Transcriptome-wide Association Studies
Lecture 2

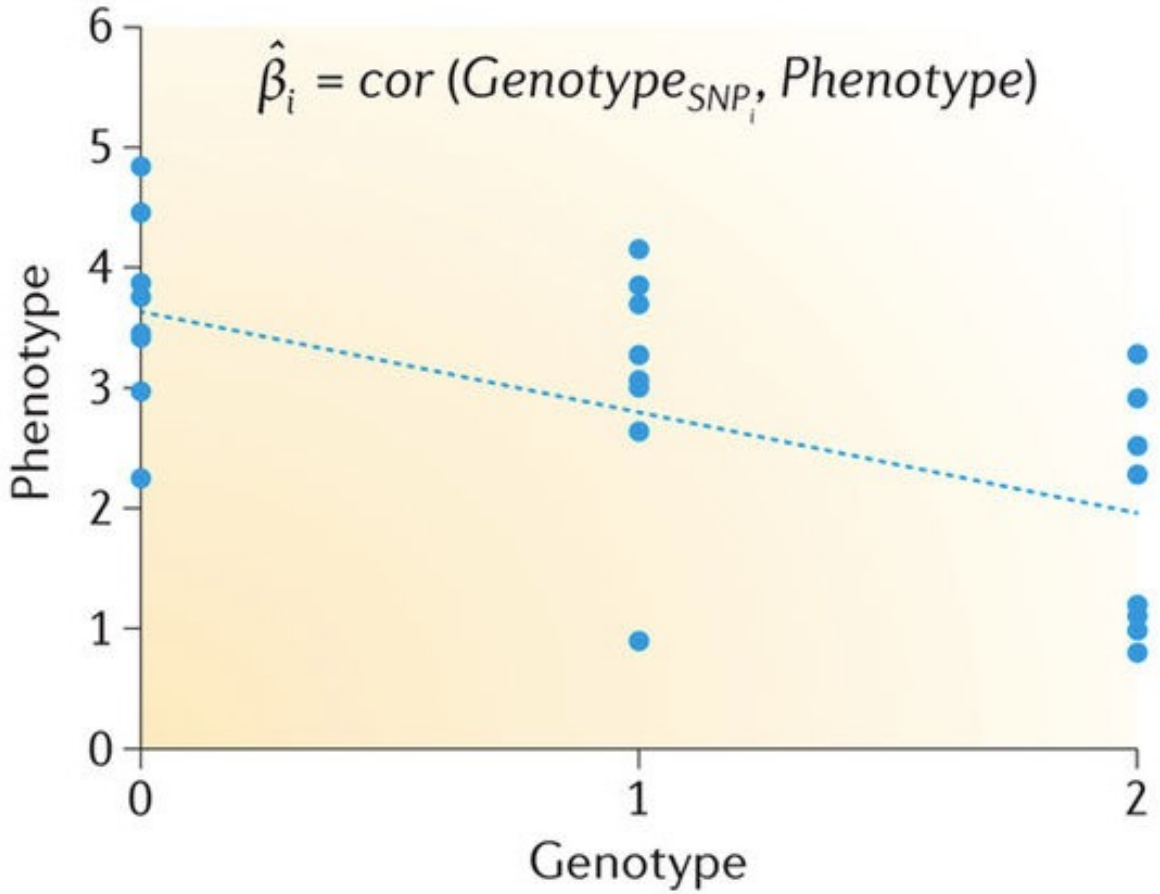
Outline

- Connection between GWAS and eQTL Analysis
- OTTERS: Leveraging eQTL Summary Data for TWAS
- Fine-mapping TWAS Results

Genome-wide Association Study (GWAS)

- Study quantitative trait by linear regression model
  - $Y = \beta_0 + \alpha C + \beta_1 X + \epsilon,\ \epsilon \sim N(0, \sigma^2)$
  - Y represents the quantitative trait values
  - X represents the genotype data (0, 1, 2) for additive genetic model
  - C represents the confounding covariates or other environmental variables
  - ϵ represents the error term, other unknown factors
- $H_0: \beta_1 = 0\ ;\ H_a: \beta_1 \neq 0$
- P-values can be obtained by Wald Test

Linear Regression Model



Expression Quantitative Trait Loci (eQTL) Analysis

- Consider the profiled gene expression levels as the quantitative trait E_g in the following single variant linear regression model:

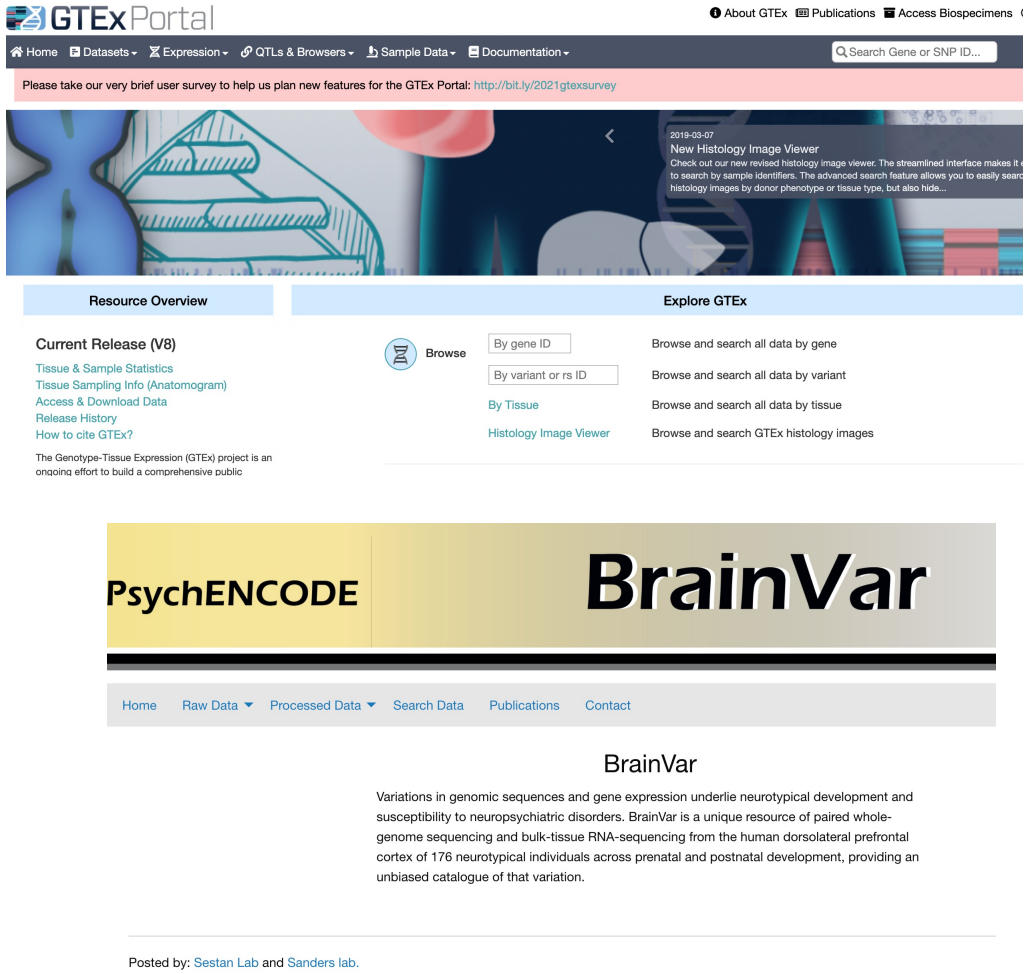
$$E_g = \beta_0 + \alpha C + \beta_1 X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

  - X represents the genotype data (0, 1, 2) or dosage [0, 2] of the test SNP
  - C represents the confounding covariates or other environmental variables
  - ϵ represents the error term, other unknown factors
- eQTL: SNPs significantly associated with a gene expression quantitative trait ($H_0: \beta_1 = 0$ is significantly rejected) are referred as **e**xpression **Q**uantitative **T**rait **L**oci (eQTL)
- Cis-eQTL : eQTL nearby the test gene (e.g., located within the +-1MB region of the transcription starting site; thousands cis-SNPs per gene).
- Trans-eQTL : eQTL distant from the test gene (e.g., located out of the +-1MB region of the transcription starting site, or on different chromosome; ~10M trans-SNPs per gene).

Reference Transcriptomic Data

- Leverage reference datasets possessing both genetic and transcriptomic data

- Transcriptome-wide Association Study (TWAS):
  - Exploit SNP-expression relationships in reference data to impute gene expression in GWAS study
  - Test association between imputed gene expression (GReX) and phenotype



Other Expression Resources Exist

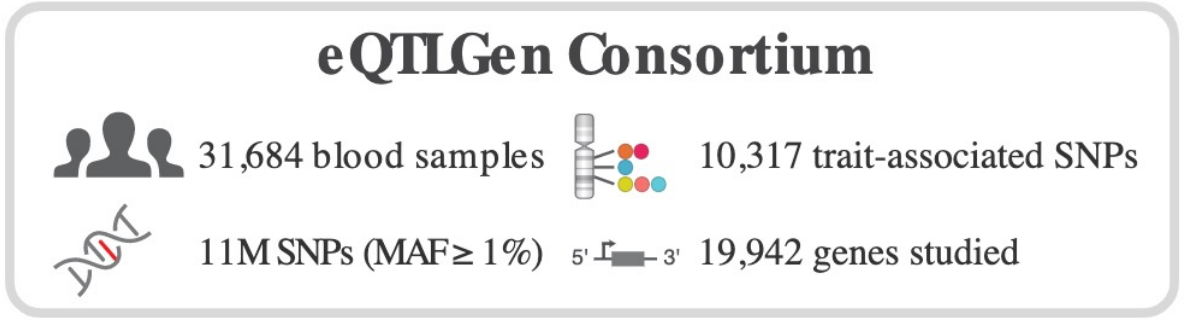
- Consortia exist that provide summary-level eQTL results
  - Larger sample size for training GReX imputation model
  - Enable TWAS using more reference expression data

ARTICLES
https://doi.org/10.1038/s41588-021-00913-z
nature genetics
Check for updates

Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression

Trait-associated genetic variants affect complex phenotypes primarily via regulatory mechanisms on the transcriptome. To investigate the genetics of gene expression, we performed *cis*- and *trans*-expression quantitative trait locus (eQTL) analyses using blood-derived expression from 31,684 individuals through the eQTLGen Consortium. We detected *cis*-eQTL for 88% of genes, and these were replicable in numerous tissues. Distal *trans*-eQTL (detected for 37% of 10,317 trait-associated variants tested) showed lower replication rates, partially due to low replication power and confounding by cell type composition. However, replication analyses in single-cell RNA-seq data prioritized intracellular *trans*-eQTL. *Trans*-eQTL exerted their effects via several mechanisms, primarily through regulation by transcription factors. Expression of 13% of the genes correlated with polygenic scores for 1,263 phenotypes, pinpointing potential drivers for those traits. In summary, this work represents a large eQTL resource, and its results serve as a starting point for in-depth interpretation of complex phenotypes.

