SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification

Chong Wu

Department of Biostatistics

The University of Texas MD Anderson Cancer Center

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

University of Hawaii Cancer Center

March 21, 2023

Research goal

My long-term research goal is to develop new methods, theories, and software to:

- identify likely causal risk factors and biomarkers for a complex disease (prostate cancer, Alzheimer's disease, etc.)
- enhance risk prediction to advance precision medicine

Research Interests: causal inference (Mendelian randomization), machine learning, statistical genetics (polygenic risk score, integrative analysis, TWAS, PWAS)

Data we work on: UK Biobank (genotype, risk factors, & disease status), GTEx (splicing, gene expression, & genotype), ROS/MAP (protein & genotype), GWAS summary data, functional annotations, DNA methylation

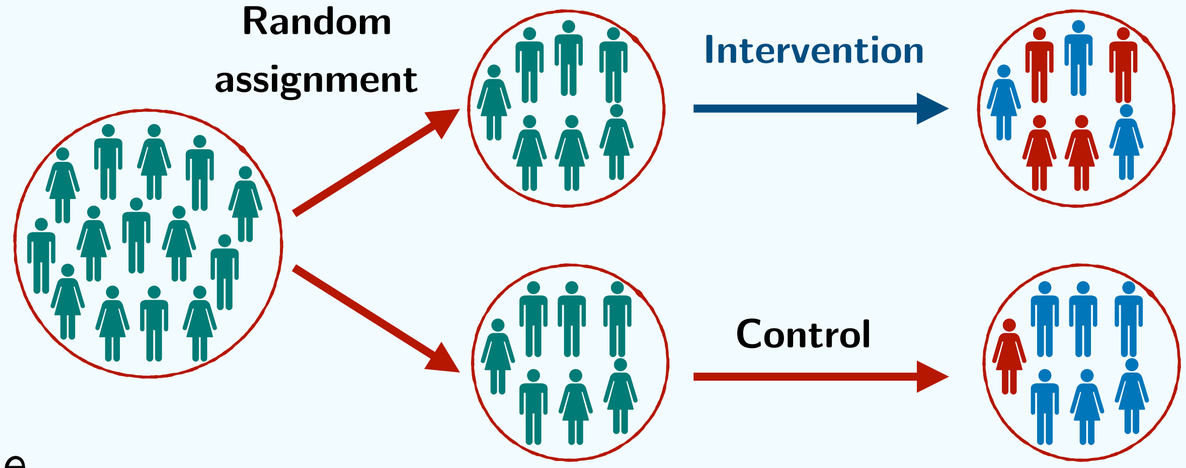
Outline

- **Background**
- New method: SUMMIT
- Results
- Extension

Causal inference in observational data

Does X (risk factor) cause Y (complex disease)?

- **Example:** Does smoking cause lung cancer?
- Randomized clinical trial
  - Gold standard
  - Randomization balances participant characteristics between the groups
- **Challenges:** randomized clinical trial would be both not feasible and unethical



Causal inference in observational data

Example: identify causal biomarkers for a complex disease

Why:

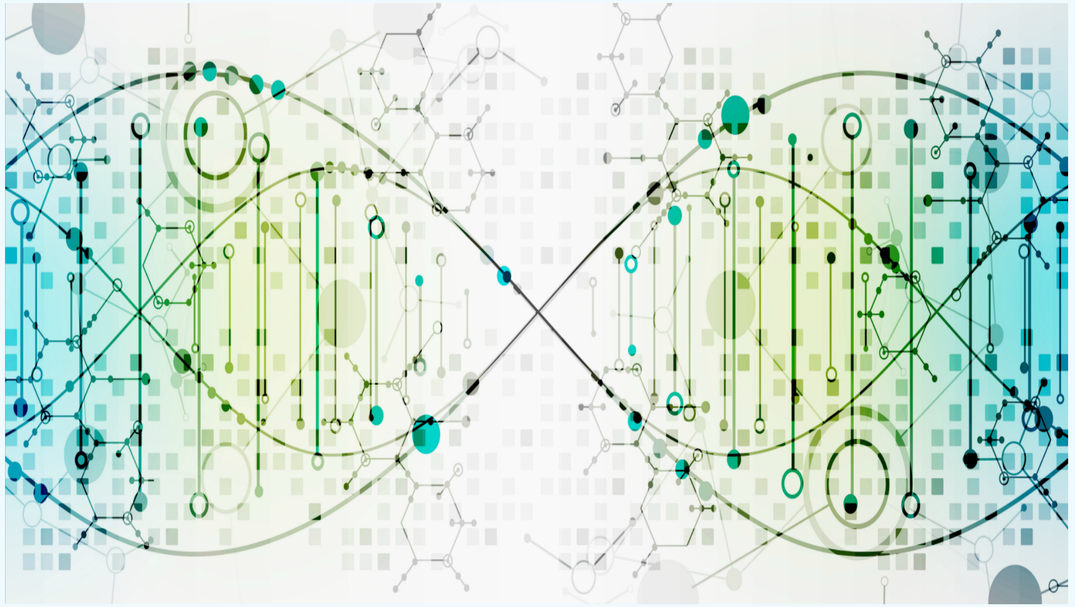
- understand the etiology
- drug development

Challenges:

- the number of biomarkers is large
- biomarkers are correlated

Goal:

identify likely causal biomarkers by using observational data

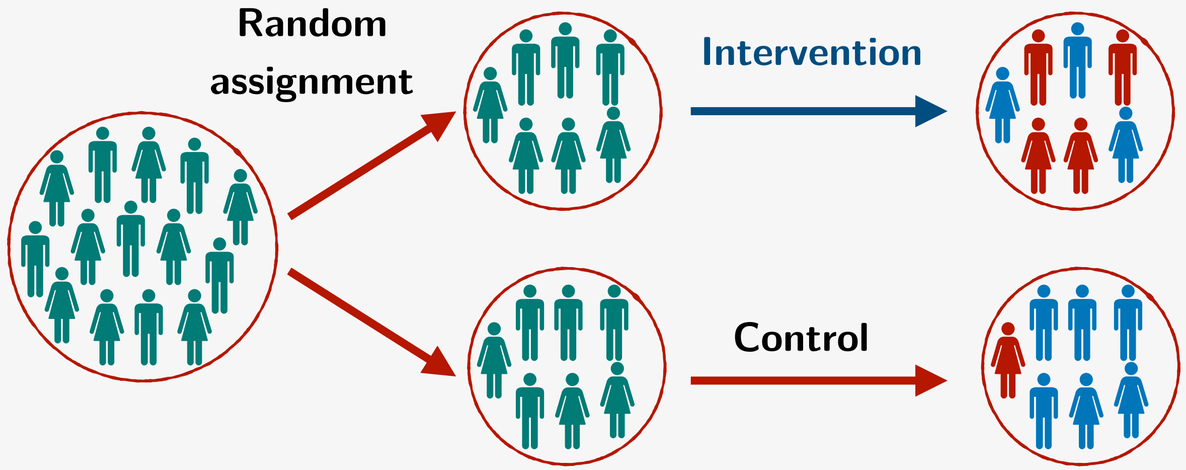


This figure is downloaded from Google Image

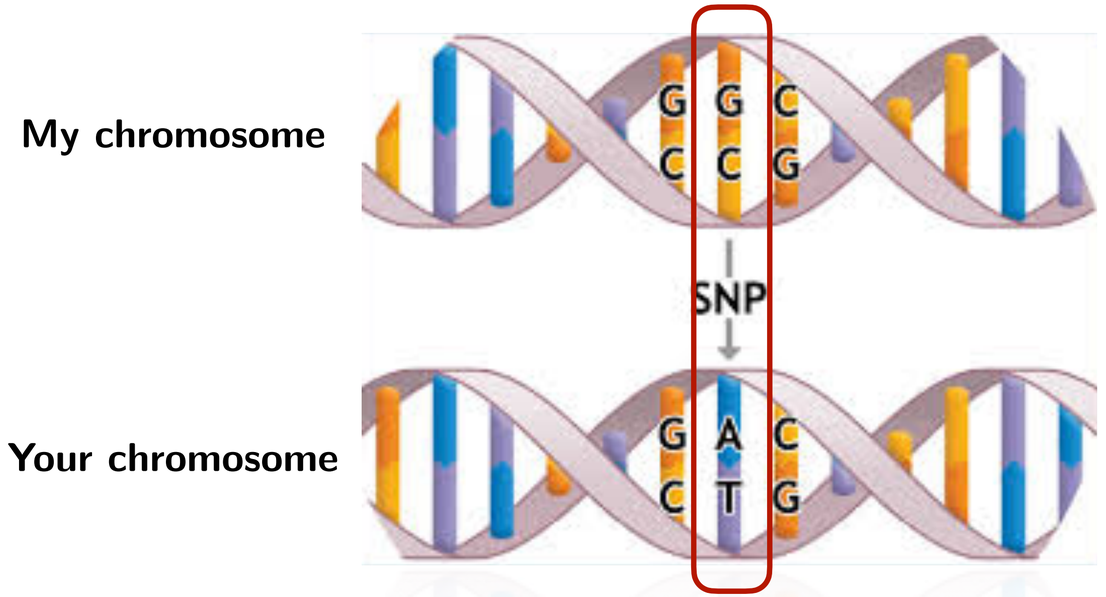
Mendelian randomization

Randomized clinical trial

- Gold standard
- Randomization balances participant characteristics between the groups



- Genome: genetic information encoded in 23 chromosome pairs
- SNP
  - variation in a single base pair
  - inherited randomly and fixed at conception

