A powerful fine-mapping method for transcriptome-wide association studies

Chong Wu
Department of Statistics
Florida State University

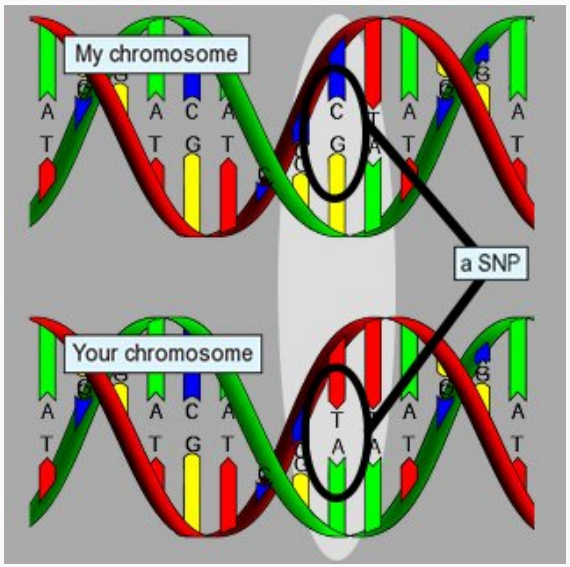
Chong Wu & Wei Pan
JSM 2020

Aug. 5, 2020

Outline

- Background
- Methods
- Results
- Discussion

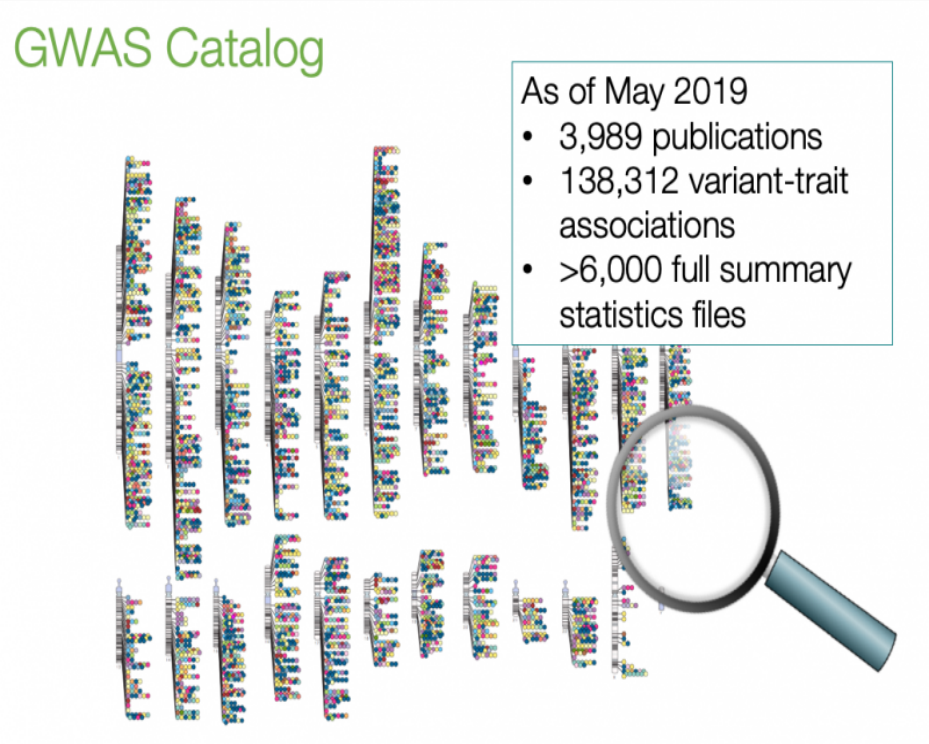
Genome-wide association study (GWAS)



copyright @ John Fouts (2016)

- Genome: the set of genetic information encoded in 23 chromosome pairs
- SNP: Variation in a single base pair
- Genetic score (additive) for each SNP and a person:
  AA = 0, AB = 1, BB = 2

GWAS



Question

How do we understand GWAS associations?