- eQTLGen (N=~32K) provides cis-eQTL summary data of blood
  - Increased sample size relative to GTeX (N=~100s) --> increased TWAS power



Limitations of TIGAR/PrediXcan/FUSION

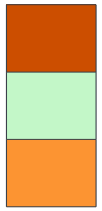
Limited Sample Sizes:
GTEx V8 n < 1K

Training Stage

Individual-level eQTL panel

Multivariable regression model

Expression gene A



~

eQTL Genotype

A	T	G	T
A	A	C	T
C	T	G	A

Large-scale Meta-analyses:
eQTLGen n ~ 32K

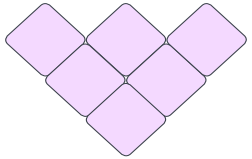
Summary-level eQTL panel

OTTERS:

eQTL summary statistics

	Z-Score
eQTL 1	0.6
eQTL 2	-0.7
eQTL 3	-1.8

reference LD pattern



Stage I: Training Imputation Models (Estimate $\boldsymbol{w}$) from Summary-level eQTL Data

- **Goal:** Training gene expression prediction model with summary-level eQTL data:

$$\boldsymbol{E}_g = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$$

- **Summary-level eQTL Data**
  - Effect sizes and p-values from single-variant regression models used in eQTL analysis:

$$\boldsymbol{E}_g = \boldsymbol{X}_j\widetilde{w_j} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \boldsymbol{I}), j = 1, \dots, m.$$

$\boldsymbol{X}_j, \widetilde{w_j}$: Genotype vector, marginal effect size for cis-eQTL j

  - Reference LD panel of cis-eQTLs from population of similar ancestry such as the 1000 Genomes Project or GTEx WGS data

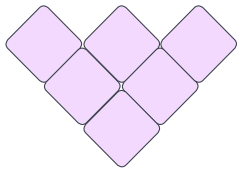
Connections between GWAS and eQTL Summary Data

?

eQTL summary statistics

	Z-Score
eQTL 1	0.9
eQTL 2	-1.7
eQTL 3	1.3

reference LD pattern



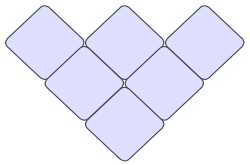
estimate joint effect sizes of eQTLs on the gene expression

Polygenetic Risk Score (PRS) Methods

GWAS summary statistics

	Z-Score
SNP 1	0.9
SNP 2	-1.7
SNP 3	1.3

reference LD pattern



estimate joint effect sizes of SNPs on the trait

Training Imputation Models (Estimate $\boldsymbol{w}$)

- Considered four complementary summary-level training methods used in traditional **Polygenic Risk Score (PRS)** analysis of complex traits
  - P-value thresholding (*P+T*)
  - Summary-statistic LASSO (*lassosum*)
  - Summary-level Bayesian Dirichlet process regression (*SDPR)*
  - PRS with continuous shrinkage prior (*PRS-CS)*

P-value thresholding (*P*+*T*)

- Given eQTL summary data from:

$$\boldsymbol{E}_g = \boldsymbol{X}_j \widetilde{w_j} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \boldsymbol{I}), j = 1, \ldots, m.$$

- Pick a p-value threshold (e.g., 0.001) to filter out eQTL with p-value less significant than the threshold
- Pick a R^2 threshold (e.g., 0.3) to select the top significant eQTL among all eQTL with genotype R^2 greater than the threshold
- Take the marginal eQTL effect sizes $\widetilde{w_j}$ as variant weights for Stage-II gene-based association test

PLINK Tool: https://zzz.bwh.harvard.edu/plink/

PRS Methods using Summary Data

- Recall multiple linear regression model for gene expression prediction (or deriving polygenic risk score):

$$\boldsymbol{E_g = Xw + \varepsilon}$$

- Assume $\boldsymbol{E_g}$ and columns of $\boldsymbol{X}$ are centered with mean 0 and standardized with variance 1
- GWAS/eQTL summary data from marginal single variant linear regression model:

$$\boldsymbol{E_g = X_j \widetilde{w_j} + \epsilon,\ \epsilon \sim N(0, \sigma_\epsilon^2 I), j = 1, \dots, m}$$

- **Goal:** Estimates $\widehat{\boldsymbol{w}}$ in the multiple linear regression model with marginal effect size estimates $\left\{\widetilde{w_j}, = \frac{E_g^T X_j}{n}, j = 1, \dots, m\right\}$, marginal Z-score statistics $\{Z_j, j = 1, \dots, m\}$, sample size n, reference genotype correlation matrix (LD) $\boldsymbol{R}$
- **Note:** $\boldsymbol{\widetilde{w} = \frac{E_g^T X}{n}; R = \frac{X^T X}{n}; Z = \widetilde{w}\sqrt{n}}$

Summary-statistic LASSO (*lassosum*)

- Multiple linear regression model for gene expression prediction:

$$\boldsymbol{E_g = Xw + \varepsilon}$$

- LASSO estimates $\boldsymbol{w}$ in the multiple linear regression model by minimizing the following loss function with a penalty parameter λ:

$$\mathrm{f}(\mathbf{w}) = \frac{\mathbf{E}_\mathrm{g}^\mathrm{T}\mathbf{E}_\mathrm{g}}{\mathrm{n}} + \mathbf{w}^\mathrm{T}\mathbf{R}\mathbf{w} - \mathbf{2}\mathbf{w}^\mathrm{T}\tilde{\mathbf{w}} + 2\lambda \left\|\mathbf{w}\right\|_1$$

- Lassosum takes $R = (1-s)R_r + sI, 0 < s < 1; R_r = \frac{X_r^T X_r}{n_r}$;

Mak et al., Genetic Epidemiology 2017
https://choishingwan.github.io/PRS-Tutorial/lassosum/

Summary-level Bayesian Dirichlet Process Regression (*SDPR)*

- Multiple linear regression model for gene expression imputation

$$\boldsymbol{E_g = Xw + \varepsilon}$$

- Note that the least square estimate based on the multiple linear regression model is

$$\widehat{\boldsymbol{w}} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{E_g^T X} = \left(\frac{\boldsymbol{R^{-1}}}{\boldsymbol{n}}\right) \widetilde{\boldsymbol{w}}\mathbf{n} = \boldsymbol{R^{-1}}\widetilde{\boldsymbol{w}}; \qquad \widetilde{\boldsymbol{w}} = \boldsymbol{R}\widehat{\boldsymbol{w}}$$

- Nonparametric Dirichlet Process Prior is assumed for $\boldsymbol{w_j \sim N\left(0, \sigma_w^2\right), \sigma_w^2 \sim DP(H, \alpha)}$,
  - Base distribution $\boldsymbol{H = IG(a_0, b_0)}$
  - Concentration parameter $\boldsymbol{\alpha}$ controlling the shrinkage of the distribution on $\boldsymbol{\sigma_w^2}$ toward $\boldsymbol{H}$.

Summary-level Bayesian Dirichlet Process Regression (*SDPR)*

- SDPR uses marginal eQTL effect sizes $\tilde{w}$ as input data

$$\frac{\tilde{w}}{c}|w \sim N\left(Rw, \frac{R + naI}{n}\right)$$

  - Scale marginal effect size by a constant c to correct for deflation of summary statistics if double genomic control was applied for generating the summary data by meta-analysis, which can be estimated by SumHer (Speed D. et al. Nat. Genetics 2019)
  - Assume correlation between two SNPs is $\frac{R_{ij}}{1+na}$ rather than R_{ij} , with parameter a to shrink the correlation between two SNPs;
  - $\boldsymbol{I}$ is an identify matrix
- Estimate $\boldsymbol{w}$ by MCMC

Zhou et al., PLoS Genet. 2021
https://github.com/eldronzhou/SDPR

PRS with continuous shrinkage prior (*PRS-CS)*

- PRS-CS assumes a normal prior for $\boldsymbol{w_j}$ and non-informative scale-invariant Jeffreys prior on residual variance σ_ε^2 in the multiple linear regression model

$$\boldsymbol{E_g = Xw + \epsilon}, \qquad \boldsymbol{\epsilon \sim N(0, \sigma_\epsilon^2 I)}, \qquad \boldsymbol{p(\sigma_\epsilon^2) \propto \sigma_\epsilon^{-2}}$$

- Priors:

