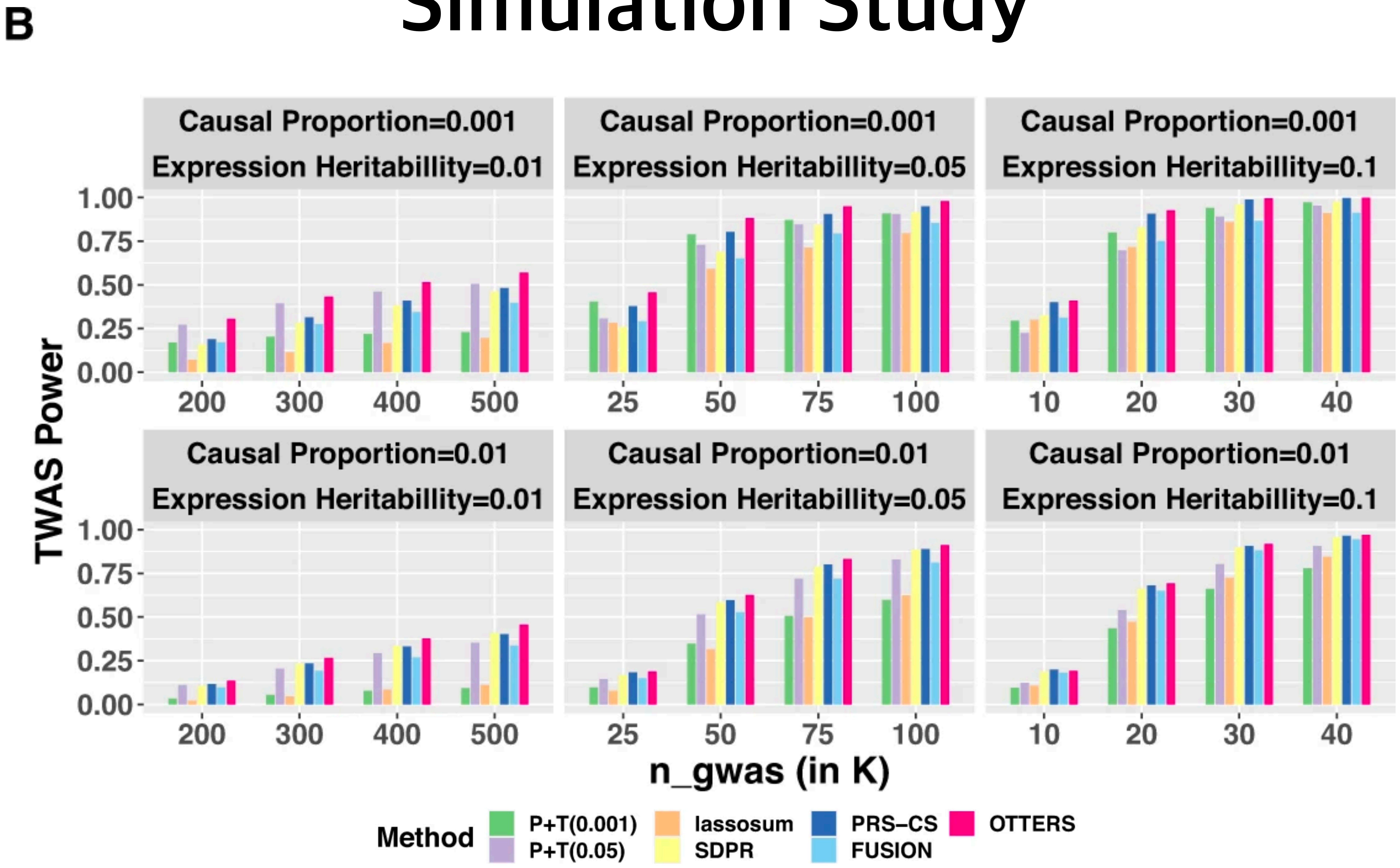
$$\mathbf{e}_g = \mathbf{X}_g \mathbf{w} + \boldsymbol{\epsilon}_g$$
$$\boldsymbol{\epsilon}_g \sim N(0, (1 - h_e^2)\mathbf{I})$$
$$\mathbf{y} = h_p\left(\mathbf{X}_g \mathbf{w}\right) + \boldsymbol{\epsilon}_p, \boldsymbol{\epsilon}_p \sim N(0, \mathbf{I}).$$