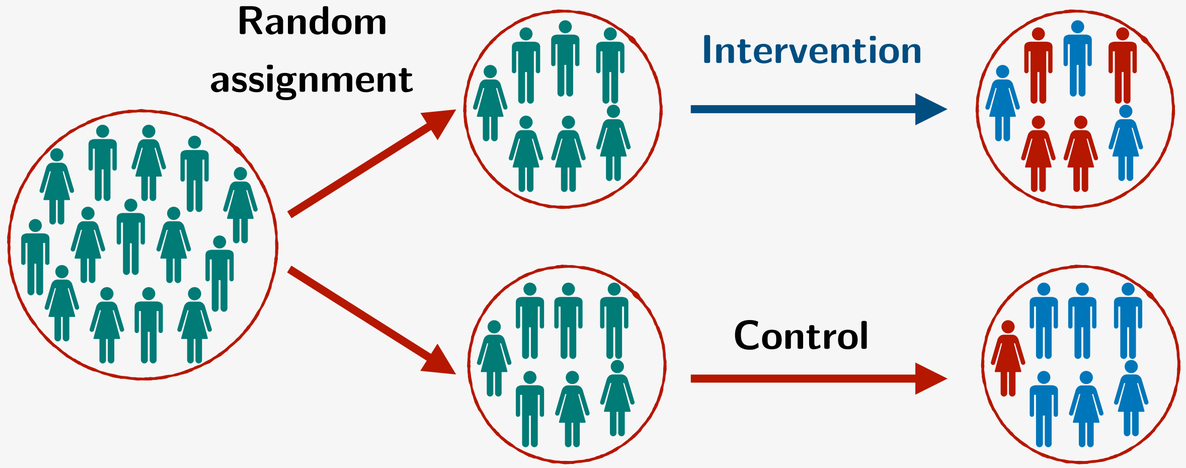


This figure is downloaded from Google Image

Mendelian randomization

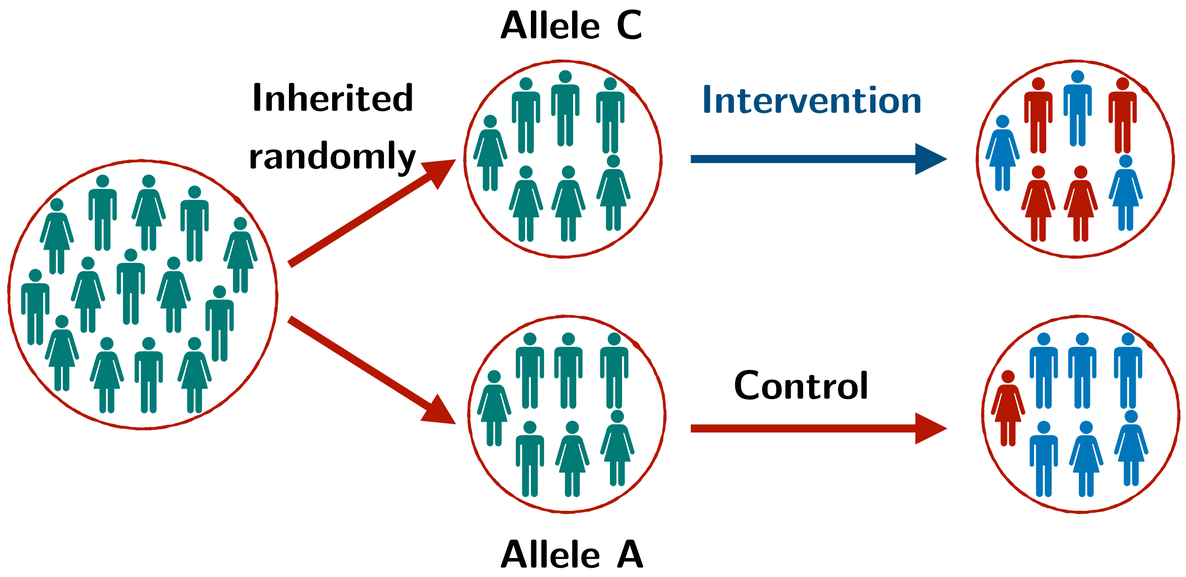
Randomized clinical trial

- Gold standard
- Randomization balances participant characteristics between the groups

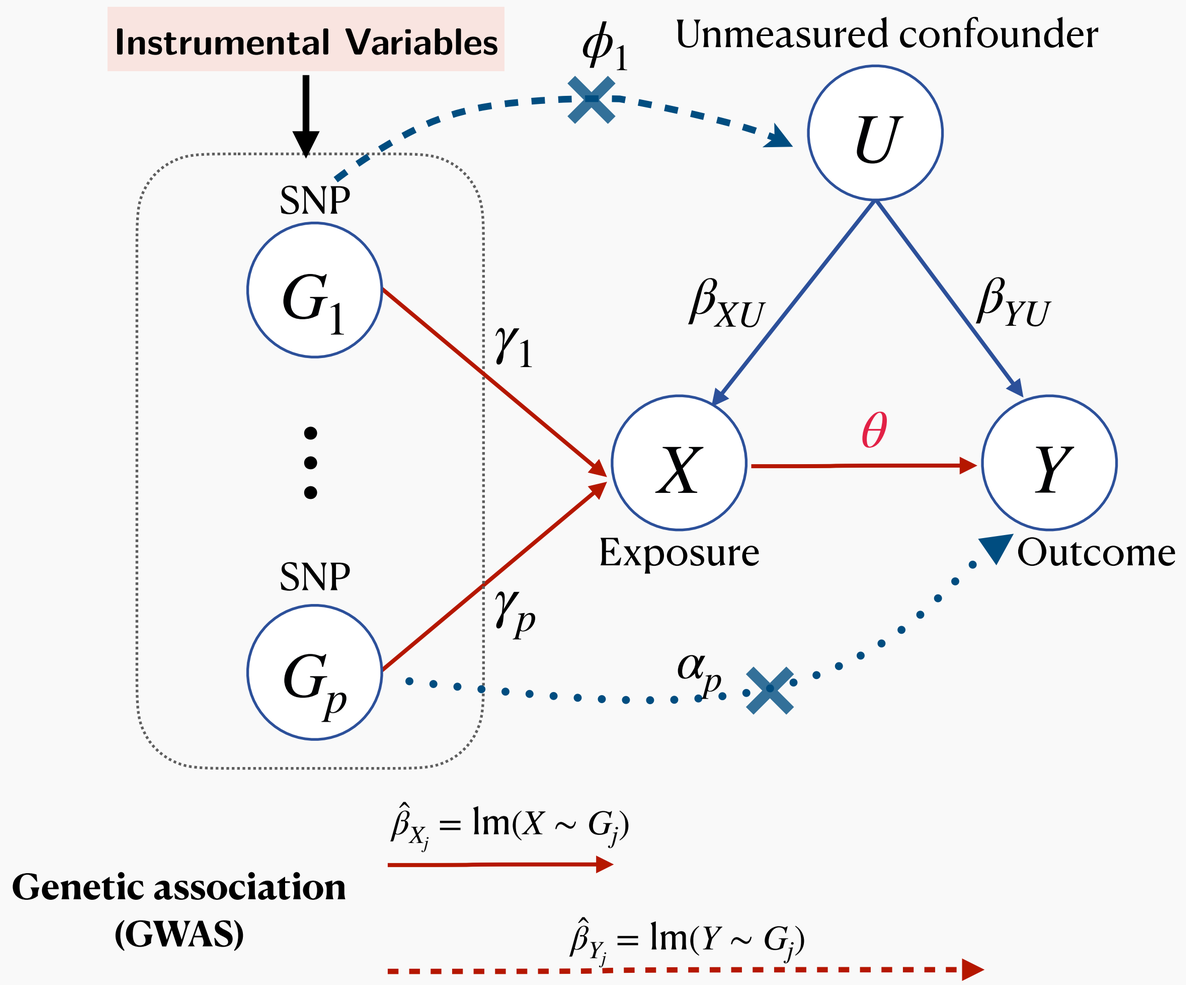


Hypothetical example

- Allele A: not smoking
- Allele C: smoking
- Not associated with unmeasured confounding factors (e.g., drinking)
- No direct effect on the outcome (e.g., lung cancer)



Mendelian randomization



Structure equation model:

$$\beta_{X_j} = \gamma_j + \phi_j \cdot \beta_{XU}$$

$$\beta_{Y_j} = \beta_{Y_j,\mathrm{M}} + \beta_{Y_j,\mathrm{D}} = \theta \cdot \beta_{X_j} + (\alpha_j + \phi_j \cdot \beta_{YU})$$

SNP j is a valid instrumental variable (IV) if

- **Relevance:** $\gamma_j \neq 0$
- **Independence:** $\phi_j = 0$
- **Exclusion restriction:** $\alpha_j = 0$

For a valid IV SNP j:

$$\beta_{X_j} = \gamma_j$$

$$\beta_{Y_j} = \theta \cdot \beta_{X_j}$$

Two-sample summary-data MR

Two-sample MR setup:

	Original data	Summary data
Exposure GWAS	$\left\{(X_i^*, G_{ij}^*)\right\}_{i=1}^{n_X}$	$\left\{(\hat{\beta}_{X_j}, \sigma_{X_j})\right\}_{j=1}^{p}$
Outcome GWAS	$\left\{(Y_i, G_{ij})\right\}_{i=1}^{n_Y}$	$\left\{(\hat{\beta}_{Y_j}, \sigma_{Y_j})\right\}_{j=1}^{p}$

Strengths of two-sample MR:

- Increase the power
- Expand the scope of MR studies

Inverse variance weighted (IVW) estimator:

- Assume all IVs are valid
- Assume no measurement error: $\hat{\beta}_{X_j} = \beta_{X_j}$
- $\hat{\beta}_{Y_j} = \theta \cdot \hat{\beta}_{X_j} + \epsilon_j$
- The IVW estimator:

$$\hat{\theta}_{\texttt{IVW}} = \frac{\sum_{j=1}^{p} \hat{\beta}_{X_j}\hat{\beta}_{Y_j}/\sigma_{Y_j}^2}{\sum_{j=1}^{p} \hat{\beta}_{X_j}^2/\sigma_{Y_j}^2}$$

Identify likely causal gene expression

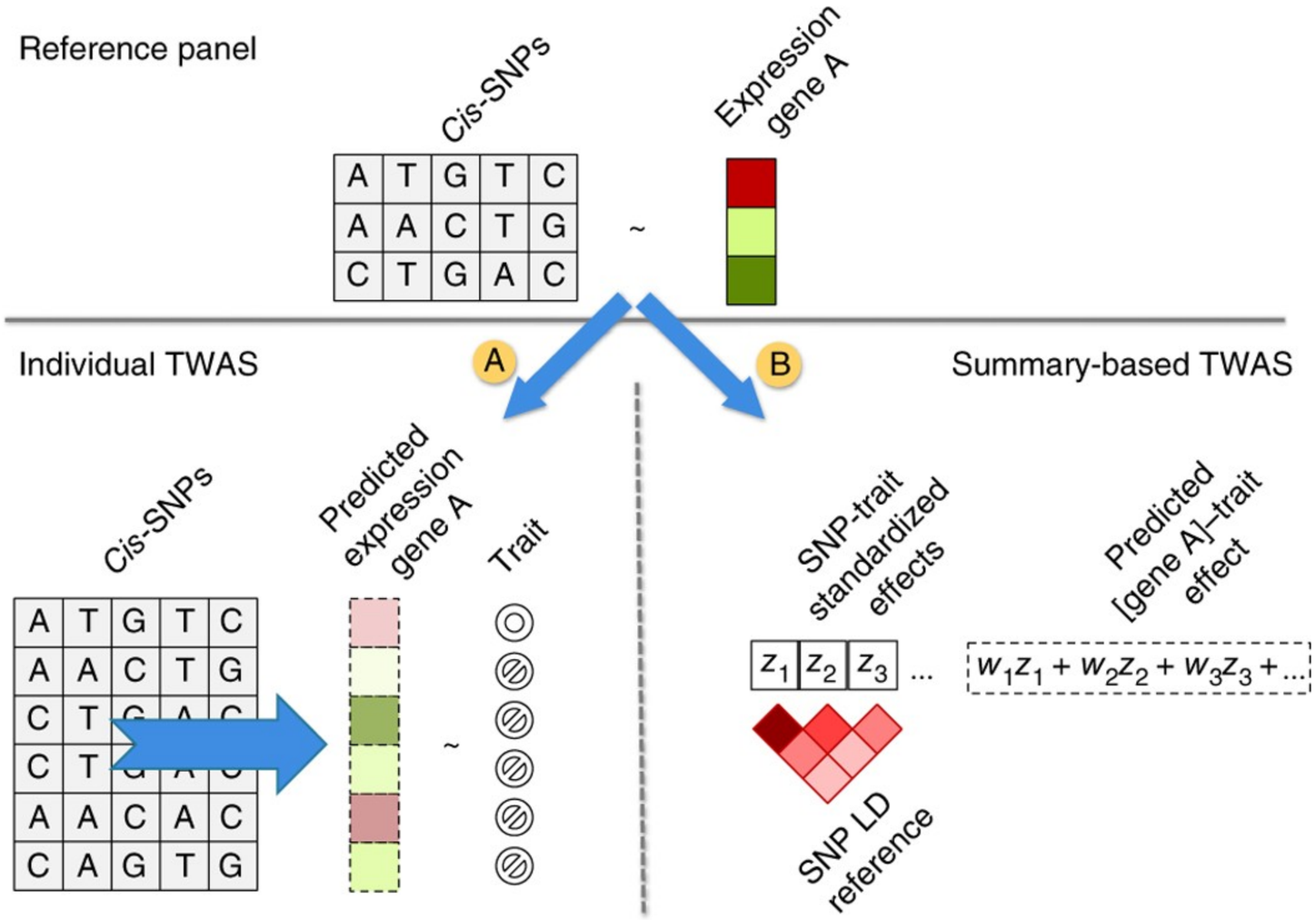
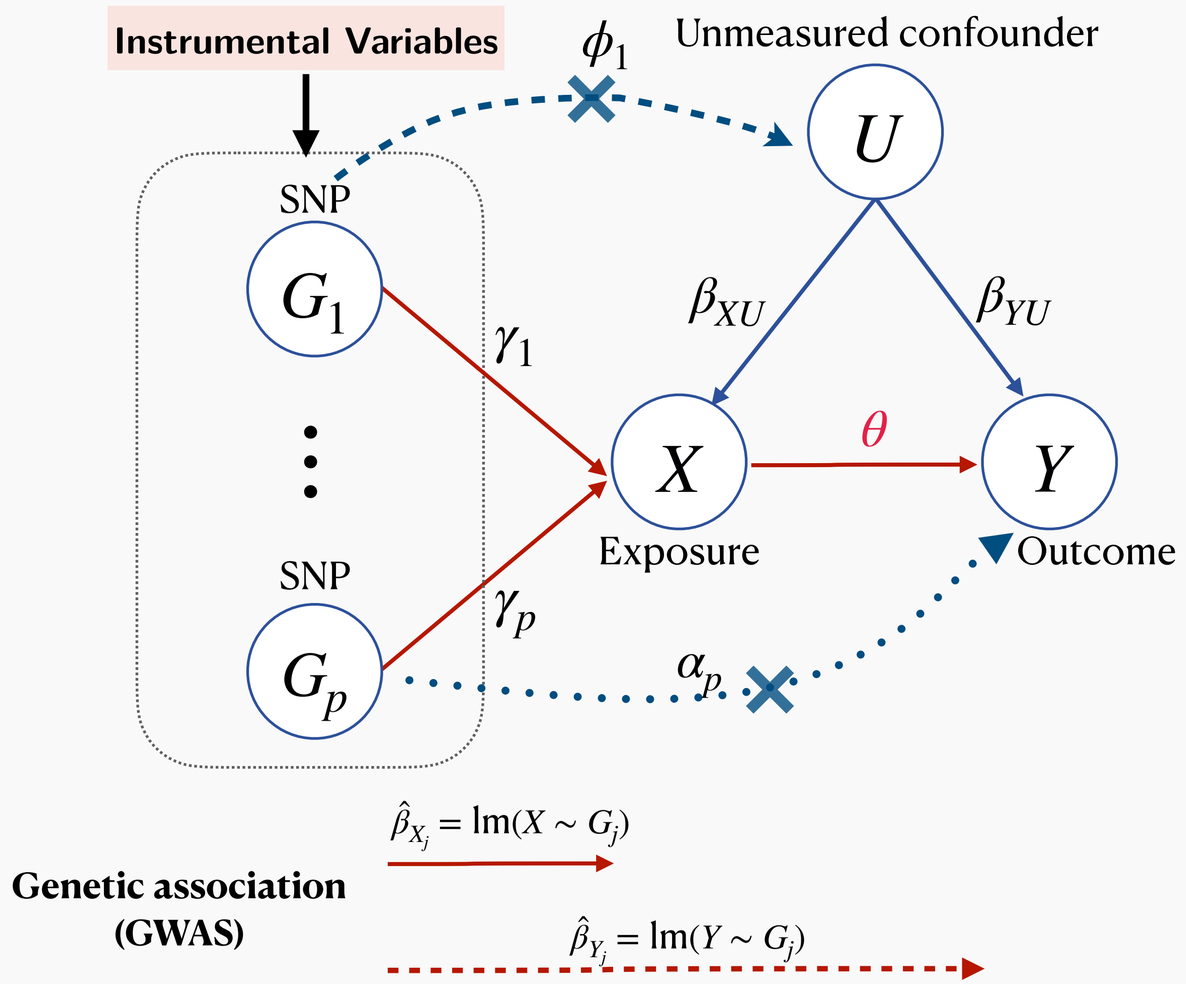