Transcriptome-Wide Association Study (TWAS)

Goal:

Estimate the association between gene expression and disease

- SNP $\rightarrow$ Gene expression $\rightarrow$ disease
- The sample size of gene expression data is usually small

Transcriptome-Wide Association Study (TWAS)

- We have two separate datasets: 1) transcriptome data (with SNP and gene expression); 2) GWAS data (with SNP and disease status)
- TWAS idea:
  - predict/impute gene-expression with SNPs as predictors;
  - test association b/w a trait and imputed gene-expression;

TWAS/PrediXcan idea cont.

- Build a prediction model for genetically regulated expression (GRex): $Y^* = \sum_{j=1}^{p} w_j X_j^* + \epsilon$, where Y^* is gene-expression.
- for a given gene for subject i, predict the GReX of the gene using the SNPs around that gene: $\widehat{\text{GReX}}_i = \sum_{j=1}^{p} \hat{w}_j X_{i,j}$;
- test association between a trait and predicted gene-expression: $g(E(Y_i)) = \beta_0 + \widehat{\text{GReX}}_i \beta_c = \beta_0 + \sum_{j=1}^{p} \hat{w}_j X_{i,j} \beta_c$ with null hypothesis H_0: $\beta_c = 0$.

More details

- Consider a GLM: $g(E(Y_i)) = \beta_0 + \beta' X_i = \beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j$ with $H_0 : \beta = (\beta_1, ..., \beta_p)' = 0$;
- replace $X_{i,j}$ by the weighted genotype scores $\hat{w}_j X_{i,j}$;
- PrediXcan = TWAS = Sum test (Pan 2009).
- $U^* = (U_1^*, ..., U_p^*)' = \sum_{i=1}^{n} X_i'(Y_i - \hat{\mu}_i^0)$;
  $U = (U_1, ..., U_p)' = WU^* = \sum_{i=1}^{n} WX_i'(Y_i - \hat{\mu}_i^0)$,
  where $W = \text{Diag}(\hat{w}_1, ..., \hat{w}_p)$

Challenges and opportunities in TWAS

PERSPECTIVE
https://doi.org/10.1038/s41588-019-0385-z
nature genetics

Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg[1], Nasa Sinnott-Armstrong[2], Nicholas Mancuso[3], Alvaro N. Barbeira[4], David A. Knowles[5,6], David Golan[2], Raili Ermel[7], Arno Ruusalepp[7,8], Thomas Quertermous[9], Ke Hao[10], Johan L. M. Björkegren[8,10,11,12*], Hae Kyung Im[4*], Bogdan Pasaniuc[3,13,14*], Manuel A. Rivas[15*] and Anshul Kundaje[1,2*]

Transcriptome-wide association studies (TWAS) integrate genome-wide association studies (GWAS) and gene expression datasets to identify gene–trait associations. In this Perspective, we explore properties of TWAS as a potential approach to prioritize causal genes at GWAS loci, by using simulations and case studies of literature-curated candidate causal genes for schizophrenia, low-density-lipoprotein cholesterol and Crohn's disease. We explore risk loci where TWAS accurately prioritizes the likely causal gene as well as loci where TWAS prioritizes multiple genes, some likely to be non-causal, owing to sharing of expression quantitative trait loci (eQTL). TWAS is especially prone to spurious prioritization with expression data from non-trait-related tissues or cell types, owing to substantial cross-cell-type variation in expression levels and eQTL strengths. Nonetheless, TWAS prioritizes candidate causal genes more accurately than simple baselines. We suggest best practices for causal-gene prioritization with TWAS and discuss future opportunities for improvement. Our results showcase the strengths and limitations of using eQTL datasets to determine causal genes at GWAS loci.