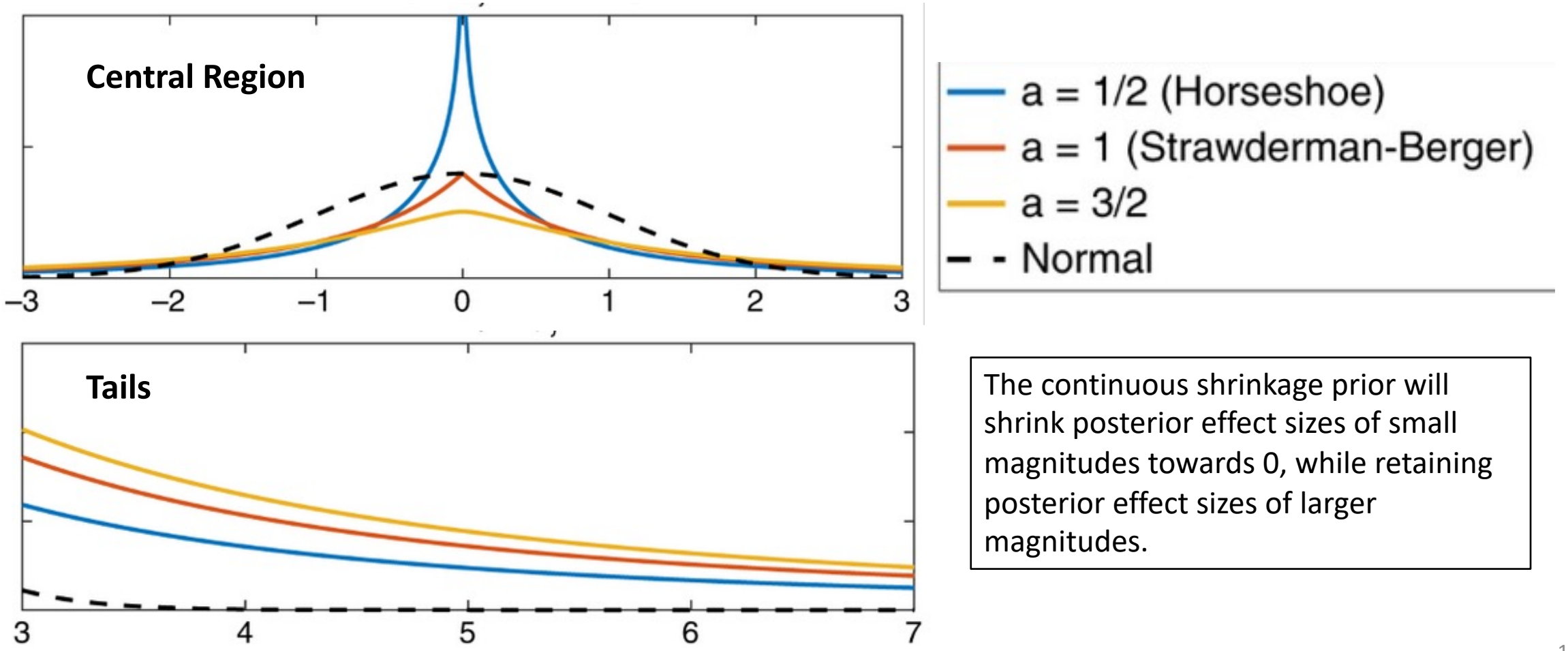
$$w_j \sim N\left(0, \frac{\boldsymbol{\sigma_\epsilon^2}}{n}\phi\psi_j\right);\ \psi_j \sim \text{Gamma}(a, \delta_j), \delta_j \sim Gamma(b, 1)$$

  - ψ_j: local shrinkage parameter with a gamma-gamma prior distribution, Gamma(shape parameter, scale parameter)
  - ϕ: global shrinkage parameter controlling the overall sparsity of $\boldsymbol{w}$

Ge et al. Nat. Commun. 2019
https://github.com/getian107/PRScs

PRS with continuous shrinkage prior (*PRS-CS)*

Marginal prior density on the effect size w_j with $\psi_j = 1, b = 0.5, a = (0.5, 1, 3/2)$



The continuous shrinkage prior will shrink posterior effect sizes of small magnitudes towards 0, while retaining posterior effect sizes of larger magnitudes.

PRS with continuous shrinkage prior (*PRS-CS)*

- Posterior estimates of $\boldsymbol{w}$ can be obtained by using only summary-level eQTL data $\tilde{\boldsymbol{w}}$ and reference genotype correlation matrix $\boldsymbol{R}$ by Gibbs Sampler

$$\mathbf{w} \mid \sigma_\epsilon^2, \boldsymbol{\Psi}, \tilde{\mathbf{w}}, \mathbf{R} \sim \mathrm{MVN}\left(\frac{\mathrm{n}}{\sigma_\epsilon^2}\boldsymbol{\Sigma}\tilde{\mathbf{w}}, \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{\Sigma} = \frac{\sigma_\epsilon^2}{\mathrm{n}}(\mathbf{R} + \boldsymbol{\Psi}^{-1})^{-1}, \boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_\mathrm{m})$$

TWAS Stage II: GReX Prediction and Testing

- Based on $\widehat{\boldsymbol{w}}$, impute gene expression in GWAS cohort as $\widehat{\boldsymbol{GReX}} = \boldsymbol{X}_{\boldsymbol{test}}\widehat{\boldsymbol{w}}$
  - $\boldsymbol{X}_{\boldsymbol{test}}$: Genotype matrix in the test (GWAS) cohort
- Test association between $\widehat{\boldsymbol{GReX}}$ and phenotype $\boldsymbol{Y}$, adjusting for covariates $\boldsymbol{Z}$, using regression model:

$$\boldsymbol{E}[\boldsymbol{g}(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{test}})] = \boldsymbol{\beta}\widehat{\boldsymbol{GReX}} + \boldsymbol{\alpha}'\boldsymbol{Z} = \boldsymbol{\beta}(\boldsymbol{X}_{\boldsymbol{test}}\widehat{\boldsymbol{w}}) + \boldsymbol{\alpha}'\boldsymbol{Z}$$

- **TWAS tests $\boldsymbol{H_0}: \boldsymbol{\beta} = \boldsymbol{0}$** using score/Wald/LR test from regression model

TWAS Stage II: Testing with Summary-level GWAS Data

- With summary-level GWAS data (Z scores), one can instead test

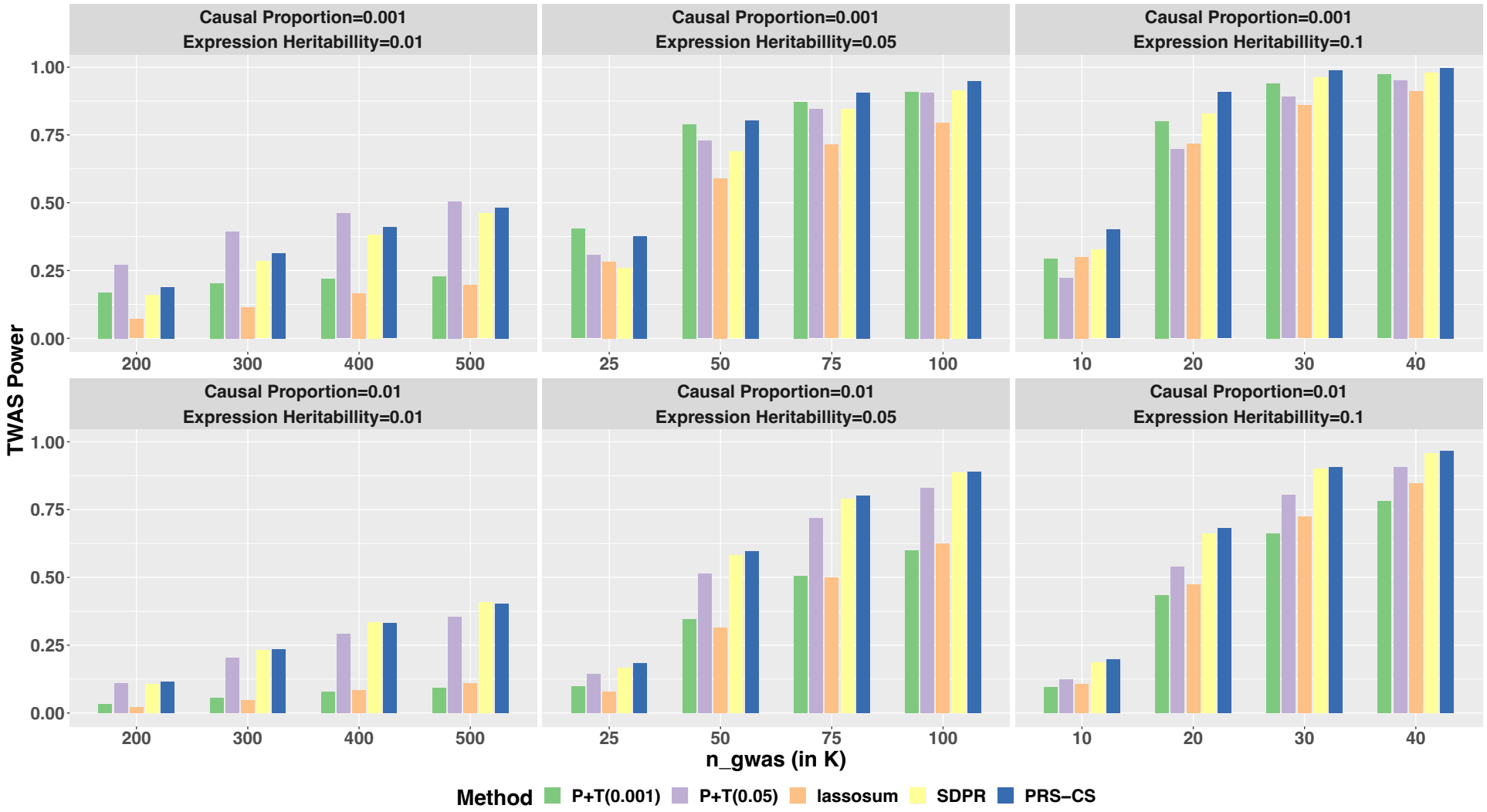
$$Z_{g,FUSION} = \frac{\sum_{j=1}^{J}(\widehat{w}_j Z_j)}{\sqrt{\widehat{\boldsymbol{w}}' \boldsymbol{V} \widehat{\boldsymbol{w}}}}$$

  - Z_j: Single-variant Z score in GWAS for test SNP ("eQTL") j .
  - $\boldsymbol{V}$: Genotype correlation matrix, which can be approximated from reference genotype panel (e.g. 1000 Genome).
  - Centered and standardized gene expression and genotypes were assumed for deriving estimates of eQTL weights.

Gusev et al. (Nat Gen: 48: 245)

Which PRS method to use?