Figure: Workflow of TWAS[1]

1 Gusev, Alexander, et al. "Integrative approaches for large-scale transcriptome-wide association studies." *Nature Genetics* 48.3 (2016): 245-252.

Mendelian randomization



Structure equation model:

$$\beta_{X_j} = \gamma_j + \phi_j \cdot \beta_{XU}$$

$$\beta_{Y_j} = \beta_{Y_j,\mathrm{M}} + \beta_{Y_j,\mathrm{D}} = \theta \cdot \beta_{X_j} + (\alpha_j + \phi_j \cdot \beta_{YU})$$

SNP j is a valid instrumental variable (IV) if

- **Relevance:** $\gamma_j \neq 0$
- **Independence:** $\phi_j = 0$
- **Exclusion restriction:** $\alpha_j = 0$

For a valid IV SNP j:

$$\beta_{X_j} = \gamma_j$$

$$\beta_{Y_j} = \theta \cdot \beta_{X_j}$$

Motivation

- The size of the expression reference panels primarily determines the number of analyzable genes, and hence the power of TWASs
- The average number of expression models increased from **4,570 (v6p)** to **7,213 (v8)** for one popular TWAS method PrediXcan when the average sample size increased from 160 (v6p) to 332 (v8)
- The existing methods are based on individual-level expression reference panel with limited sample size;
- eQTLGen consortium has conducted the largest meta-analysis involving 31,684 blood samples from 37 cohorts

Q: How can we build expression prediction models using summary-level expression reference panel with large sample size?

Outline

- Background
- **New method: SUMMIT**
- Results
- Extension

SUMMIT: Overview

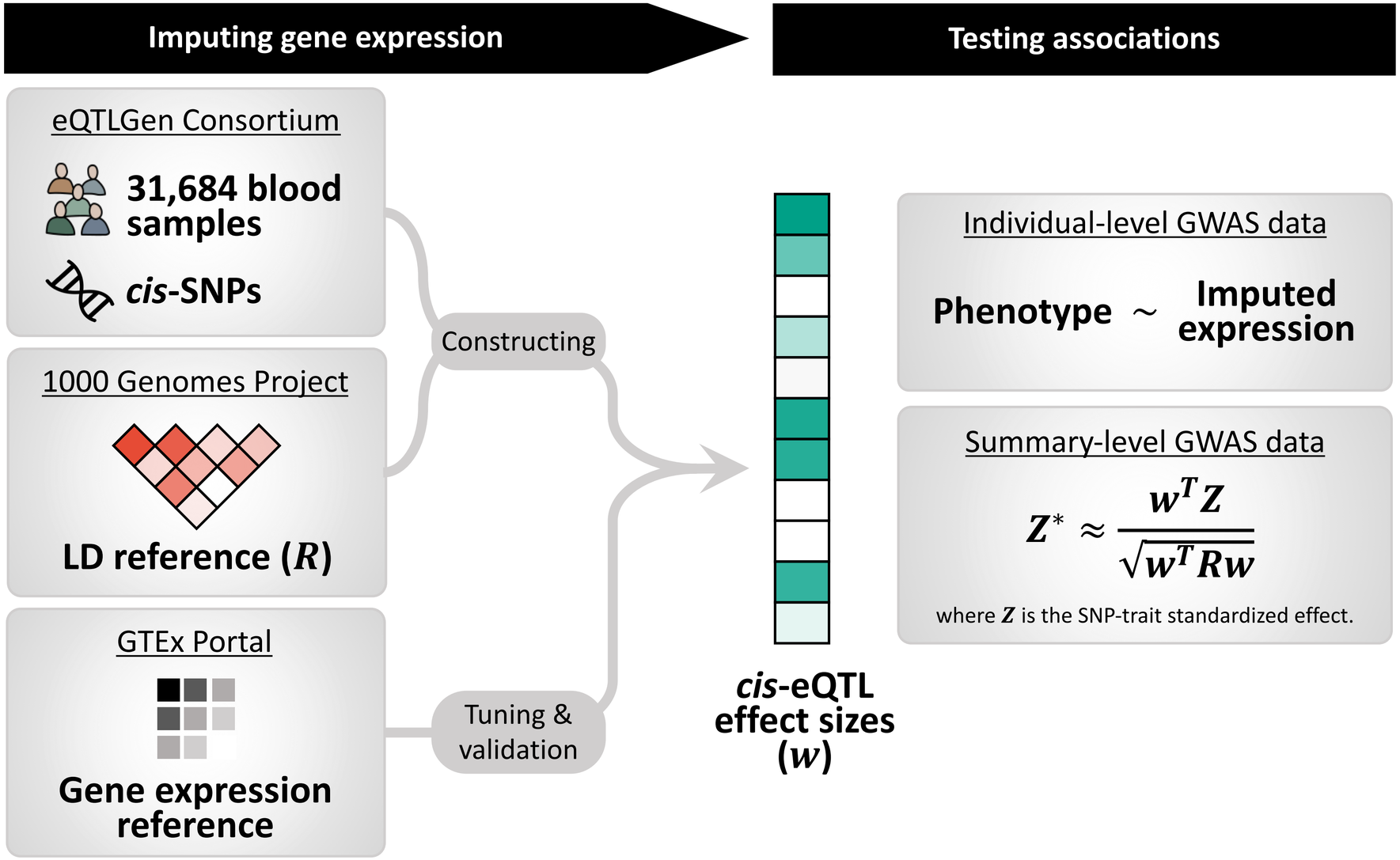


Figure: Workflow of SUMMIT

SUMMIT

Notation and model setup

$$\mathbf{Y} = \sum_{j=1}^{p} w_j \mathbf{X}_j + \epsilon$$

- $\mathbf{Y}$ is the gene expression levels; $\mathbf{X} = (\mathrm{X}_1^{'}, \cdots, \mathrm{X}_p^{'})'$ is the $N \times p$ standardized genotype matrix of p cis-SNPs around the gene; $\mathbf{w} = (w_1, \cdots, w_p)'$ is the cis-eQTL effect size, which can be estimated by

$$f(\mathbf{w}) = \frac{(\mathbf{Y} - \mathbf{X}\mathbf{w})'(\mathbf{Y} - \mathbf{X}\mathbf{w})}{N} + J_\lambda(\mathbf{w}) = \frac{\mathbf{Y}'\mathbf{Y}}{N} + \mathbf{w}'\left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)\mathbf{w} - 2\mathbf{w}'\frac{\mathbf{X}'\mathbf{Y}}{N} + J_\lambda(\mathbf{w})$$

SUMMIT

Notation and model setup

$$f(\mathbf{w}) = \frac{\mathbf{Y}'\mathbf{Y}}{N} + \mathbf{w}'\mathbf{R}\mathbf{w} - 2\mathbf{w}'\mathbf{r} + J_\lambda(\mathbf{w}),$$

Not depend on w

- $J_\lambda(\cdot)$ is a penalty term; such as LASSO, elastic net, MCP, SCAD, and MNet
- $\mathbf{r} = \mathbf{X}'\mathbf{Y}/N = (r_1, \cdots, r_p)'$ is p-dimensional vector of standardized marginal effect size for cis-SNPs (i.e., correlation between cis-SNPs and gene expression levels)
- $\mathbf{R} = \mathbf{X}'\mathbf{X}/N$ is the linkage disequilibrium (covariance) matrix of the cis-SNPs.
- The objective function is

$$\tilde{f}(\mathbf{w}) = \mathbf{w}'\tilde{\mathbf{R}}\mathbf{w} - 2\mathbf{w}'\tilde{\mathbf{r}} + \theta\mathbf{w}'\mathbf{w} + J_\lambda(\mathrm{w})$$

Ensure a unique solution upon optimization

SUMMIT

Estimating the standardized marginal effect size $\tilde{r}$:

$$\tilde{r}_j = Z_j / \sqrt{N_j - 1 + Z_j^2},$$

- where Z_j and N_j are the z-score and sample size for cis-SNP j, respectively.
- Z_j and N_j are provided by eQTL summary-level data (such as eQTLGen; **publicly available**)

Estimating the LD matrix $\tilde{R}$:

We can estimate LD matrix $\tilde{R}$ from a reference panel (such as 1000 Genomes Project data; **publicly available**)

High dimensionality problem:

- Instead of using sample correlation matrix, we use the shrinkage estimator of the LD matrix
- Stabilize results by shrinking the off-diagonal entries toward zero (the magnitude depends on the genetic distance)

SUMMIT

When individual-level GWAS data (genotype data $\mathsf{X}_{\mathsf{new}}$, phenotype $\mathsf{P}_{\mathsf{new}}$, and covariance matrix $\mathsf{C}_{\mathsf{new}}$) are available

- one can apply a generalized linear regression model to test $H_0 : \beta = 0$

$$f(E[\mathsf{P}_{\mathsf{new}} \,|\, \mathsf{X}_{\mathsf{new}}, \mathsf{C}_{\mathsf{new}}]) = \alpha \mathsf{C}_{\mathsf{new}} + \beta \mathsf{X}_{\mathsf{new}} \hat{\mathsf{w}},$$

- where $\mathsf{X}_{\mathsf{new}} \hat{\mathsf{w}}$ is the predicted genetically regulated expression for the trait of interest.