Challenges and opportunities in TWAS

Goal:

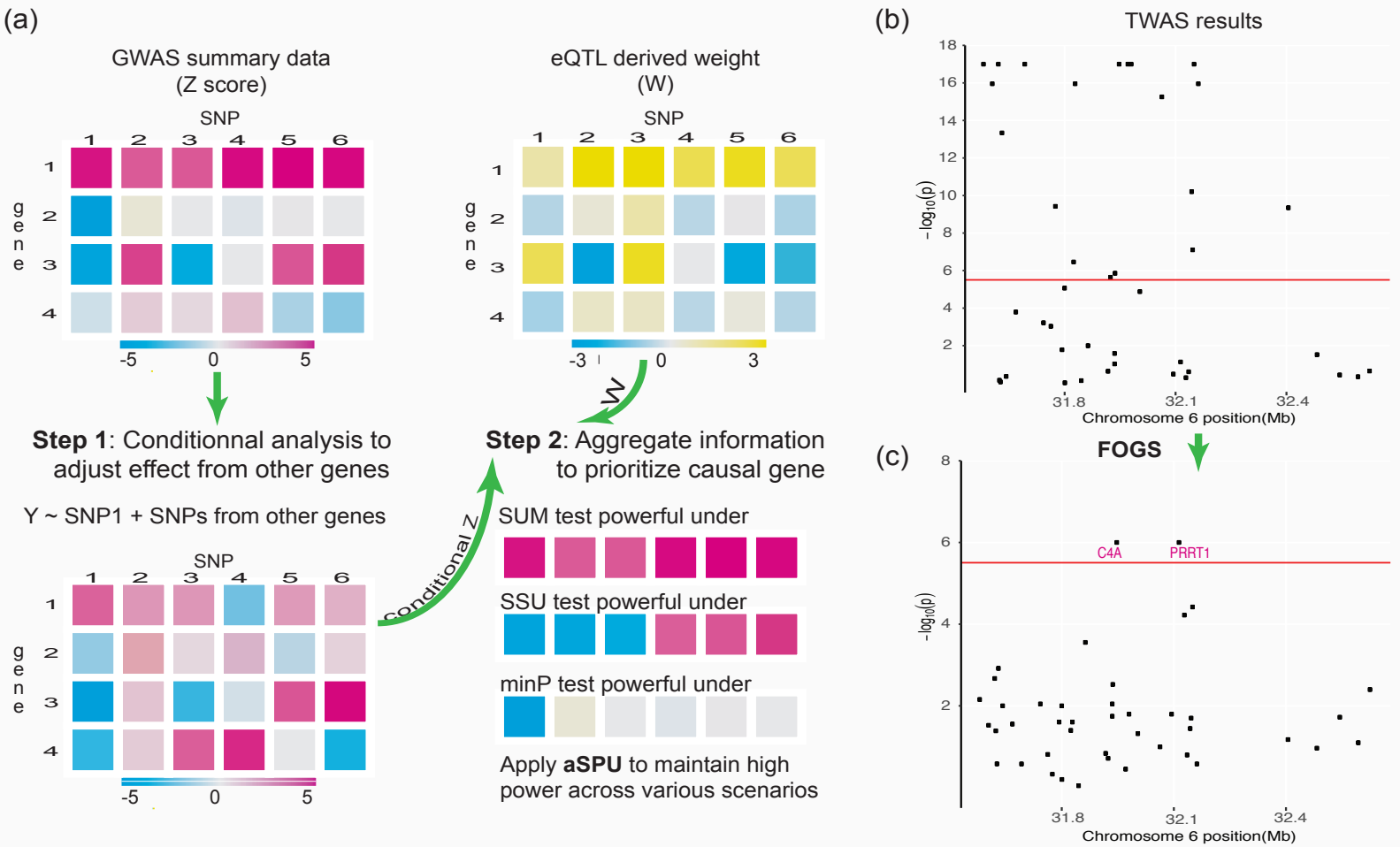
Distinguishing co-regulated genes through fine-mapping

- Fine-mapping multiple associated TWAS models at a locus;
- Fine-mapping is a method that prioritizes the most likely causal SNP/gene at a locus
- Fine-mapping is conditional analysis; Not a causal inference method