Evaluating using test $R^2$
(The squared Pearson correlation coefficient between inputed GReX and simulated gene expression)

# Simulation Study



# Simulation Study



# Real data evaluation

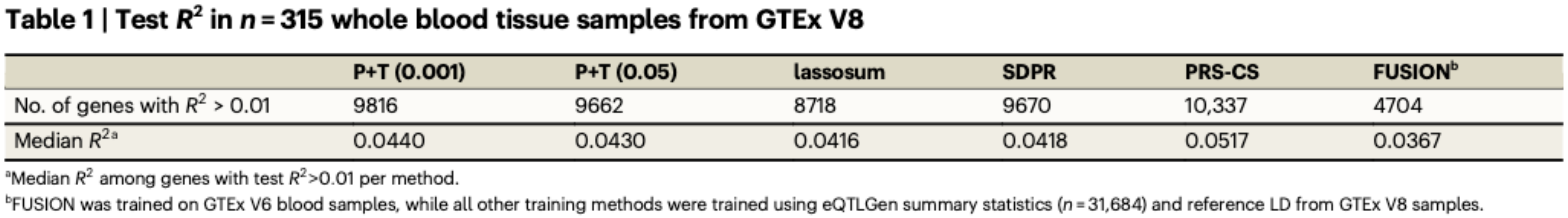
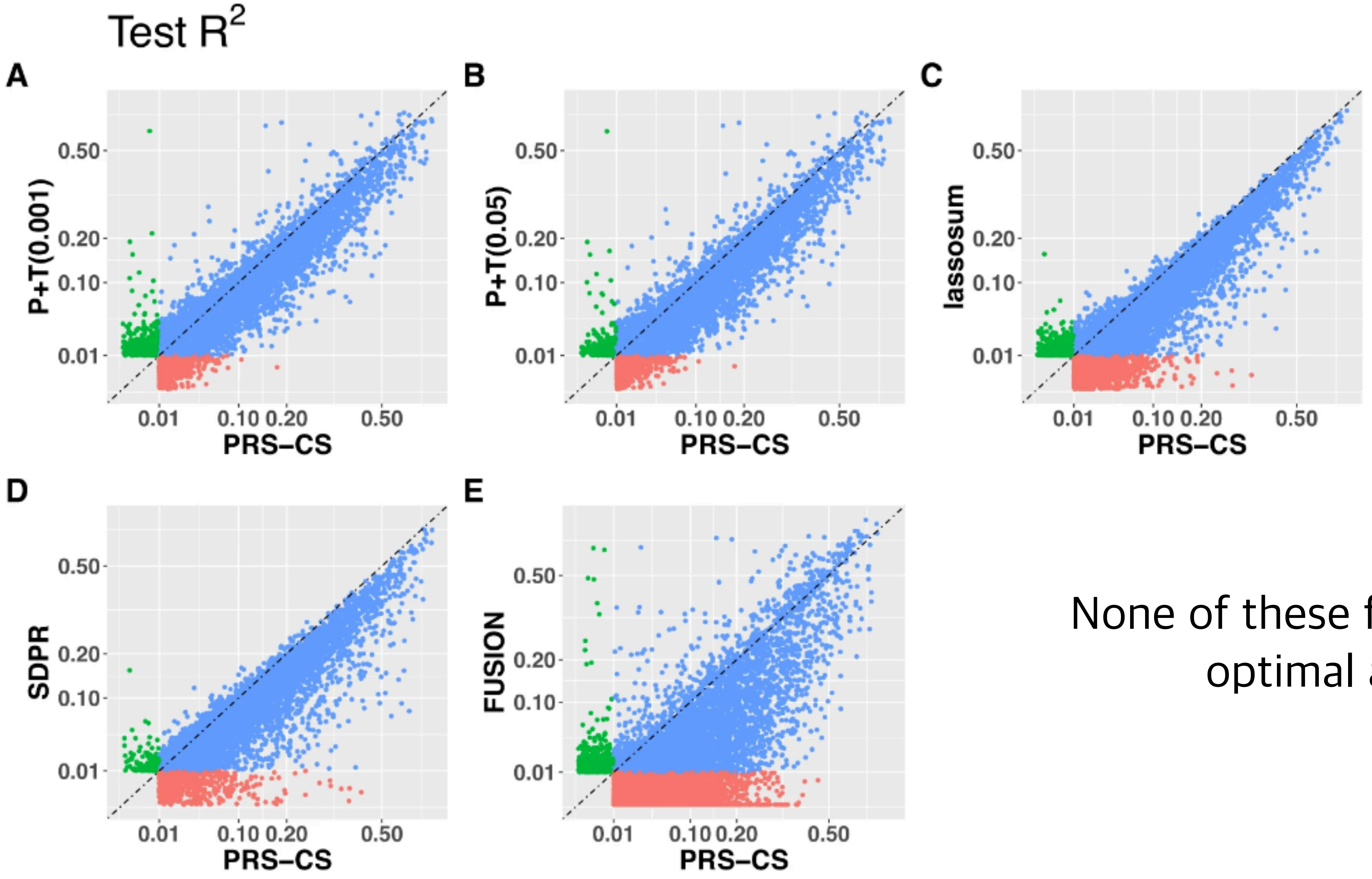