When only summary-level GWAS data are available

- one can apply a burden-type test:

$$\tilde{Z} = \mathsf{Z} \hat{\mathsf{w}} / \sqrt{\hat{\mathsf{w}}' \mathsf{V} \hat{\mathsf{w}}},$$

- where Z is the vector of z-scores for all cis-SNPs and V is the LD matrix of analyzed SNPs

SUMMIT

Cauchy combination test to integrate information from K models

$$T = \sum_{j=1}^{K} \tilde{R}_j^2 \tan\{(0.5 - p_j)\pi\},$$

- where p_j the p-value for model j and $\tilde{R}_j^2$ is calculated by $R_j^2 / \sum_{j=1}^{k} R_j^2$.
- T approximately follows a standard Cauchy distribution, and the p-value can be calculated as $0.5 - \arctan(T)/\pi$.
- The Cauchy combination test has been widely used, key benefit:
  - p-value approximation is accurate for highly significant results (which are of interest),
  - no need to estimate the correlation structure among the combined p-values.

Outline

- Background
- New method: SUMMIT
- **Results**
- Extension

Methods to be compared

SUMMIT: SUMMIT with the cis-eQTL summary-level data from eQTLGene (31,684 blood samples)

Lassosum: a popular polygenic risk score method Lassosum with the eQTLGene

Single tissue method:

PrediXcan: Elastic Net with GTEx v8 samples (individual-level data; 670 blood samples)

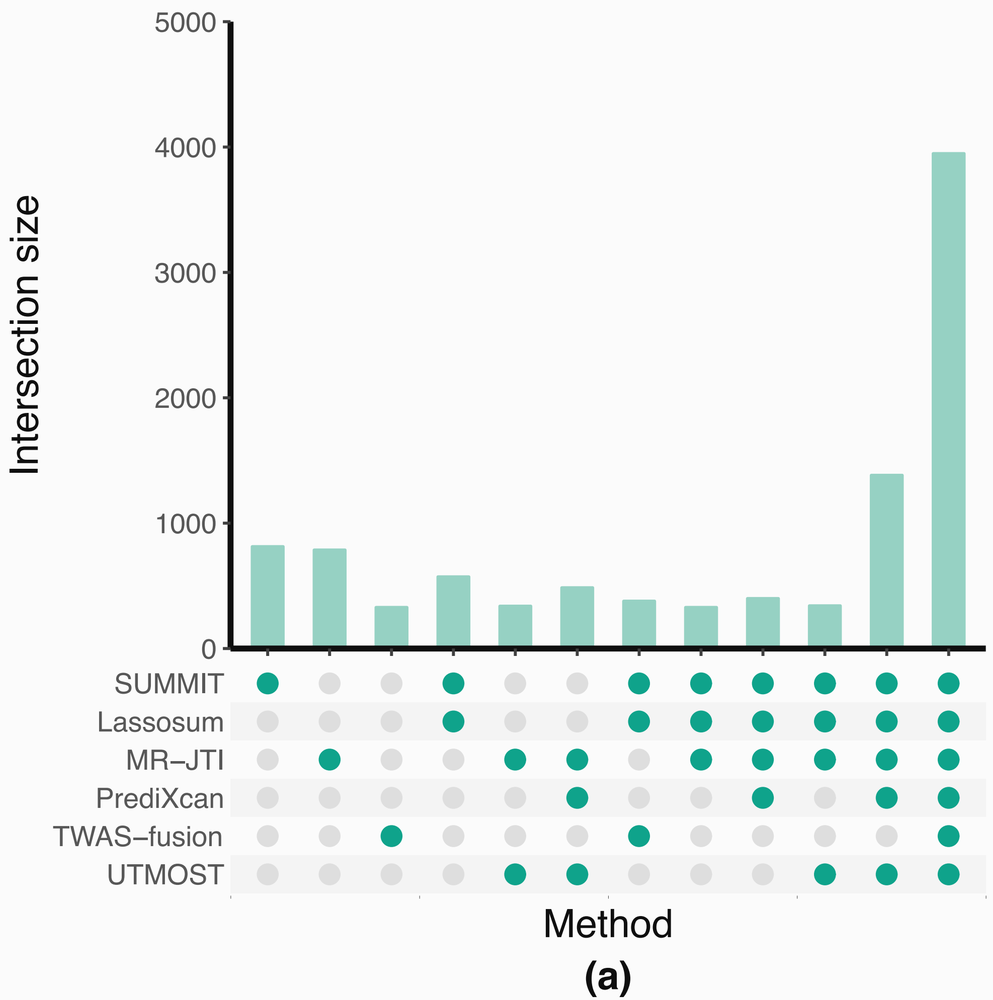
TWAS-fusion: several methods, including BLUP, BSLMM, Elastic Net, LASSO, and TOP1 with GTEx v8 samples

Cross-tissue method:

MR-JTI: GTEx v8 samples (all available tissues)

UTMOST: GTEx v8 samples (all available tissues)

SUMMIT improves the expression imputation accuracy

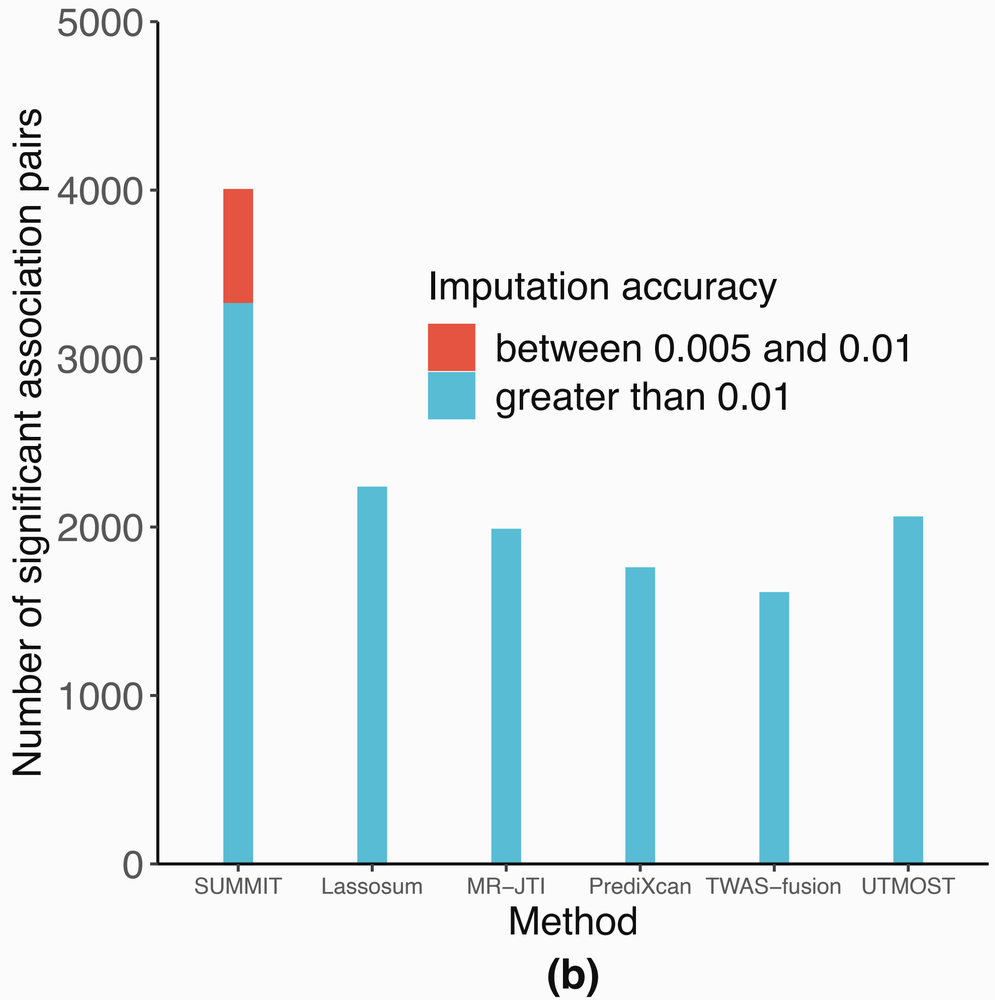


Number of genes with $R^2 \geq 0.01$:

- SUMMIT: 9,749
- Bechmark methods: Lassosum: 8,249; MR-JTI: 9,576; TWAS-Fusion: 5,411; PrediXcan: 7,512; UTMOST: 7,236

SUMMIT achieved higher prediction accuracy in different quantiles compared with all benchmark methods (by Kolmogorov-Smirnov test)

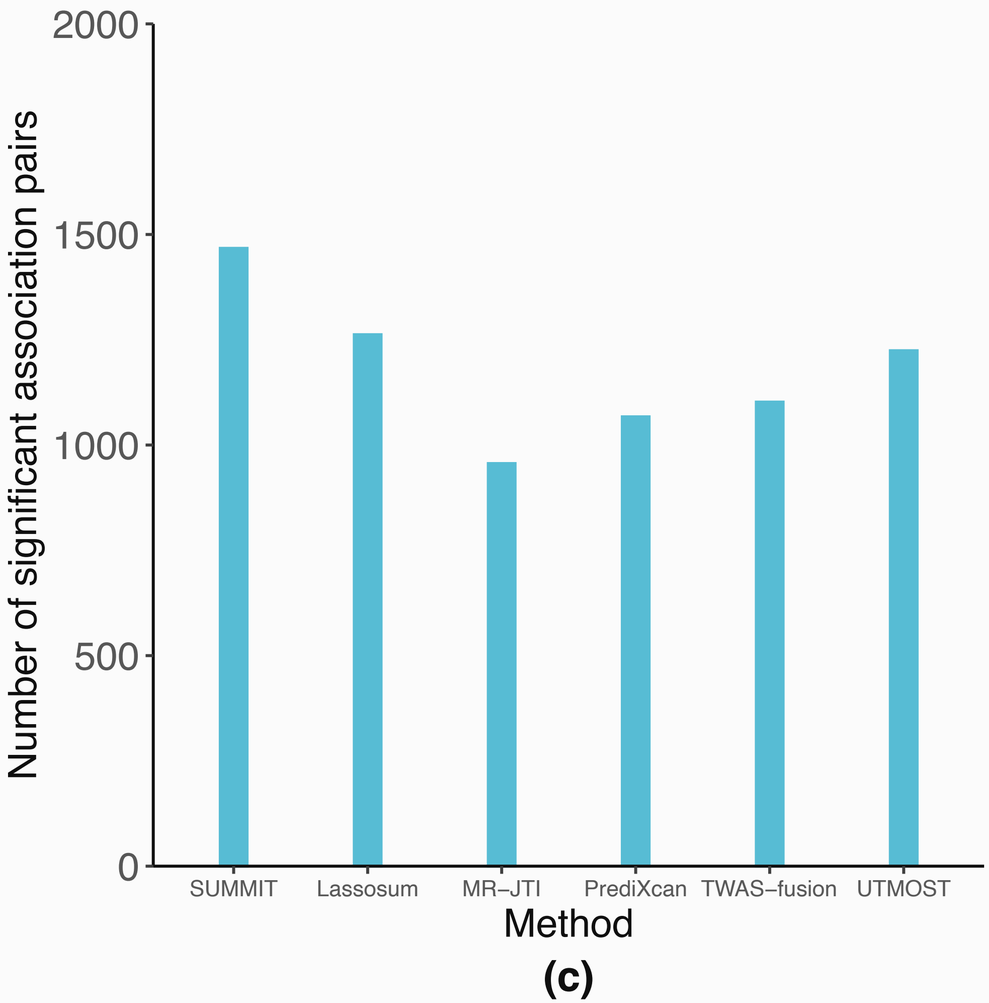
SUMMIT identifies more associations than competing methods