- **Classical:** P+T (0.001) P+T (0.05)
- **Penalized Regression:** lassosum
- **Bayesian Regression:** SDPR, PRS-CS



OTTERS Motivation

- **Issue:** Performance using individual training method dependents on the underlying genetic architecture of gene expression
  - Idea: Create an omnibus test by combining individual TWAS p-values together using Aggregated Cauchy Association Test (ACAT)

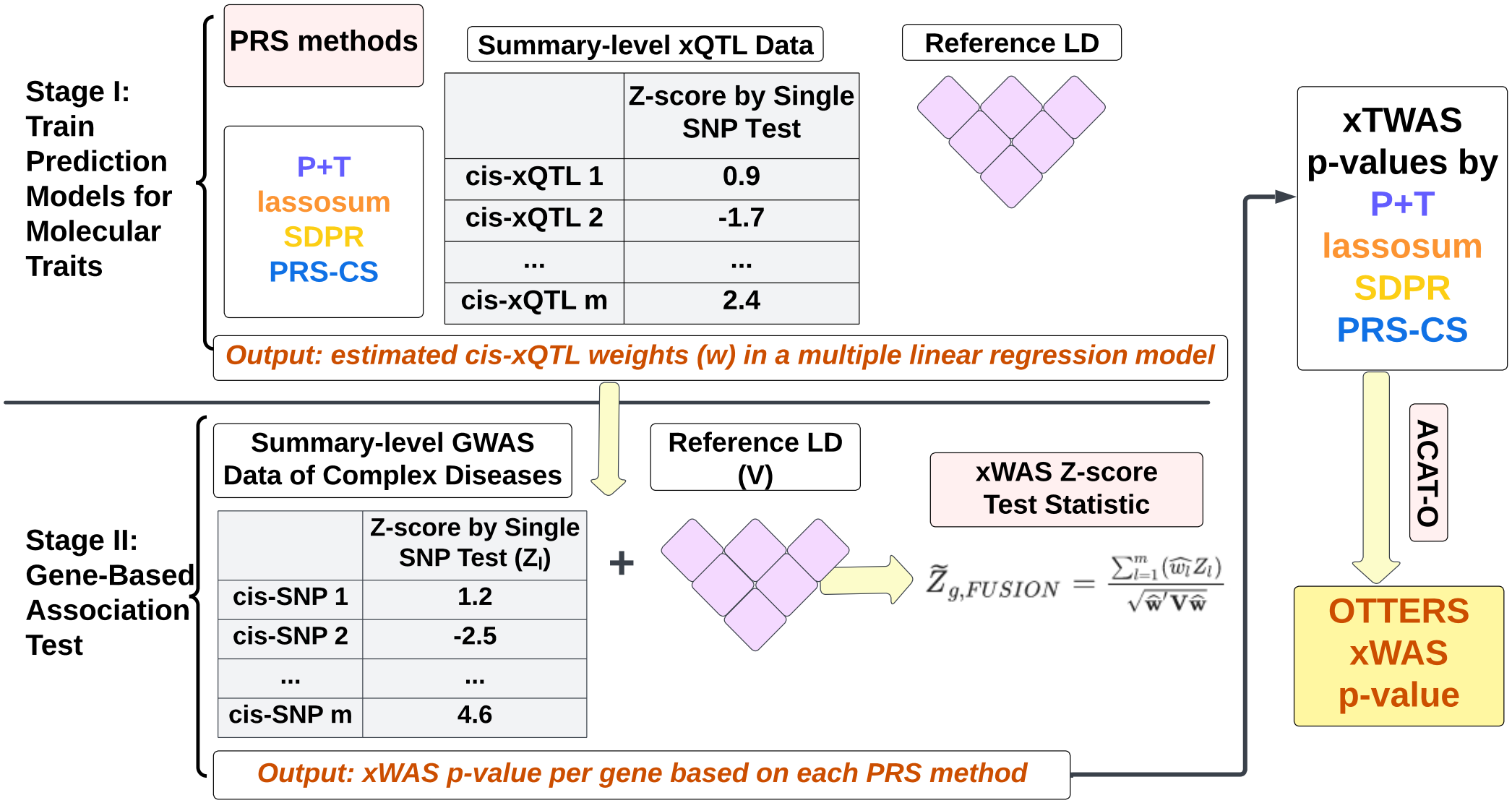
- **O**mnibus **T**ranscriptome **T**est using **E**xpression **R**eference **S**ummary data (**OTTERS**)

Liu et al. [AJHG 104: 410]

OTTERS p-value

- $\boldsymbol{p_k}$: TWAS p-value with eQTL weights trained under k^{th} method
  - $\tan\{(0.5 - p_k)\}$ follows Cauchy under null
- $\boldsymbol{T_{OTTERS}} = \frac{1}{K}\sum_{k=1}^{K} \tan\{(0.5 - p_k)\pi\}$, $\pi \approx 3.14$
  - Can be approximated by Cauchy distribution
- **OTTERS p-value**: $p_{OTTERS} \approx \frac{1}{2} - \frac{\{\arctan(T_{OTTERS})\}}{\pi}$

OTTERS framework



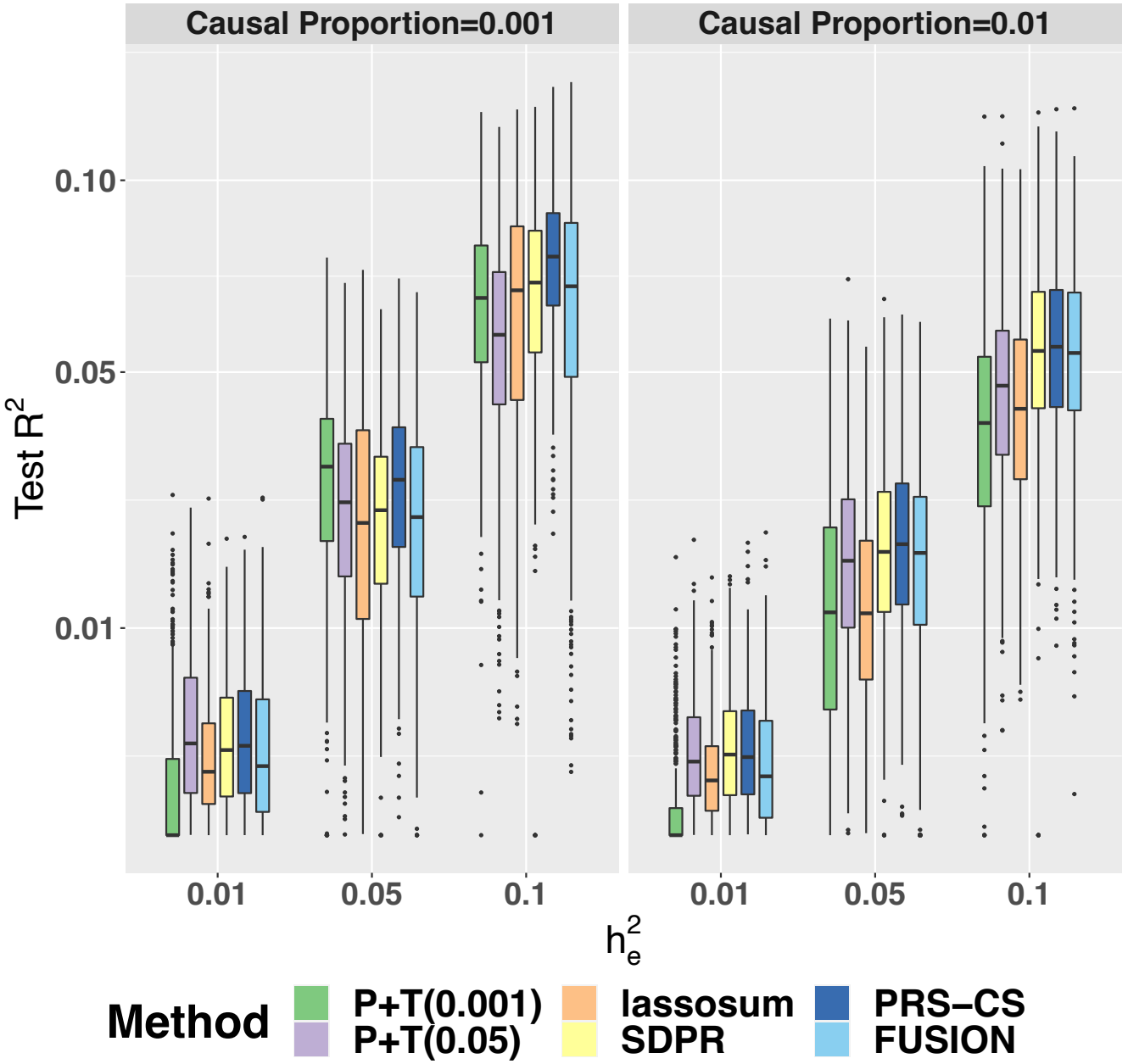
Simulations

- 1894 WGS European subjects
  - Randomly select 30% for training and 70% for prediction to mimic real training sample size
- Genotypes: randomly select 500 genes and their cis-SNPs
  - Randomly select causal eQTLs: $p_{causal} = (0.001, 0.01)$
  - Proportion of expression variance explained by causal eQTLs: $h_e^2 = (0.01, 0.05, 0.1)$
- Generated GWAS Z-scores: $Z \sim MVN(\mathrm{R_g w}\sqrt{n_{gwas} * h_p^2},\ \mathrm{R_g})$
  - $h_p^2 = 0.025$: Proportion of phenotypic variance explained by simulated expression $(\mathrm{X_g w})$
  - n_{gwas}: GWAS sample size
  - $\mathrm{R_g}$: Correlation matrix of standardized genotype matrix $\mathrm{X_g}$