**Table 1 | Test $R^2$ in $n = 315$ whole blood tissue samples from GTEx V8**

| | P+T (0.001) | P+T (0.05) | lassosum | SDPR | PRS-CS | FUSION[b] |
|---|---|---|---|---|---|---|
| No. of genes with $R^2 > 0.01$ | 9816 | 9662 | 8718 | 9670 | 10,337 | 4704 |
| Median $R^2$[a] | 0.0440 | 0.0430 | 0.0416 | 0.0418 | 0.0517 | 0.0367 |

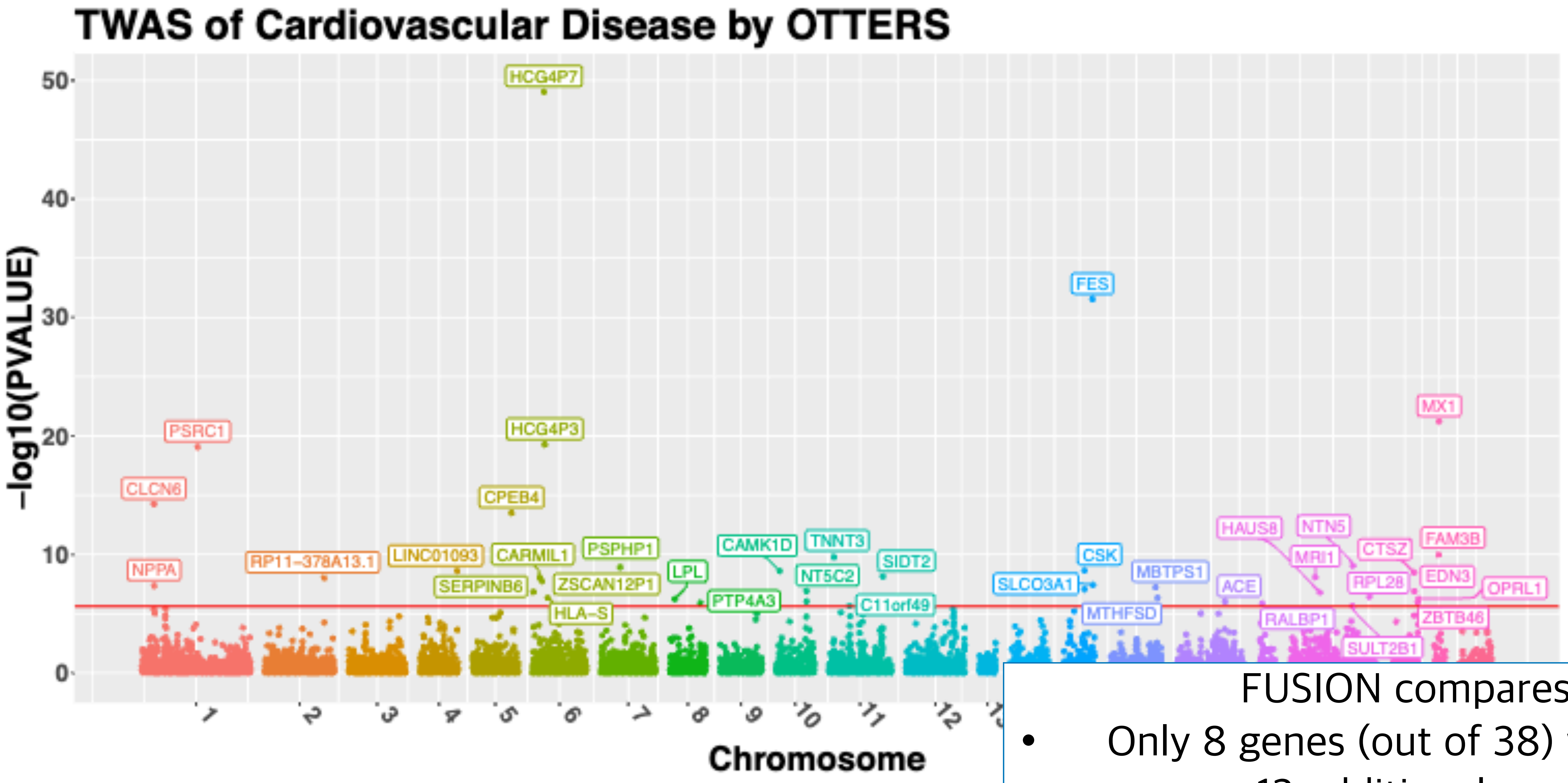
[a] Median $R^2$ among genes with test $R^2$>0.01 per method.

[b] FUSION was trained on GTEx V6 blood samples, while all other training methods were trained using eQTLGen summary statistics ($n = 31{,}684$) and reference LD from GTEx V8 samples.



None of these four PRS method were optimal across all genes

# Real data evaluation



FUSION compares to OTTERS:

- Only 8 genes (out of 38) were found in FUSION
- 13 additional genes were found

| Method | OTTERS | P+T (0.001) | P+T (0.05) | Lassosum | SDPR | PRS-CS | FUSION |
|---|---|---|---|---|---|---|---|
| # independently significant TWAS gene | 38 | 17 | 11 | 10 | 41 | 12 | 21 |