Outline

- Background
- Methods
- Results
- Discussion

Overview of Methods



Fine-mapping Of Gene Sets (FOGS)

$$y = V\alpha + X\beta + \epsilon$$

- $y = \{y_i\}$ is a centered $n \times 1$ vector of phenotypes
- $X = \{x_{ij}\}$ is a centered (with mean 0) $n \times p$ genotype matrix at p SNPs with non-zero eQTL-derived weights for gene A (of interest)
- $V = \{v_{ij}\}$ is a centered $n \times q$ genotype matrix at q SNPs with non-zero eQTL weights for any of other genes in the same locus
- β and α are the joint effects for gene A (of interest) and other genes

Step 1: Estimating the conditional Z score via ridge regression

- Q: SNPs are highly correlated; Solution: using ridge regression
- Under $H_0 : \boldsymbol{\beta} = 0$, the effect of SNPs in gene A is zero; no need to adjust for them while estimating the conditional score Z_j for SNP j
- Q: only GWAS summary data are available; Solution: using reference data to estimate the covariance matrix

$$\hat{\boldsymbol{\beta}} = (\tilde{X}'\tilde{X} + \lambda I_{p*})^{-1}\tilde{X}y$$
$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma_j^2(\tilde{X}'\tilde{X} + \lambda I_{p*})^{-1}\tilde{X}'\tilde{X}(\tilde{X}'\tilde{X} + \lambda I_{p*})^{-1}$$

Step 2: Aggregating conditional Z scores to prioritize causal gene

$$\boldsymbol{U} = (U_1, \ldots, U_p)' = \boldsymbol{WZ},$$

where $\boldsymbol{W} = \text{Diag}(\hat{w}_1, \ldots, \hat{w}_p)$ are the eQTL-derived weights and $\boldsymbol{Z}$ is the conditional Z score estimated from the previous subsection.

Challenge:

Different tests will be powerful under different alternatives.

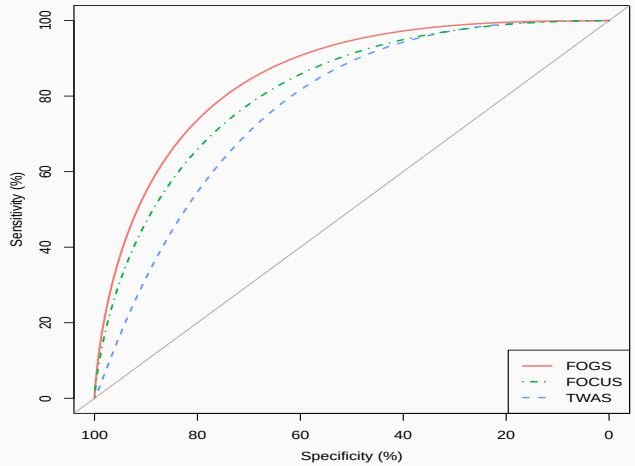
Step 2: Aggregating conditional Z scores to prioritize causal gene

- Sum test: $T_{\text{Sum}} = \sum_{j=1}^{p} U_j$
  SSU test: $T_{\text{SSU}} = U^T U = \sum_{j=1}^{p} U_j^2$
- More generally, for an integer $\gamma \geq 1$, an SPU(γ) test is defined as: $T_{\text{SPU}(\gamma)} = \sum_{j=1}^{p} U_j^{\gamma}$
- for an even integer $\gamma \to \infty$,
  $T_{\text{SPU}(\gamma)} \propto \left(\sum_{j=1}^{p} |U_j|^{\gamma} \right)^{1/\gamma} \to \max_j |U_j| = T_{\text{SPU}(\infty)}$
- $T_{\text{aSPU}} = \min_{\gamma \in \Gamma} P_{\text{SPU}(\gamma)}$, where $P_{\text{SPU}(\gamma)}$ is the p-value of the SPU(γ) test