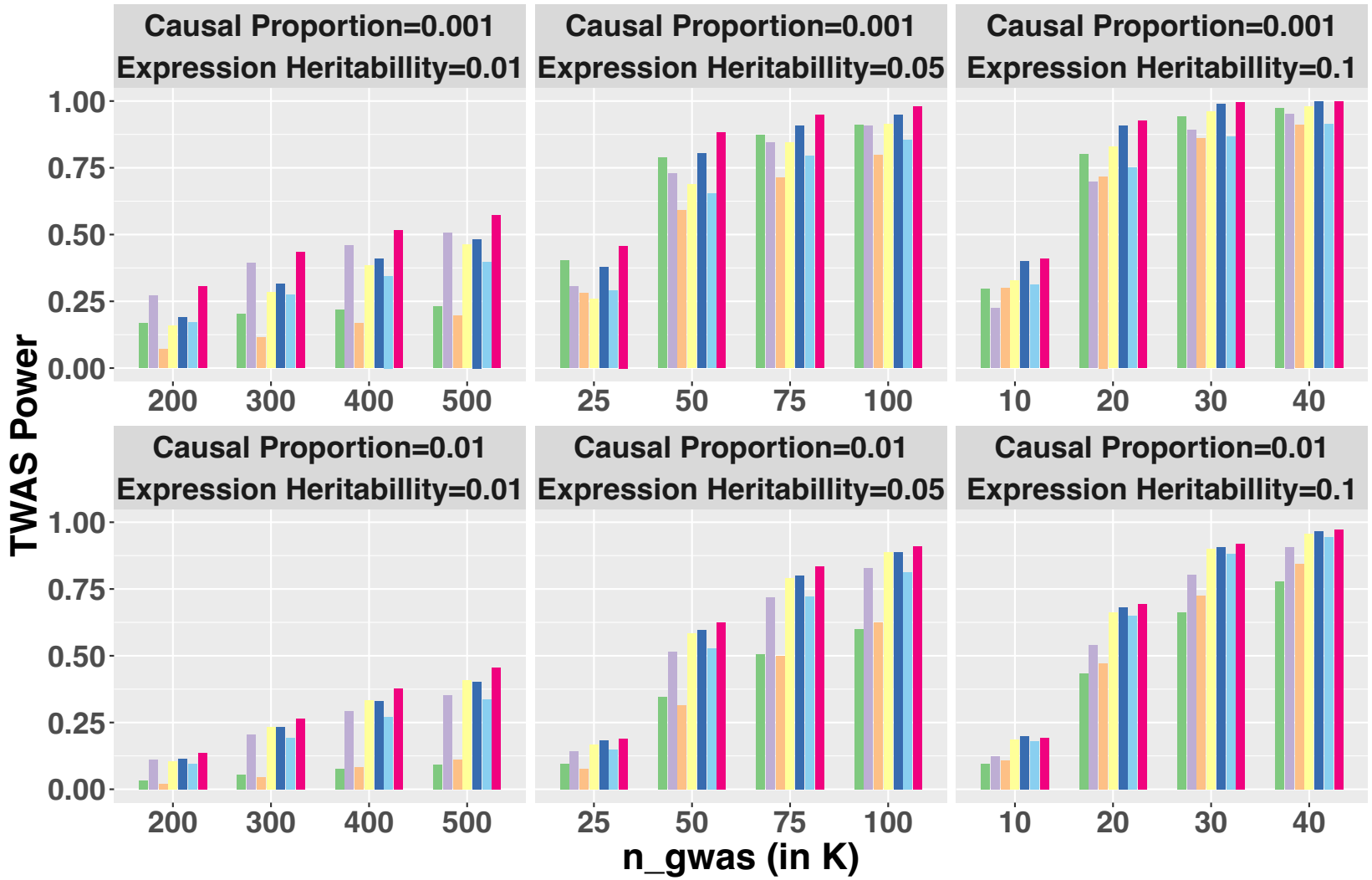
Feng et al. (PLoS Genet: 17: e1008973)

Simulation Results (Test R^2)

- 568 training samples
- 1326 test samples
- Imputation models fit using summary-level eQTL results from simulated training data
- FUSION uses individual-level reference eQTL data

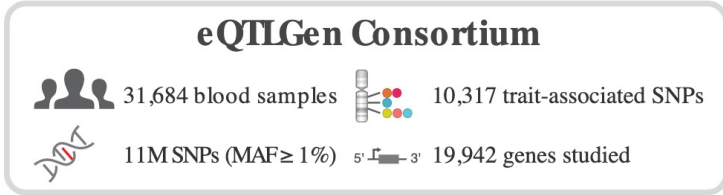


TWAS Power with h_p^2= 0.025, $\alpha = 2.5 \times 10^{-6}$

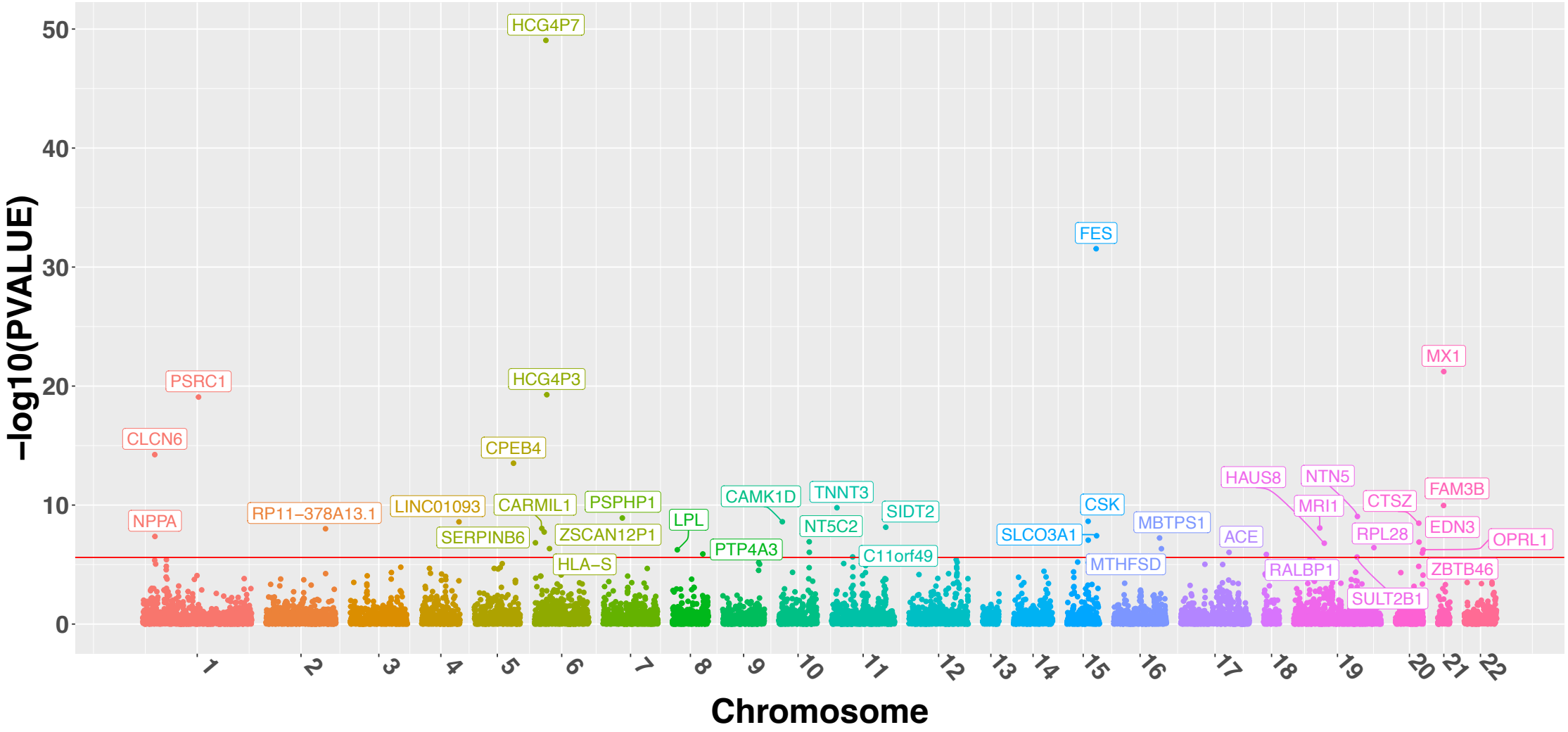


Application to UKBB GWAS Data

- Summary-level GWAS results for cardiovascular disease from UK Biobank (n=459,234; case fraction=0.319)
- eQTL summary-level reference data from eQTLGen (n=31,684)
  - Whole blood tissue
  - Trained 5 GReX models by P+T (0.001), P+T (0.05), lassosum, SDPR, PRS-CS
  - Used reference LD panel from 838 GTEx V8 WGS samples
- TWAS p-value obtained for each training model
  - Performed genomic control for statistics under each mode
- Combined adjusted p-values across training models using OTTERS