(b)

- Using Bonferroni correction for all methods
- Based on GWAS summary statistics of 24 traits ($N_{\text{total}} \approx 5{,}600{,}000$ without adjusting for sample overlap across studies
- When focused on genes with $R^2 \geq 0.01$; SUMMIT achieved better results (the differences are significant by the paired Wilcoxon rank test)
- SUMMIT can analyze genes with low heritability, which often have large causal effect sizes on the trait

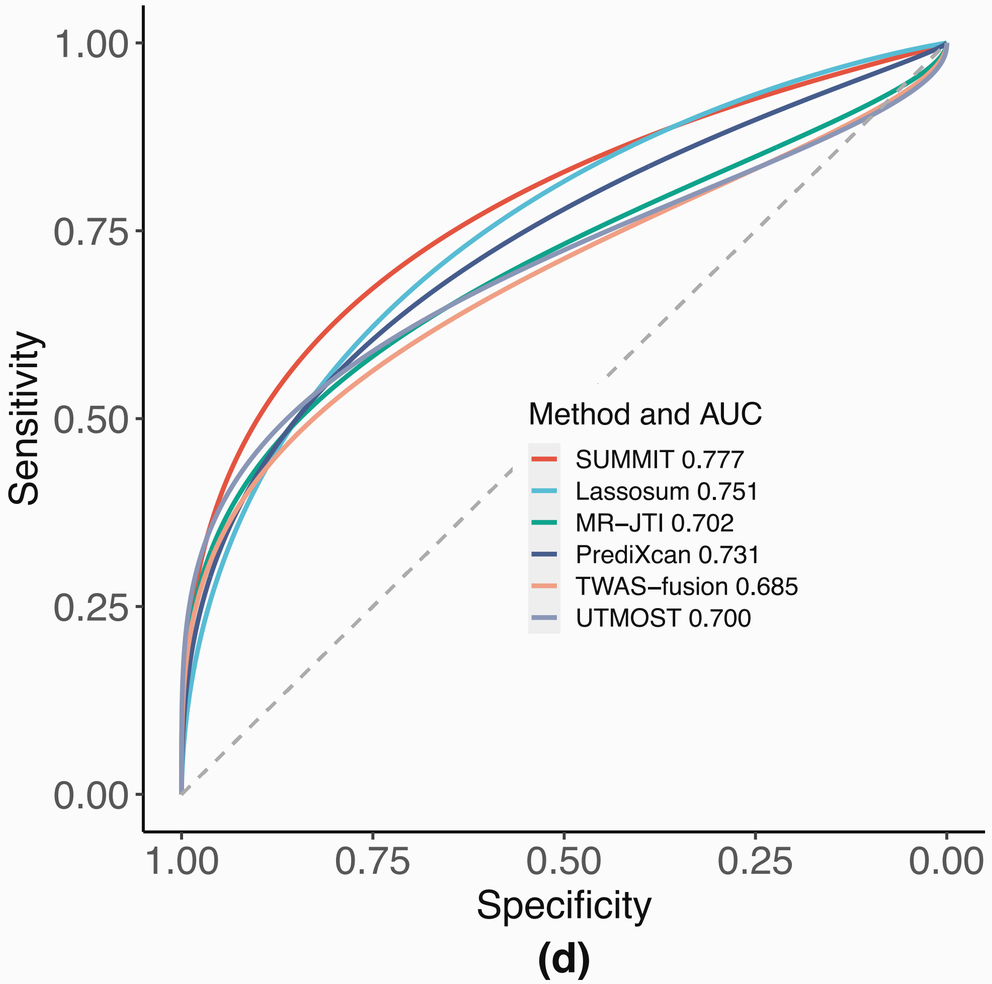
SUMMIT identifies more associations than competing methods



(c)

- When focused on genes that can be analyzed by all the methods; SUMMIT still achieved better results

SUMMIT achieves higher predictive power for identifying “silver standard” genes



(d)

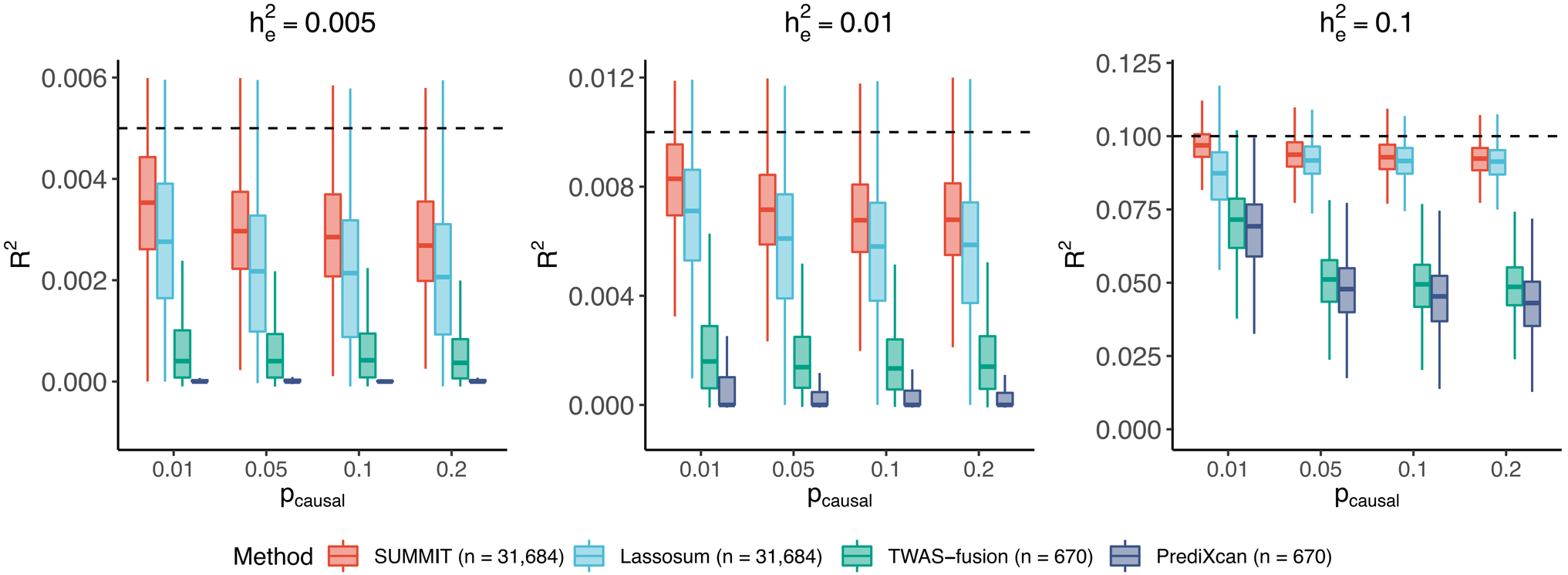
- Following Barbeira et al., we used a set of 1,258 likely causal gene-trait pairs curated by using the Online Mendelian Inheritance in Man (OMIM) database and a set of 29 gene-trait pairs based on rare variant results from exome-wide association studies
- Provide orthogonal information that is independent of the GWAS results
- All methods performed relatively good; SUMMIT achieved the highest AUC

Simulation settings

Using UK Biobank

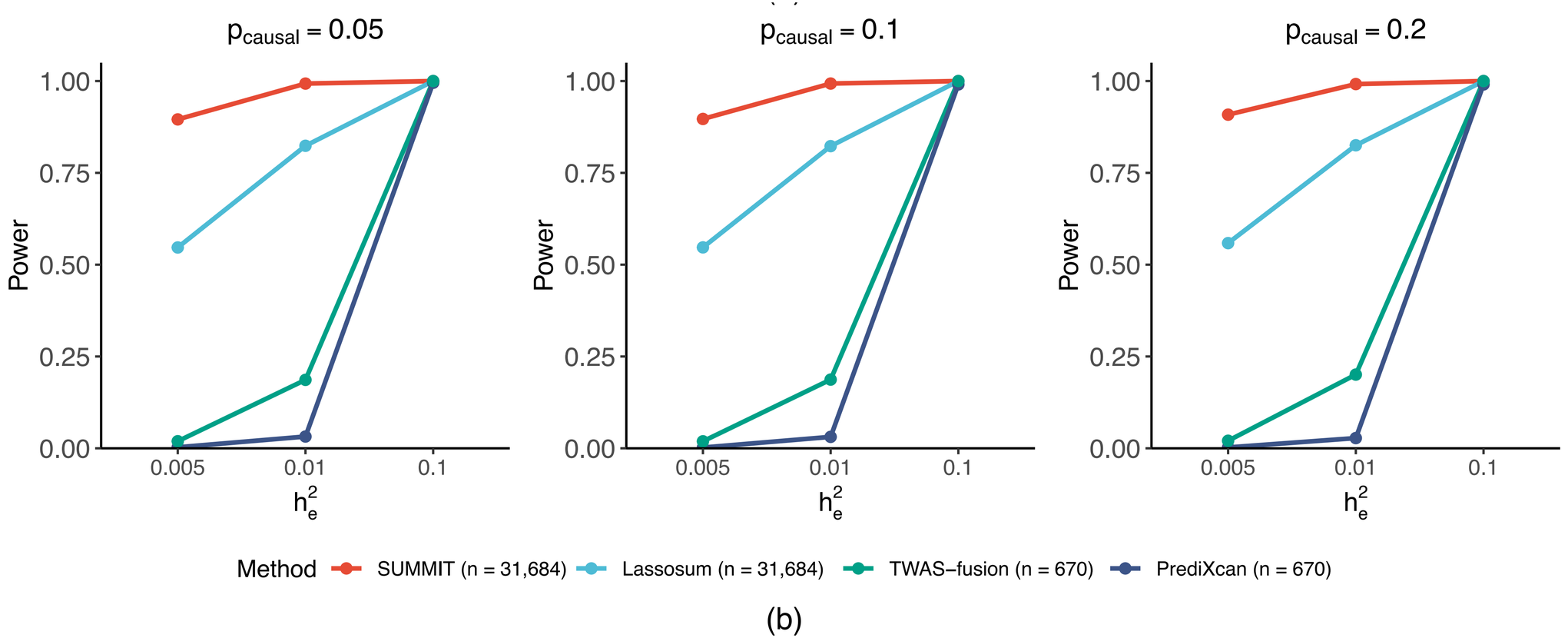
- Randomly selected genotype data from unrelated white British individuals as training data (to match with the sample size of real data analyses)
- 10,000 unrelated white British individuals as test data
- $\mathrm{E}_g = \mathrm{Xw} + \epsilon_e$
- $\mathrm{Y} = \beta \mathrm{E}_g + \epsilon_p$
- $\epsilon_e \sim N(0, 1 - h_e^2)$, and $\epsilon_p \sim N(0, 1 - h_p^2)$
- h_e^2 : expression heritability (i.e., the proportion of gene expression variance explained by SNPs)
- h_p^2: phenotypic heritability (i.e., the proportion of phenotypic variance explained by gene expression levels)