Outline

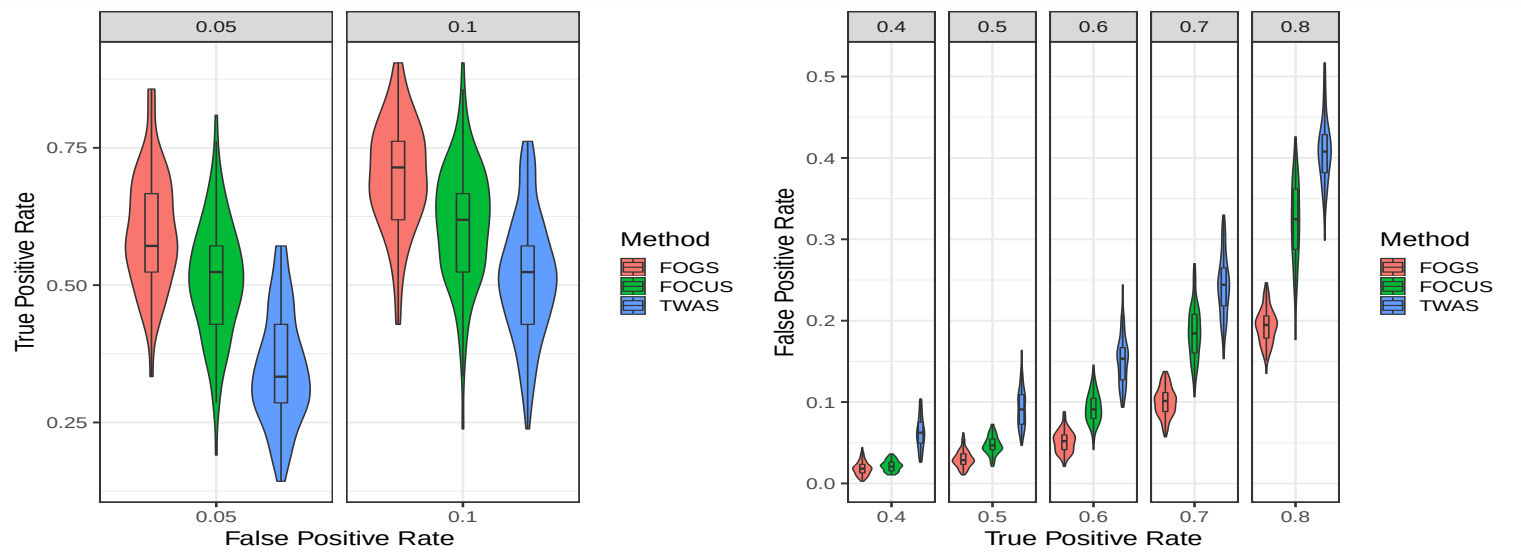
- Background
- Methods
- **Results**
- Discussion

Simulation: FOGS prioritizes and improves resolution for fine-mapping causal genes