TWAS Results of Cardiovascular Disease by OTTERS



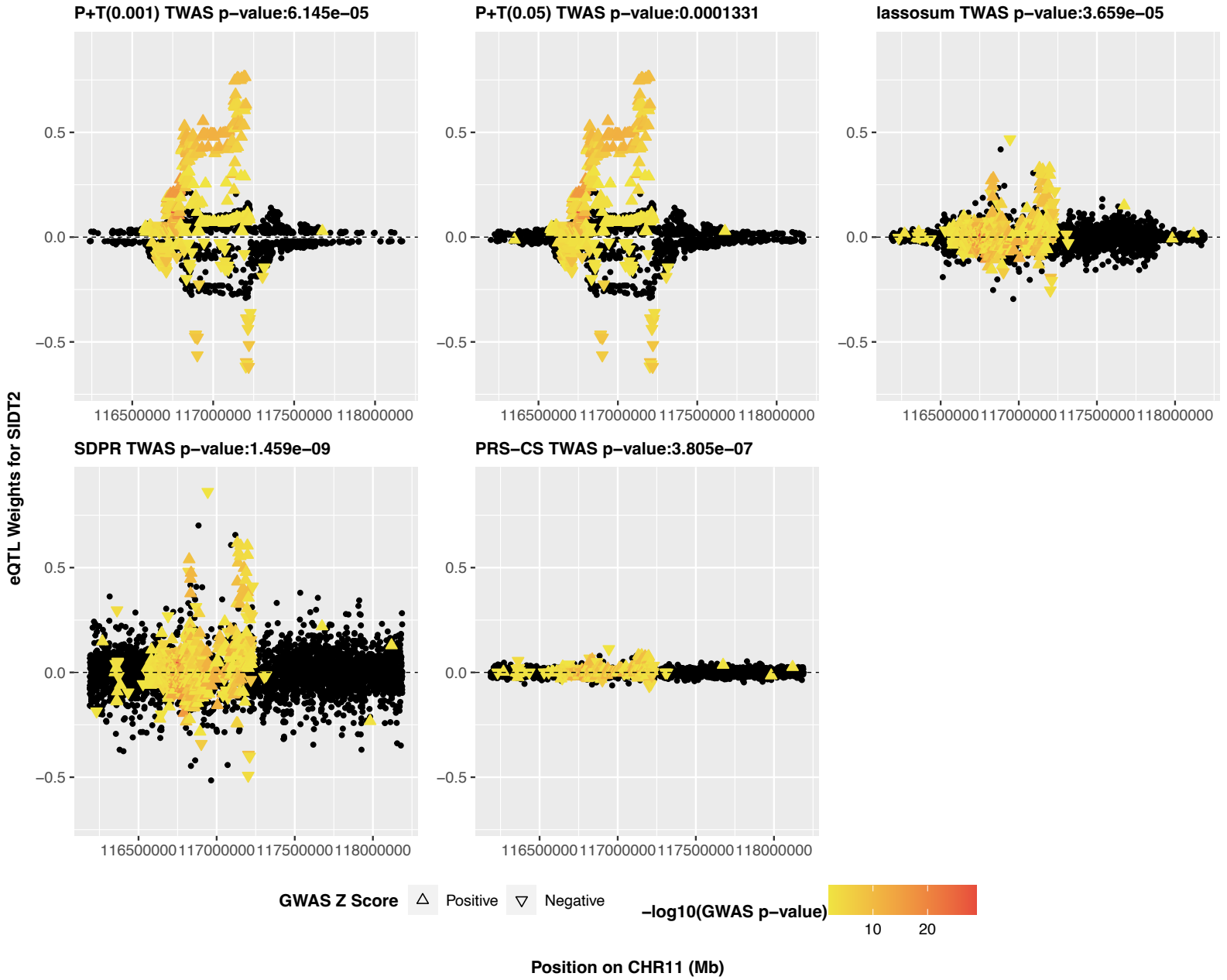
TWAS Results by OTTERS

- OTTERS identified 38 independently significant risk genes for cardiovascular disease using eQTLGen (n=31,684)
  - a: Risk genes (8) identified in TWAS-hub using GTEx whole blood tissue.
  - b: Risk genes (13) identified in TWAS-hub using other GTEx tissue types.
  - c: Novel risk genes (17)

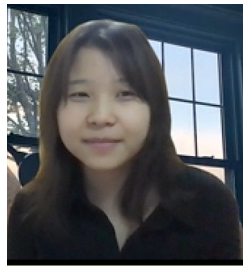
- Optimal individual training method varied by gene
  - OTTERS borrows strength across individual methods

CHROM	ID	OTTERS	P+T(0.001)	P+T(0.05)	lassosum	SDPR	PRS-CS
1	CLCN6[a]	**5.75E-15**	**4.94E-09**	**5.40E-08**	**8.77E-09**	**1.19E-15**	**1.43E-09**
1	NPPA[b]	**4.32E-08**	**1.55E-08**	**2.14E-07**	-	-	6.71E-06
1	PSRC1[a]	**8.37E-20**	**5.68E-08**	**8.46E-07**	**6.26E-11**	**1.67E-20**	**1.41E-12**
2	RP11-378A13.1[a]	**9.78E-09**	3.97E-02	4.98E-02	1.62E-05	**1.96E-09**	1.15E-04
4	LINC01093[c]	**2.57E-09**	9.85E-02	5.31E-02	**5.13E-10**	1.08E-02	2.41E-02
5	CPEB4[b]	**3.05E-14**	1.26E-02	2.05E-02	2.70E-05	**6.05E-15**	**1.60E-07**
6	SERPINB6[c]	**1.47E-07**	2.12E-01	2.24E-01	7.56E-03	**2.95E-08**	7.53E-04
6	CARMIL1[c]	**9.23E-09**	5.34E-03	3.41E-03	4.15E-03	**1.85E-09**	1.72E-03
6	ZSCAN12P1[c]	**1.84E-08**	6.00E-01	5.75E-01	4.62E-01	**3.67E-09**	3.10E-01
6	HCG4P7[c]	**8.93E-50**	3.70E-01	3.69E-01	2.30E-01	**1.79E-50**	7.26E-01
6	HCG4P3[c]	**5.33E-20**	4.20E-01	4.05E-01	5.03E-04	**1.07E-20**	2.42E-03
6	HLA-S[c]	**4.57E-07**	7.13E-01	7.31E-01	3.02E-01	**9.14E-08**	2.33E-01
7	PSPHP1[c]	**1.21E-09**	2.17E-01	2.26E-01	9.65E-03	**2.43E-10**	1.10E-01
8	LPL[c]	**5.73E-07**	1.78E-03	3.26E-03	4.44E-02	**1.15E-07**	1.05E-04
8	PTP4A3[c]	**1.28E-06**	8.13E-02	8.33E-02	6.23E-05	**2.58E-07**	1.67E-03
10	CAMK1D[a]	**2.51E-09**	3.83E-02	4.97E-02	1.23E-03	**5.03E-10**	4.97E-05
10	NT5C2[b]	**1.21E-07**	**1.69E-06**	**2.92E-06**	1.64E-05	**3.15E-07**	**2.69E-08**
11	TNNT3[b]	**1.67E-10**	**1.09E-06**	3.33E-06	**2.03E-09**	**3.40E-11**	**4.01E-07**
11	C11orf49[b]	**2.28E-06**	**8.55E-07**	**1.78E-06**	5.44E-05	-	2.93E-04
11	SIDT2[a]	**7.26E-09**	6.14E-05	1.33E-04	3.66E-05	**1.46E-09**	**3.81E-07**
15	CSK[b]	**2.30E-09**	**1.70E-07**	**2.15E-06**	**7.41E-10**	**2.80E-09**	**2.17E-09**
15	FES[b]	**2.87E-32**	**4.78E-08**	**1.23E-06**	**9.13E-24**	**5.75E-33**	**1.94E-15**
15	SLCO3A1[c]	**3.78E-08**	1.85E-02	3.15E-02	4.65E-05	**7.57E-09**	1.14E-03
16	MBTPS1[b]	**5.80E-08**	2.62E-01	3.05E-01	9.15E-04	**1.16E-08**	2.34E-03
16	MTHFSD[a]	**4.65E-07**	5.16E-02	5.94E-02	1.65E-02	**9.30E-08**	3.20E-03
17	ACE[b]	**9.42E-07**	4.93E-06	1.03E-05	4.23E-06	**9.66E-07**	**2.68E-07**
18	RALBP1[c]	**1.40E-06**	1.48E-01	1.54E-01	2.12E-04	**2.81E-07**	5.55E-03
19	MRI1[b]	**8.38E-09**	8.34E-03	1.60E-02	7.79E-03	**1.68E-09**	2.65E-03
19	HAUS8[b]	**1.60E-07**	**4.41E-08**	**1.38E-07**	**1.67E-06**	**1.42E-06**	3.29E-05
19	SULT2B1[c]	**2.32E-06**	**7.73E-07**	-	-	2.97E-02	1.10E-02
19	NTN5[a]	**9.03E-10**	**2.75E-08**	**1.16E-07**	6.23E-06	**1.85E-10**	**9.73E-09**
19	RPL28[b]	**3.76E-07**	7.33E-02	1.16E-01	6.64E-03	**7.52E-08**	4.23E-03
20	CTSZ[b]	**3.32E-09**	2.57E-02	1.99E-02	**3.40E-09**	**8.25E-10**	1.04E-01
20	EDN3[c]	**1.29E-07**	**3.61E-08**	**9.15E-08**	8.60E-06	5.90E-03	1.58E-02
20	ZBTB46[c]	**1.07E-06**	**2.83E-07**	8.35E-06	-	1.81E-03	1.27E-05
20	OPRL1[a]	**5.84E-07**	**3.44E-07**	**2.69E-06**	1.85E-03	5.51E-05	**1.90E-07**
21	FAM3B[c]	**1.08E-10**	2.28E-02	2.58E-02	8.07E-06	**2.17E-11**	1.04E-05
21	MX1[c]	**6.04E-22**	4.36E-01	3.83E-01	**3.16E-07**	**1.21E-22**	1.24E-03