Simulation results



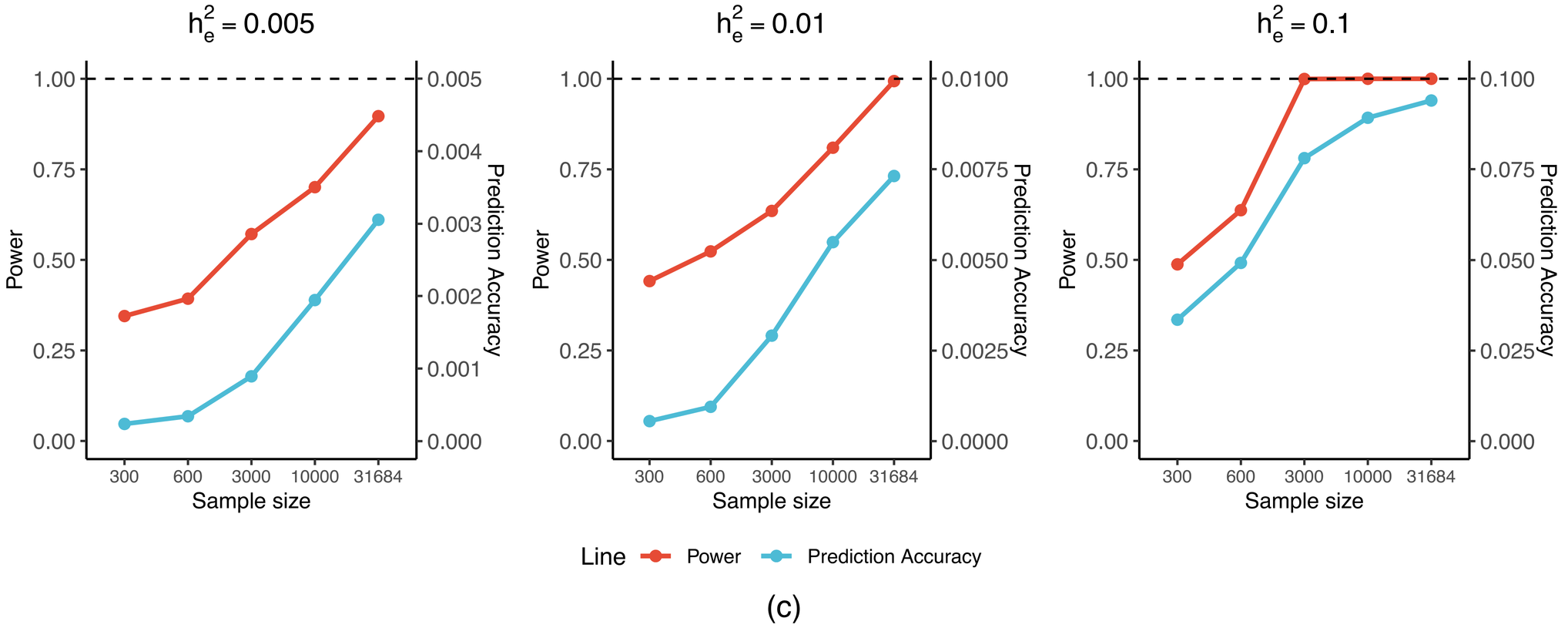
(a)

Simulation results



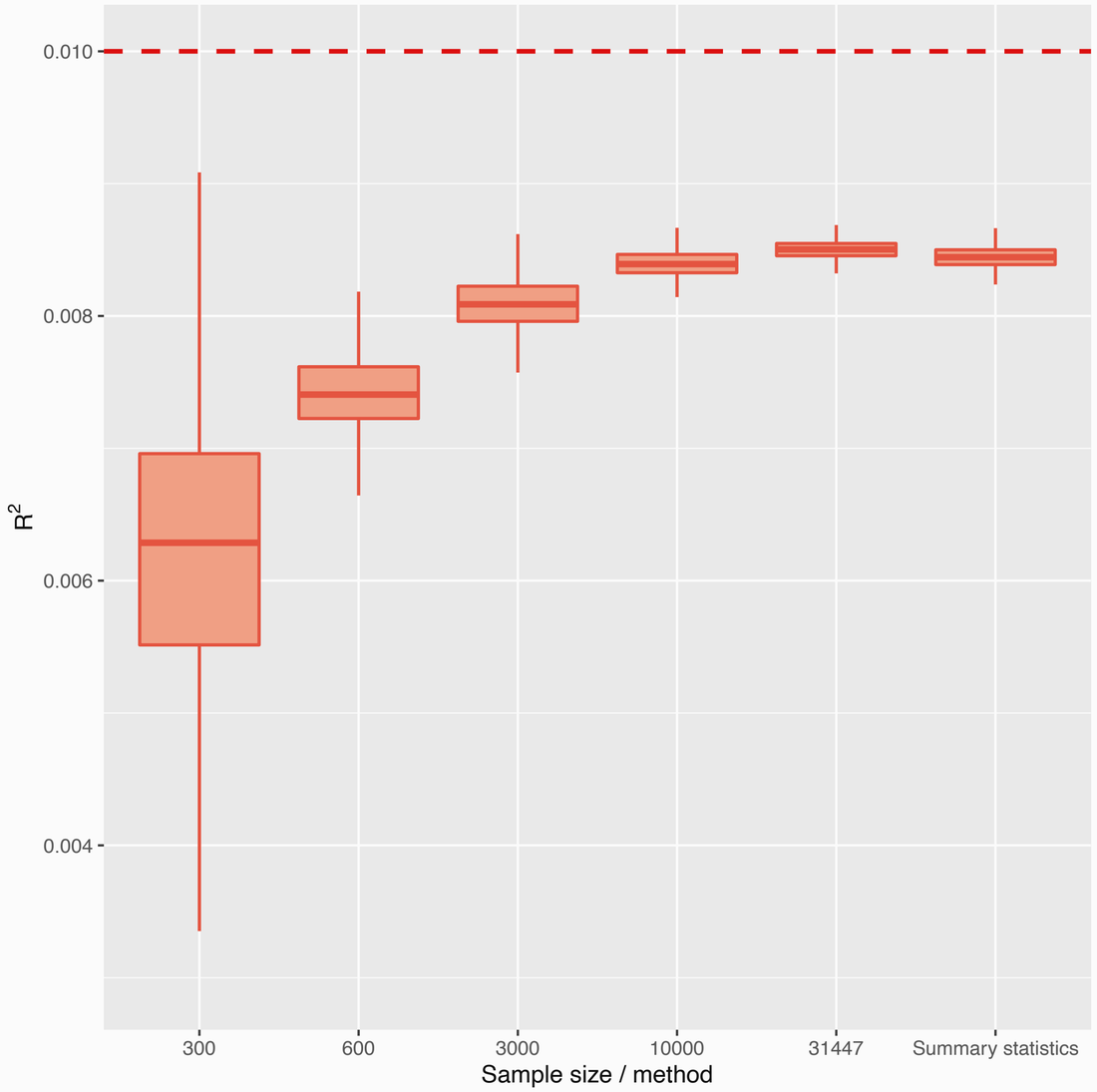
(b)

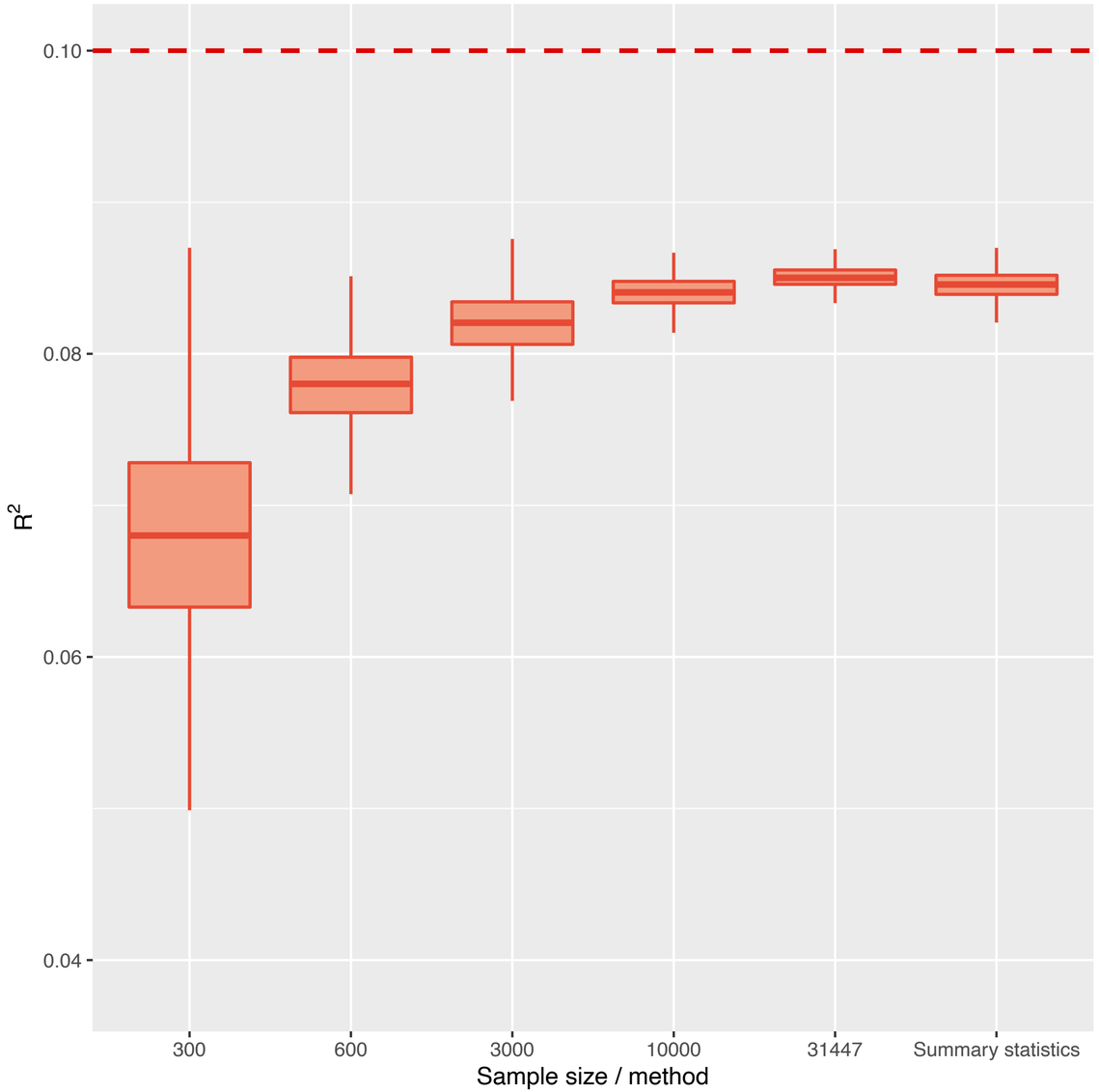
Simulation results



(c)

Simulation results





Online database

chongwulab.shinyapps.io

Mail - Wu,Chong - Outlook | SUMMIT - Online LaTeX Editor Overleaf | Integration of summary data from GWAS and eQTL studies predicts... | SUMMIT | SUMMIT: An integrative approach for better transcriptomic data imp... | screenshot on mac - Google Search