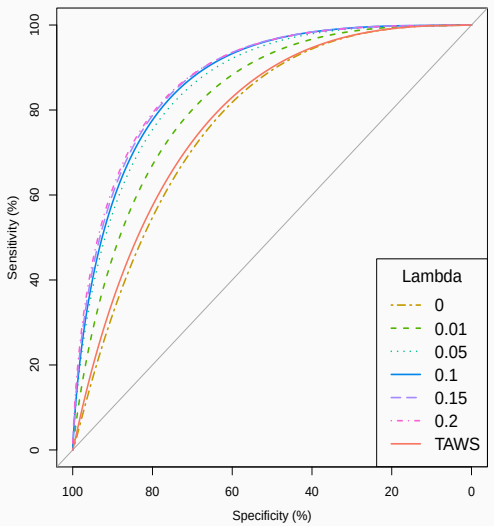


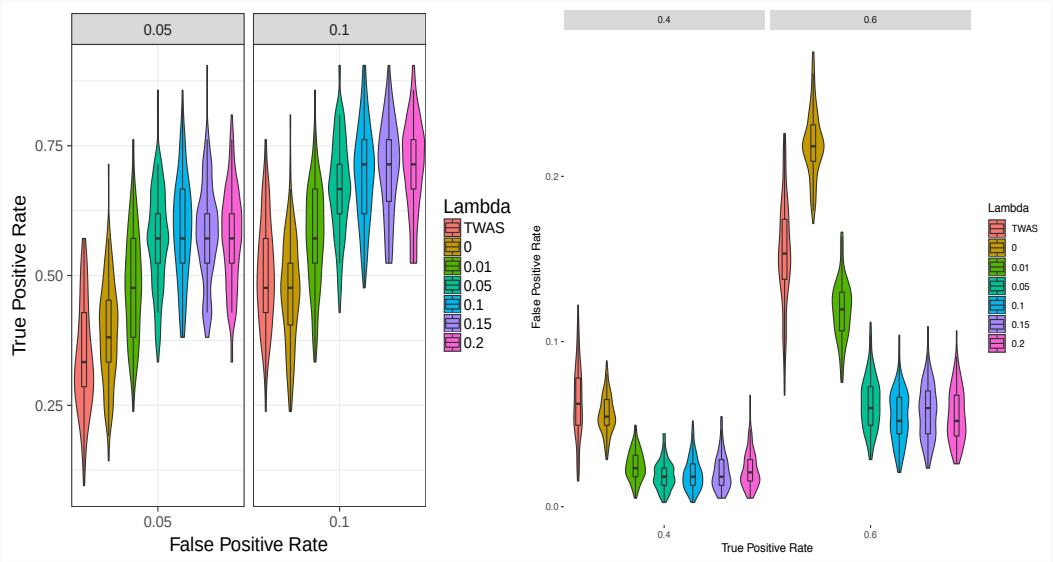
- Use chromosome 22 for all simulations; use lung health study for genotype
- We randomly selected two SNPs in one gene to be causal, and the effect size was $c = 0.1$. The estimated hertibability was about 2.5%.

Simulation: FOGS prioritizes and improves resolution for fine-mapping causal genes



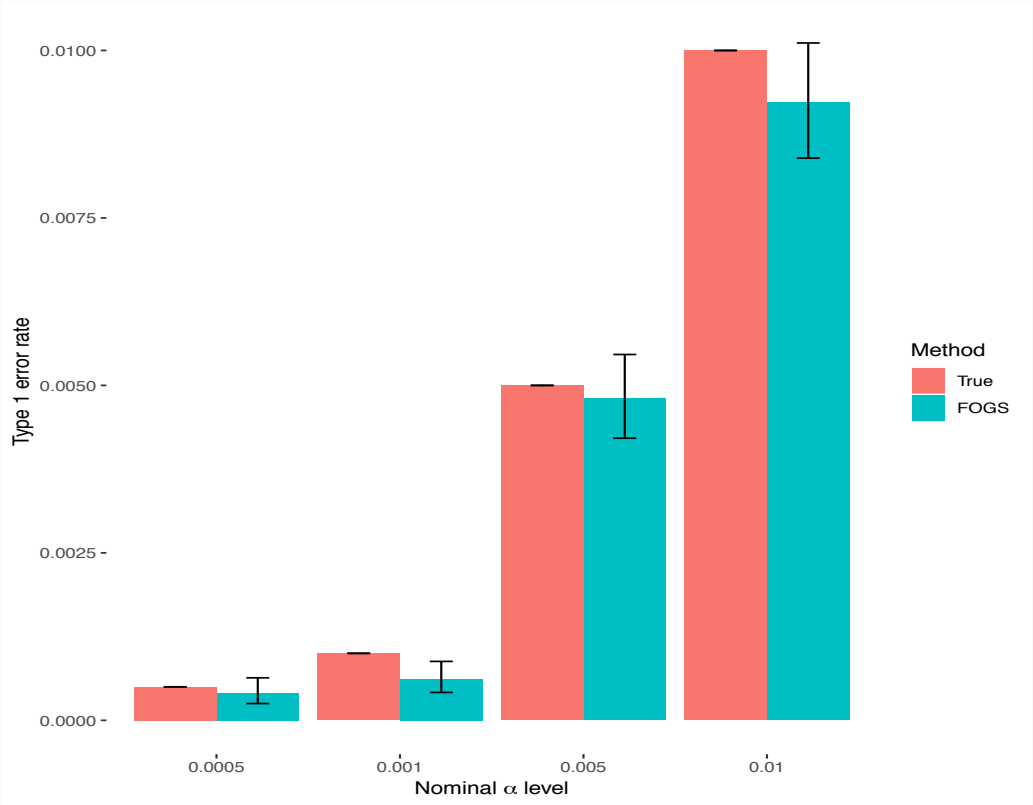
Simulation: FOGS is robust to the choice of penalty parameter λ