Example Scatter Plots of eQTL Weights



Dr. Michael Epstein



Qile Dai

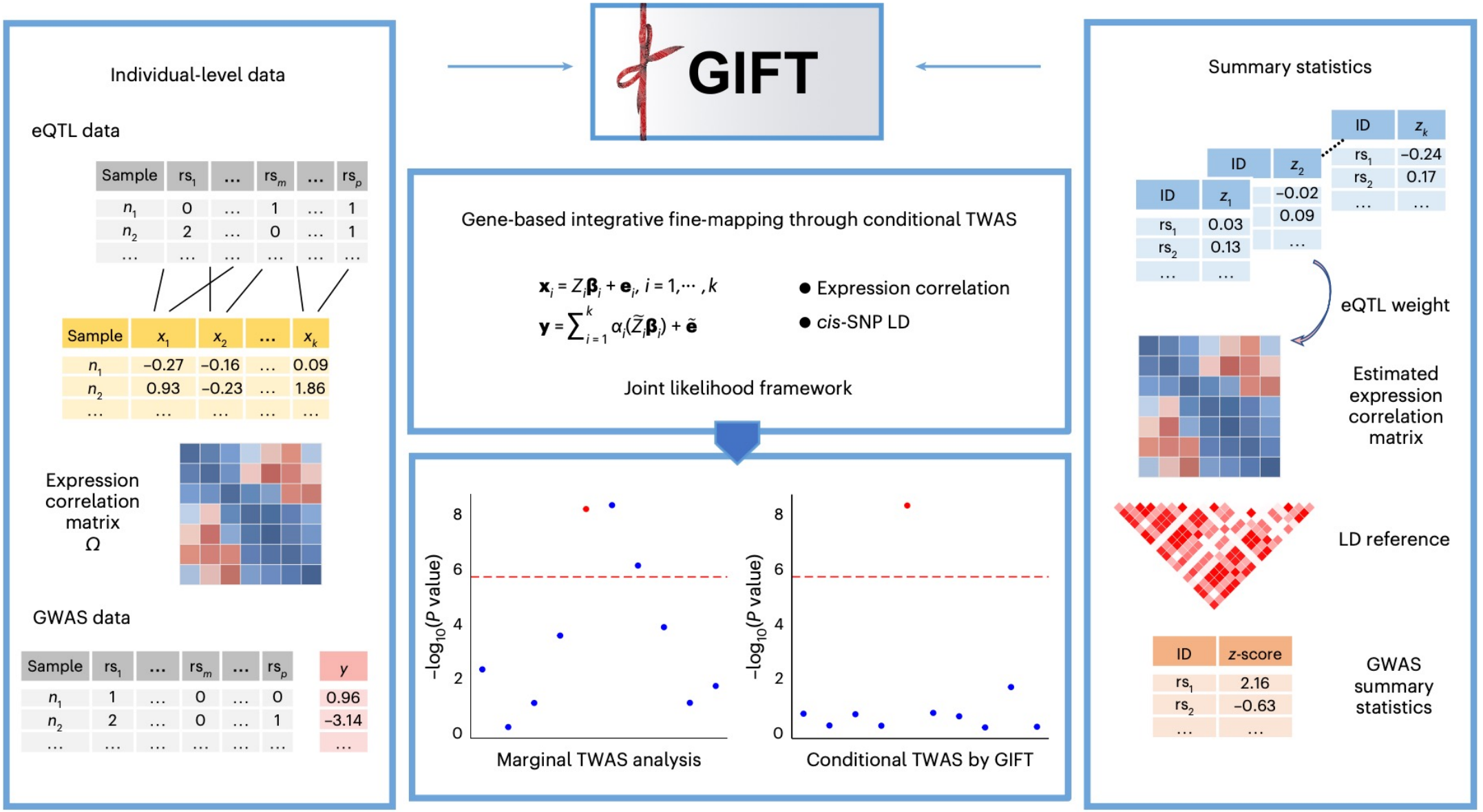
Fine-mapping TWAS

- TWAS tests genes-trait association independently
- TWAS results often contain a cluster of significant genes in the same region
  - Test SNPs are overlapped
  - Gene expression are correlated
  - Predicted genetically regulated gene expression components are correlated
- Fine map for potential “causal” or independent significant TWAS risk gene

Fine-mapping TWAS Methods

- **FOCUS** (fine-mapping of causal gene sets; Mancuso N. et al. Nat. Genet. 2019): Extends the Bayesian sparse model used in GWAS fine-mapping to TWAS
- **FOGS** (fine-mapping of gene sets; Wu C. et al. Hum. Genet. 2020): Performs conditional SNP association analysis and then aggregates the conditional SNP-trait associations into gene-trait associations
- **GIFT** (gene-based integrative fine-mapping through conditional TWAS; Liu L. et al. Nat. Genet. 2024): Models the GReX of all genes residing in the focal region and carries out TWAS conditional analysis in a maximum likelihood framework.

GIFT: Gene-based Integrative Fine-mapping through conditional TWAS



Statistical Model of GIFT

- **Gene Expression Prediction Model:**

$$\mathbf{x}_i = Z_i\boldsymbol{\beta}_i + \mathbf{e}_i, i = 1, \cdots, k,$$

  - $\boldsymbol{x_i}$: Expression level of the i^{th} gene
  - Z_i : Matrix of cis-SNPs for gene i
  - $\boldsymbol{\beta_i}$: Effect sizes of the cis-SNPs
  - e_i : Residual errors.

- **Gene-Trait Association Model:** $\mathbf{y} = \sum_{i=1}^{k} \alpha_i(\tilde{Z}_i\boldsymbol{\beta}_i) + \tilde{\mathbf{e}},$
  - $\boldsymbol{y}$: Trait outcome
  - α_i : Effect sizes of the GReX of gene i on the trait
  - $\tilde{Z}_i$: Genotype matrices in the test GWAS data
  - $\tilde{\boldsymbol{e}}$: Residual error

Note that the notations here are matched with the GIFT paper, which is different from the notations in our previous slides.

With eQTL and GWAS Summary Data

- **Gene Expression Model:** $\hat{\mathbf{z}}_{x_i} = \sqrt{(n_1 - 1)}\Sigma_{1i}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_{x_i}, i = 1, \cdots, k,$
  - $\widehat{Z_{x_i}}$: eQTL Z-score vector for the i_{th} gene
  - $\boldsymbol{\Sigma}_{1i}$: SNP-SNP correlation matrix for the i_{th} gene's with dimension $p_i \times p_i$
  - β_i: Marginal effect sizes of cis-SNPs on the i_{th} gene's expression.
  - ε_{x_i}: Residual errors for the i_{th} gene

- **Gene-Trait Model:** $\hat{\mathbf{z}}_y = \sqrt{(n_2 - 1)}\Sigma_2\left(\alpha_1\boldsymbol{\beta}_1^T, \cdots, \alpha_k\boldsymbol{\beta}_k^T\right)^T + \boldsymbol{\varepsilon}_y,$
  - $\widehat{Z_y}$: Stacked GWAS Z-score vector of all genes
  - $\boldsymbol{\Sigma}_2$ **:** SNP-SNP correlation matrix for stacked cis-SNPs across all genes with dimension $\sum_{i=1}^{k} p_i \times \sum_{i=1}^{k} p_i$
  - $\alpha_i\beta_i$: Product of GReX effect size and SNP effect sizes for the i_{th} gene
  - ε_y: Residual errors in GWAS

GIFT

- Test H_0: $\alpha_i = 0, i = 1, \ldots, k$
- Derive the likelihood framework through joint modeling the gene expression model and the gene-trait model
  - Accounts for the uncertainty of eQTL weight estimation
- Uses Expectation Maximization algorithm to maximize the joint likelihood to estimate $\{\alpha_i, i = 1, \ldots, k\}$ and calculates the maximum likelihood under the full model
- Calculates the maximum likelihood under a series of reduced model, assuming the test H_0: $\alpha_i = 0$.
- Conducts likelihood ratio tests

Apply GIFT to Study Blood Pressure and Lipid Traits with the UK Biobank Data

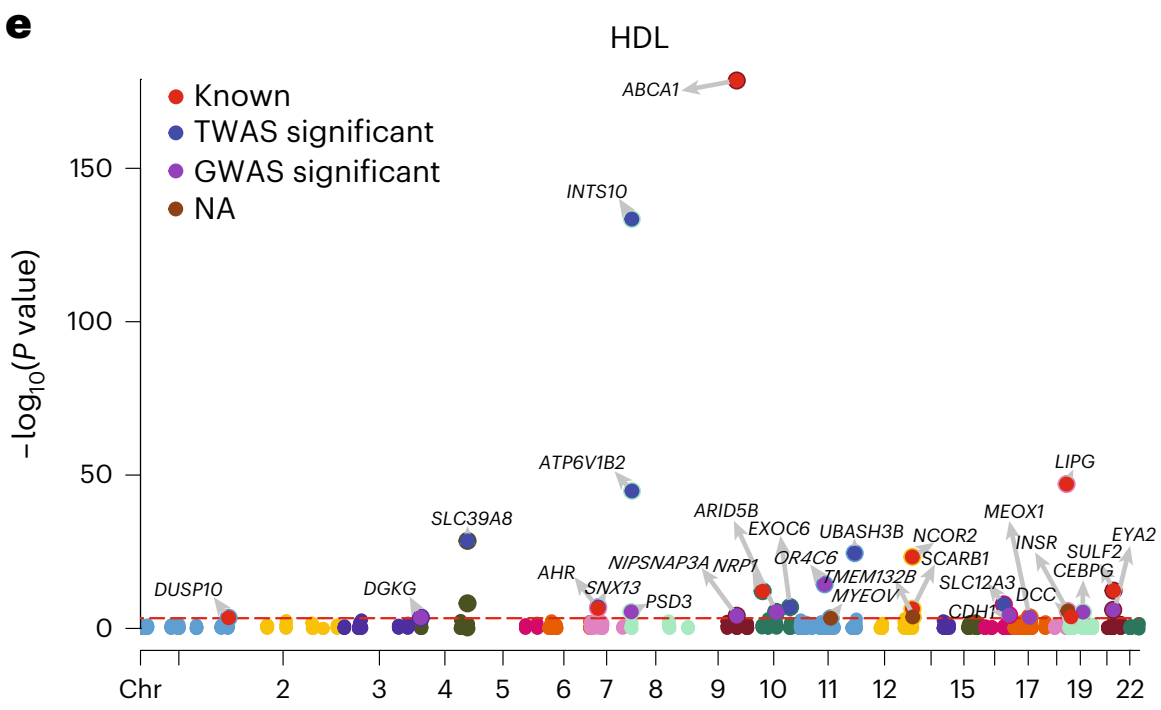
- Apply GIFT to studying 15,577 genes residing in 1,533 independent LD regions

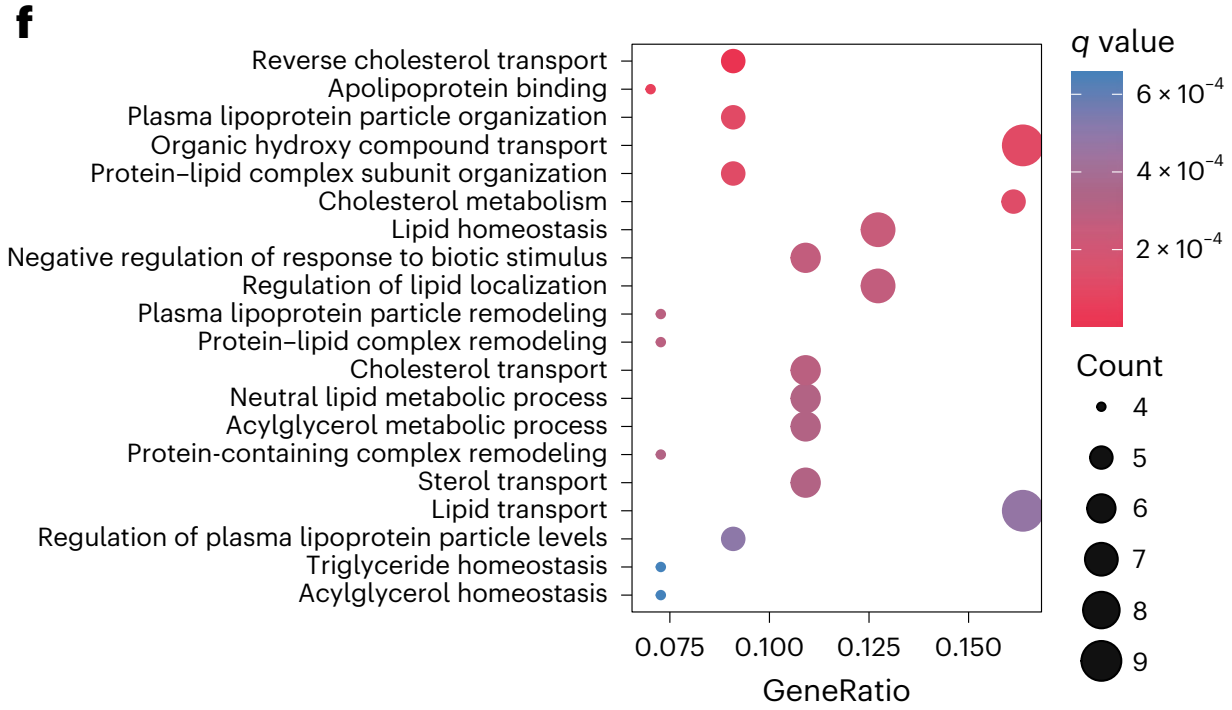
Table 1 | Summary results of fine-mapping in the real data