SUMMIT | Real-data results | Query | About

SUMMIT
chongwulab.shinyapps.io

Real-data results

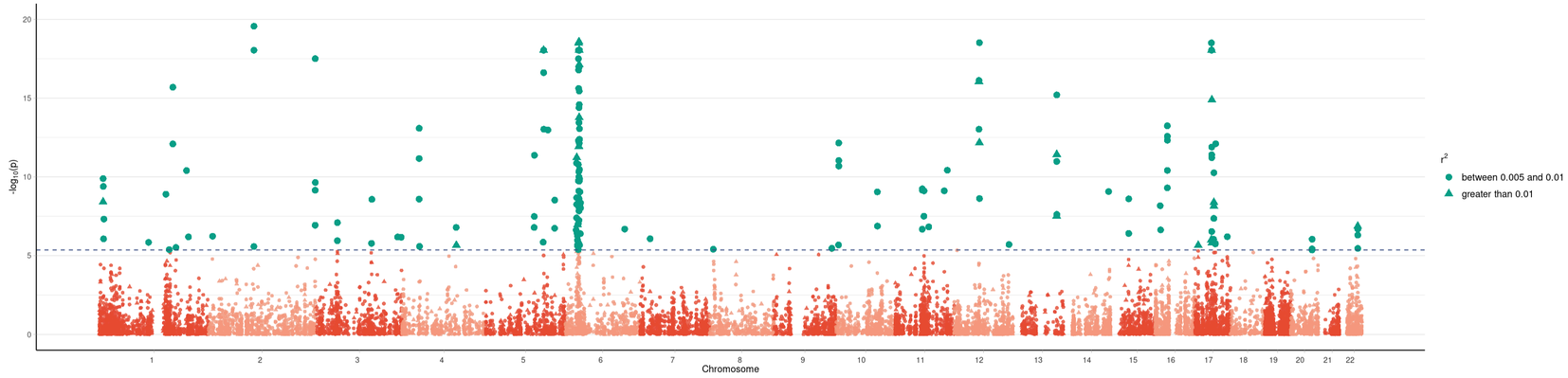
Trait/Disease: Asthma

Method: SUMMIT

Cutoff: Bonferroni-corrected cutoff

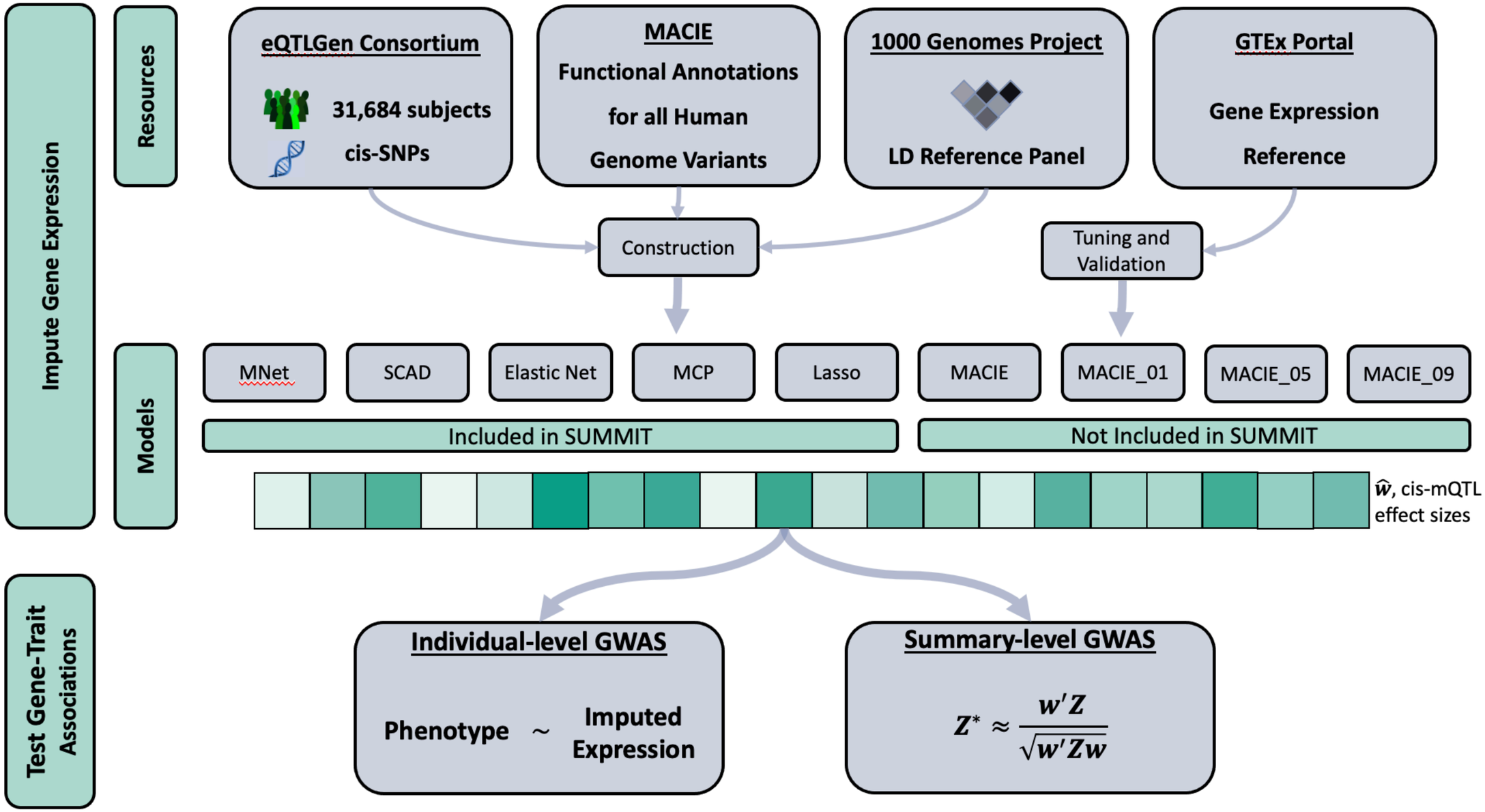
☐ Show gene annotation in Manhattan plot