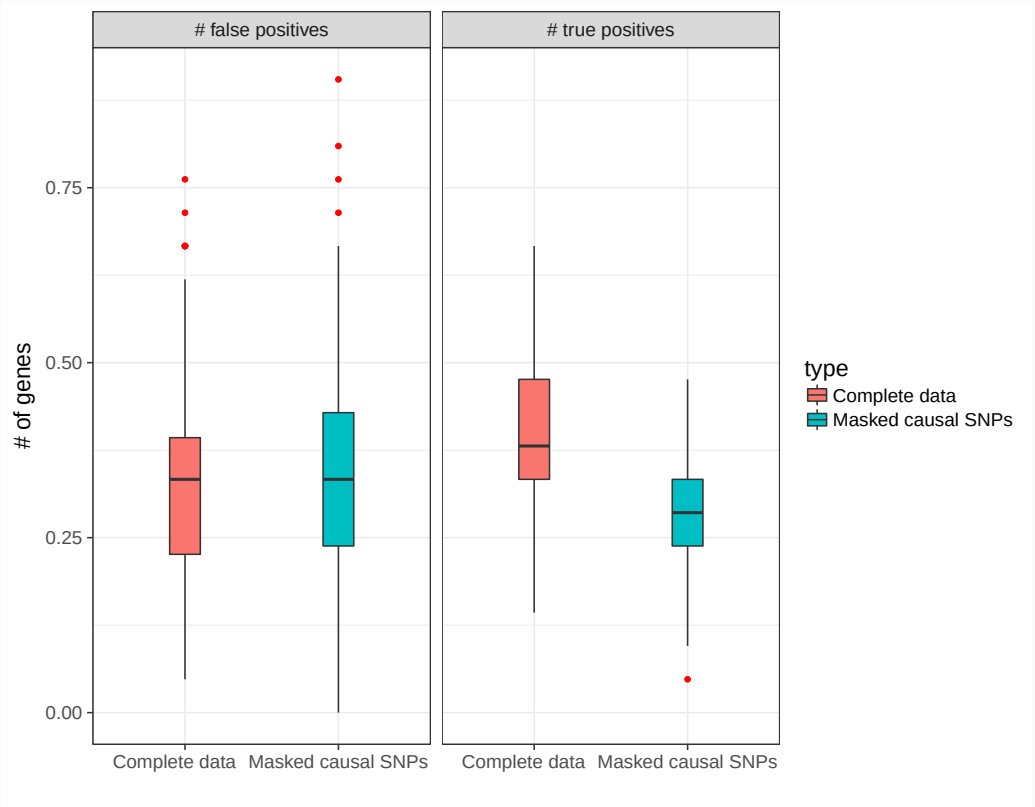
Simulation: Robustness analysis of FOGS

No SNP-trait association for all SNPs in the locus (under the null):