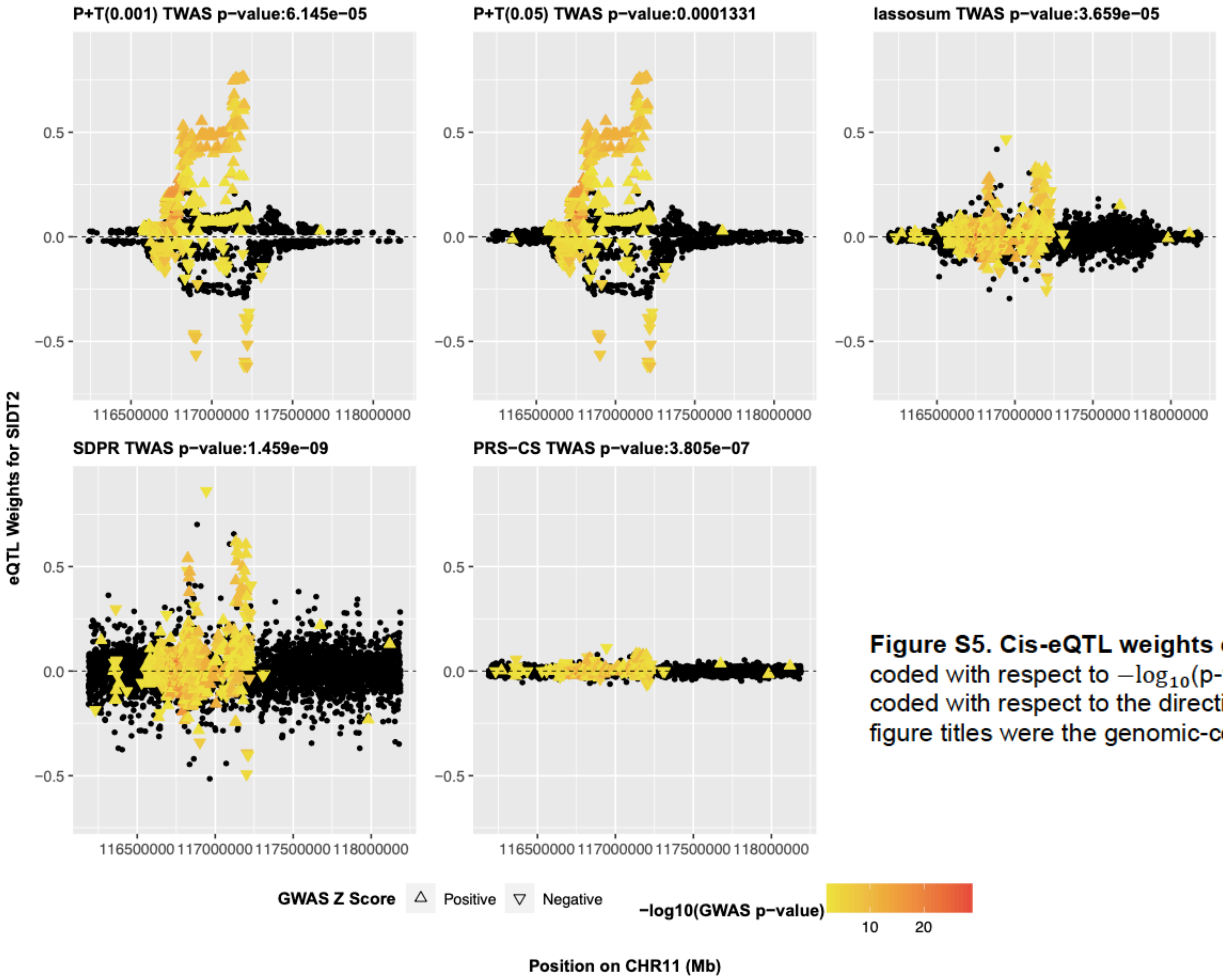


**Figure S5. Cis-eQTL weights estimated by individual methods for gene *SIDT2*.** Color coded with respect to $-\log_{10}$(p-value) from the UKBB GWAS summary statistics and shape coded with respect to the direction of GWAS Z-score test statistics. TWAS P-values shown in figure titles were the genomic-control corrected from TWAS Z-score tests (two-sided).

# Other notes and potential usage

- Adding more PRS methods in Stage 1 might give a higher TWAS power, with additional computational cost
- This method cannot provide direction of gene-phenotype association
- Could be use in other molecular QTLs like splicing QTL, methylation QTLs, metabolomics QTLs and protein QTLs

# Thanks for listening