Download gene list

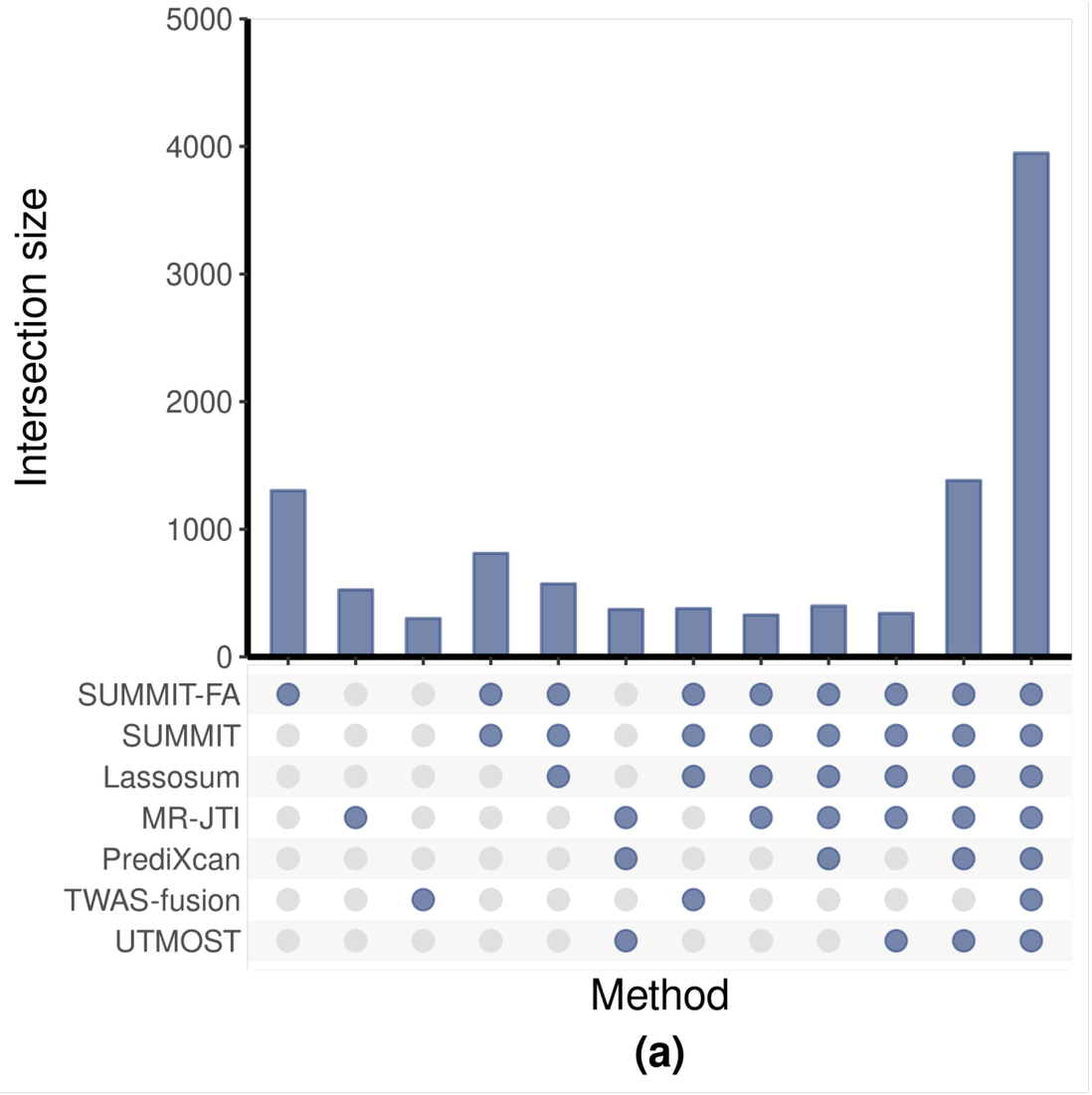


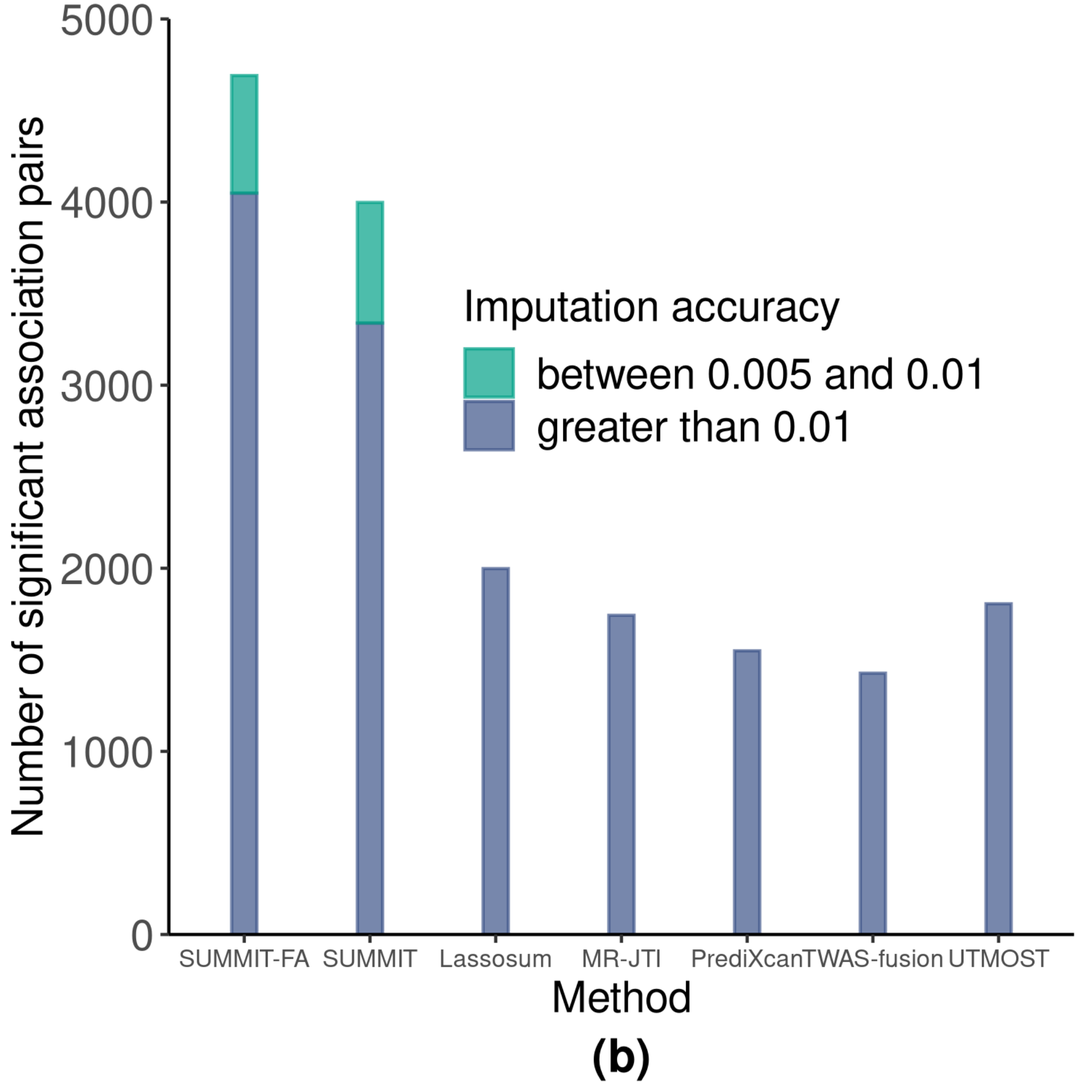
gene ID	gene name	p-value	chromosome	position	R2
ENSG00000072682	P4HA2	3.297e-22	5	131579269	0.006
ENSG00000073605	GSDMB	1.817e-57	17	38068477	0.395
ENSG00000172057	ORMDL3	1.250e-56	17	38080574	0.298
ENSG00000172346	CSDC2	1.342e-07	22	41965256	0.009
ENSG00000172568	FNDC9	1.816e-07	5	156770668	0.018
ENSG00000172638	EFEMP2	7.663e-10	11	65637487	0.072
ENSG00000172992	DCAKD	4.305e-08	17	43119590	0.253
ENSG00000174123	TLR10	2.571e-09	4	38779235	0.018
ENSG00000174130	TLR6	6.788e-12	4	38841887	0.025
ENSG00000176973	FAM89B	3.123e-08	11	65340744	0.049
ENSG00000179344	HLA-DQB1	1.893e-134	6	32631702	0.708
ENSG00000179428	AC073072.5	8.403e-07	7	22766126	0.015
ENSG00000179639	FCER1A	4.115e-06	1	159268759	0.032
ENSG00000074800	ENO1	4.035e-10	1	8930184	0.014
ENSG00000180902	D2HGDH	6.968e-10	2	242691112	0.053
ENSG00000181004	BBS12	2.188e-06	4	123659977	0.008
ENSG00000182134	TDRKH	1.253e-09	1	151753237	0.075
ENSG00000182742	HOXB4	1.780e-06	17	46655174	0.032
ENSG00000186470	BTN3A2	2.087e-09	6	26371966	0.632

Extension: SUMMIT-FA

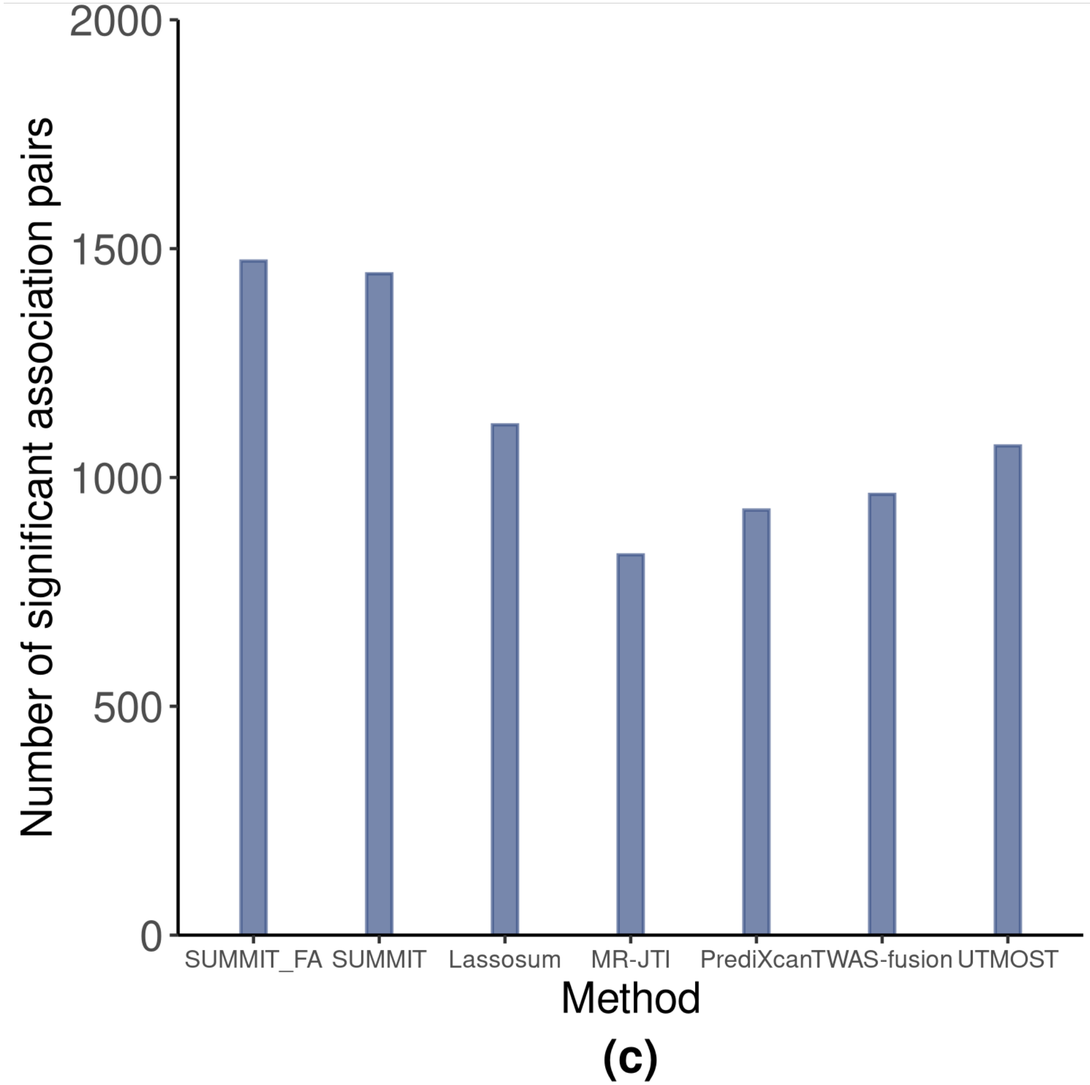


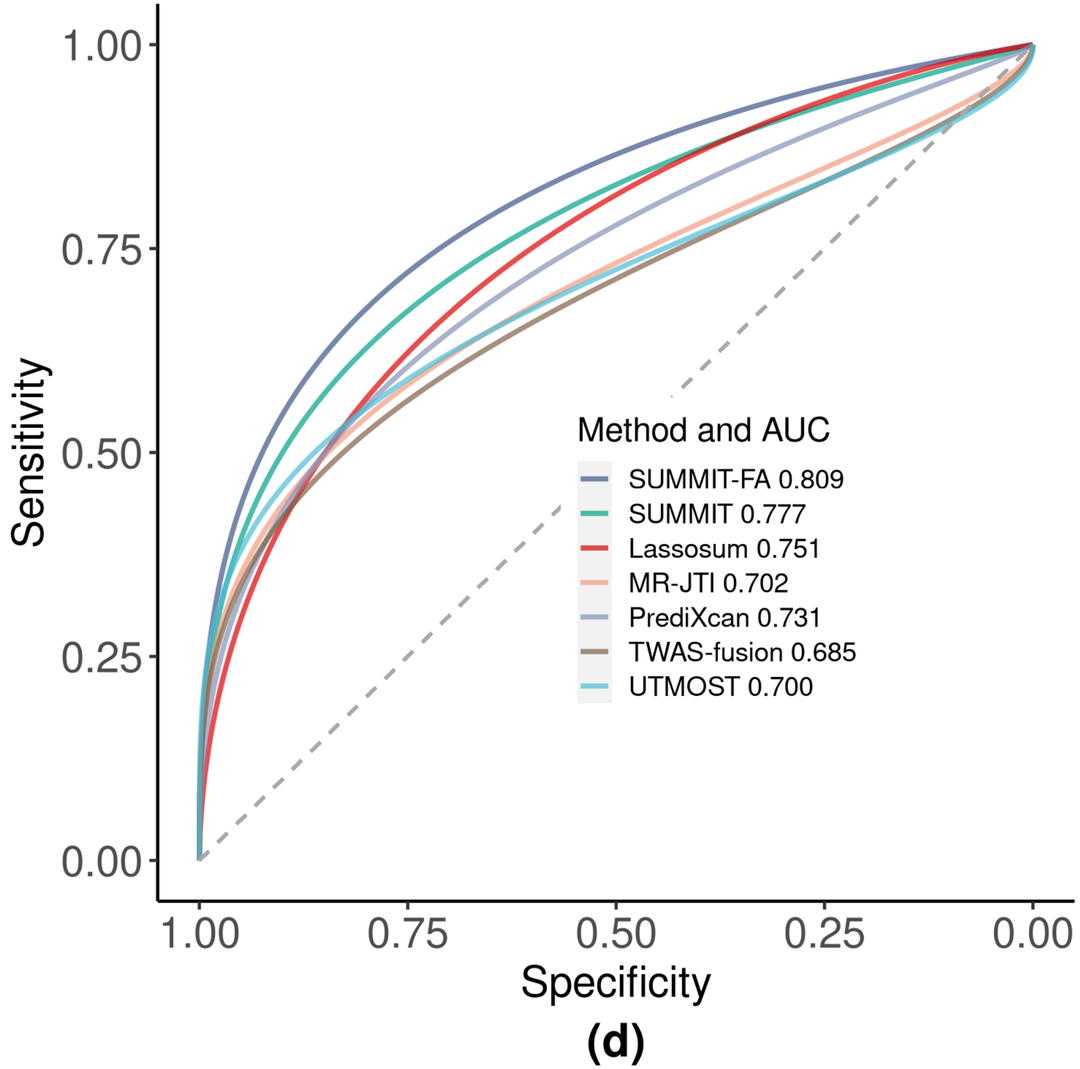
Extension: SUMMIT-FA





Extension: SUMMIT-FA





Summary

- By leveraging eQTL data with large sample-size, SUMMIT improves the accuracy of expression prediction in blood, successfully builds expression prediction models for genes with low expression heritability, and outperforms benchmark methods for identifying risk genes
- TWAS methods, including SUMMIT, can be viewed as one type of gene-based Mendelian randomization (MR) and can provide valid causal interpretations only when all genetic variants used in the expression prediction models are valid instrumental variables **(Strong and uncheckable assumption)**

Summary

- Besides complementary analyses (such as fine-mapping and colocalization), robust inference with weak assumptions are needed
- SUMMIT can be extended to other omics data (proteins, DNA methylation, and metabolites)
- Multi-ethnicity: Improve the robustness and performance (transfer learning)
- Multi-ethnicity: Identify ethnicity-specific and pan-ethnicity likely causal biomarkers

Acknowledgements

- Zichen Zhang @ Florida State University
- Ye Eun Bae @ Florida State University
- Jon Bradley @ Florida State Statistics
- Lang Wu @ University of Hawaii Cancer Center

Zhang, Zichen, Ye Eun Bae, Jonathan R. Bradley, Lang Wu, and Chong Wu. "SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification." *Nature Communications* **13**, 6336 (2022).

NIH

biobank uk

Thank you!

Chong Wu

Email: cwu18@mdanderson.org

Website: https://wuchong.org