Trait	GWAS risk regions	Significant TWAS genes	TWAS risk regions	GWAS risk regions with TWAS significant genes	GIFT	FOCUS	FOGS	MV-IWAS
SBP	147	146	85	53	7	72	350	409
DBP	191	181	87	67	13	75	356	691
TC	147	286	108	82	19	136	599	804
HDL	231	473	152	120	28	178	937	1,320
LDL	122	226	93	72	16	116	455	679
TG	200	393	124	96	31	149	753	922

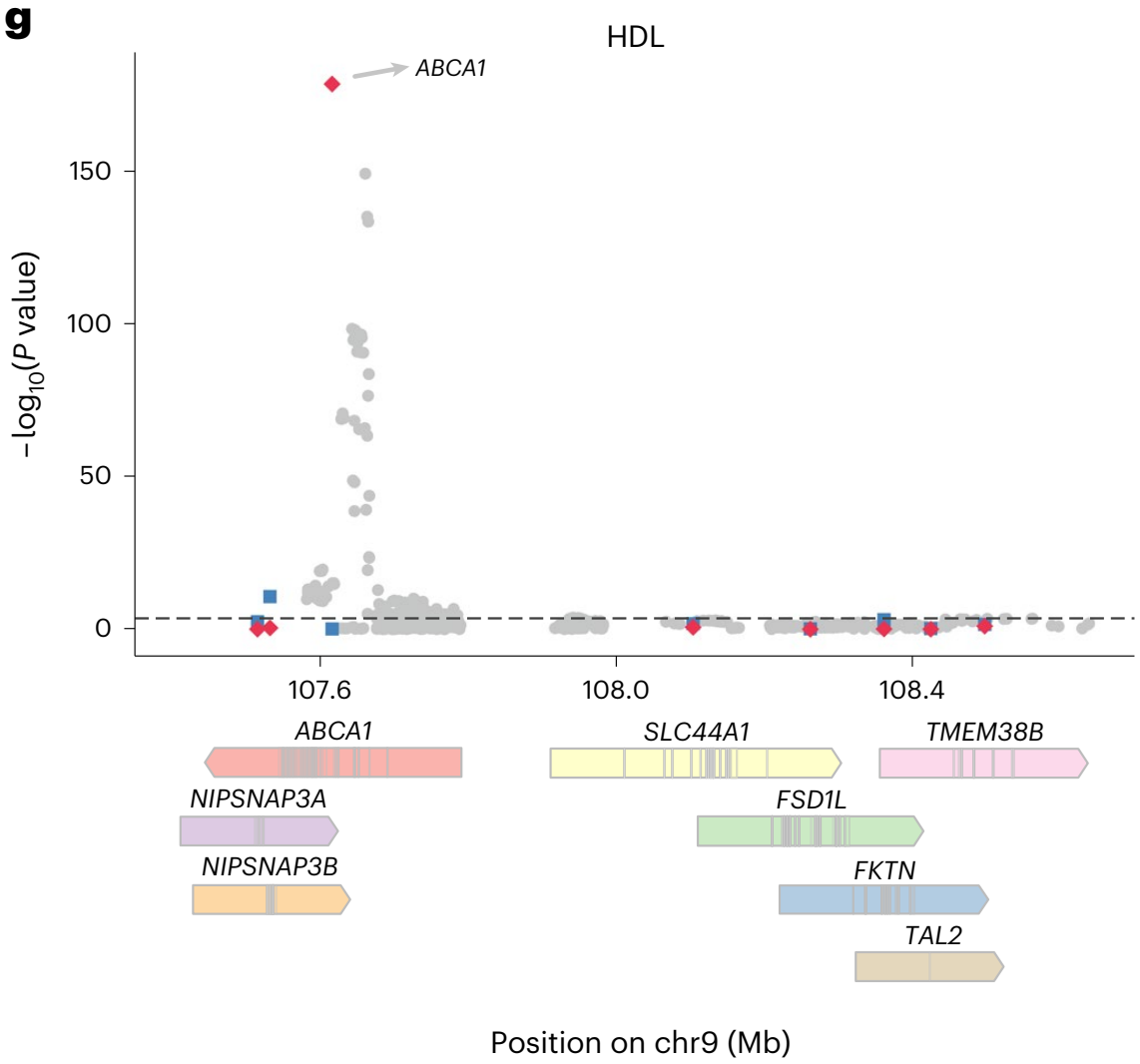
The table shows the number of discoveries for each of the six traits (rows) in TWAS fine-mapping analysis carried out by each of the four methods (columns). A GWAS risk region (first column) is defined as an LD block that harbors at least one genome-wide significant SNP ($P<5\times10^{-8}$). A significant TWAS gene (second column) is defined as a gene with a marginal TWAS $P<0.05/15{,}577$. A TWAS risk region (third column) is defined as an LD block that harbors at least one marginal TWAS significant gene. The last four columns list the number of genes discovered by each of the four methods for the six traits. We used an empirical FDR threshold of 0.05 to declare significance for all methods in the fine-mapping analysis.

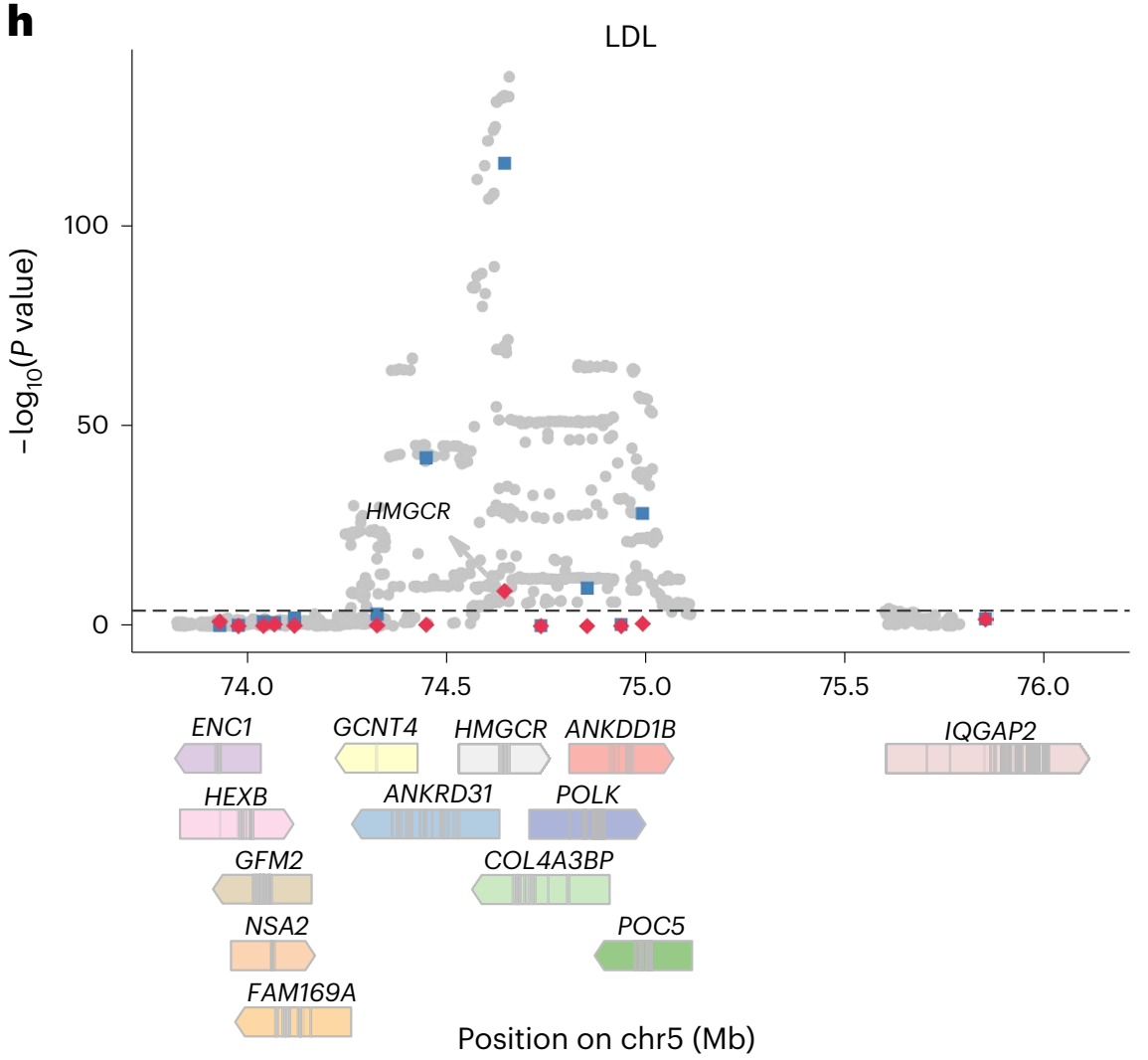
GIFT Results for HDL





Example GIFT Results for HDL and LDL





Web Resources

- OTTERS
  - https://github.com/daiqile96/OTTERS
- GIFT
  - https://yuanzhongshang.github.io/GIFT/