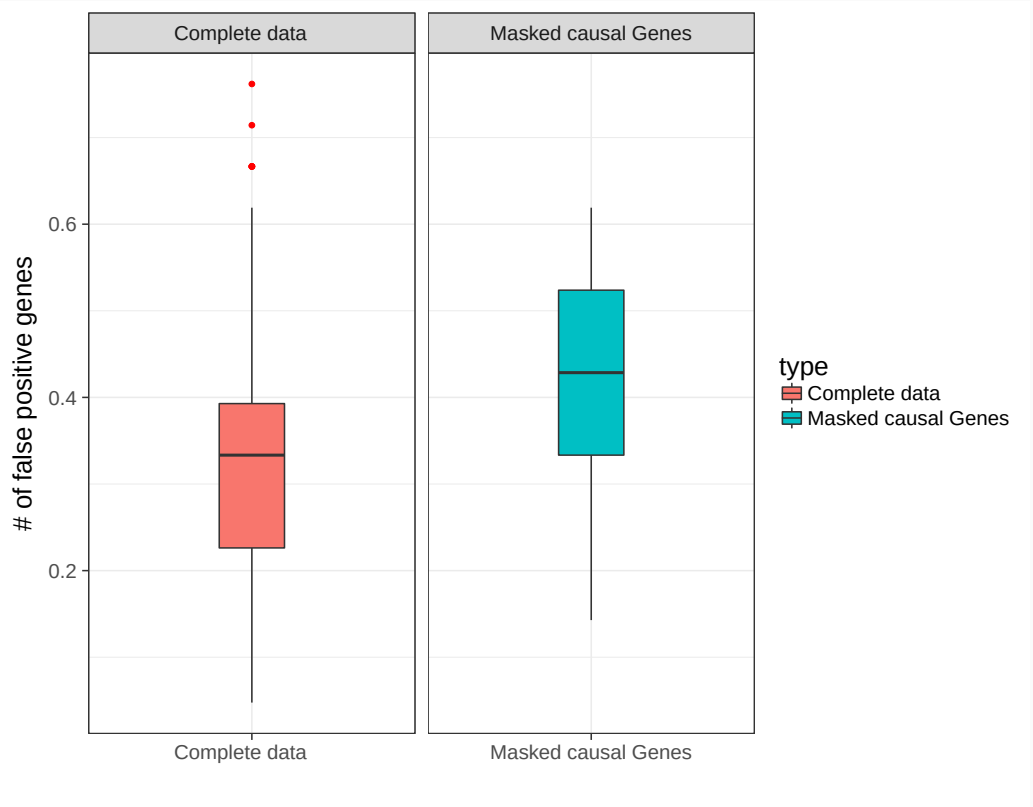
Simulation: Robustness analysis of FOGS

The two causal SNPs were missing:



Simulation: Robustness analysis of FOGS

The causal gene (with all its SNPs) was missing:



Application to a schizophrenia GWAS summary dataset

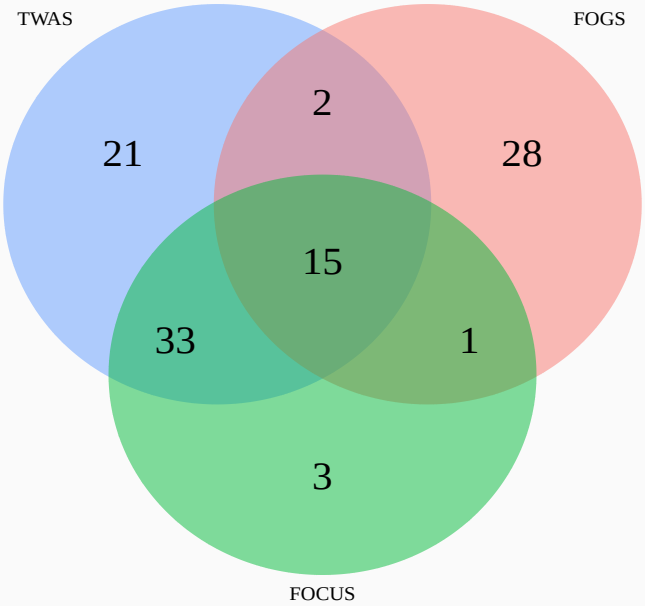


Figure 1: Diagram of the putative causal genes prioritized by different methods for the risk regions that contained at least two genes

Application to a schizophrenia GWAS summary dataset

- Both FOGS and FOCUS identified positive control: *C4A*
- FOGS identified some putatitve causal genes (such as *RGS6* and *B3GAT1*) that have some biological support but ignored by FOCUS.

Outline

- Background
- Methods
- Results
- Discussion

Discussion

- We introduce FOGS, a new method to prioritize putative causal genes for TWAS
- FOGS adequately controls Type I error rates and achieves high power under various alternatives
- Software: `https://github.com/ChongWu-Biostat/FOGS`
- Manuscript: Wu, C, Pan, W. (2020) A powerful fine-mapping method for transcriptome-wide association studies. *Human Genetics*, 139(2), 199–213.

Acknowledgment

- MSI@UMN
- Supported by NIH
- We appreciate the availability of the dbGaP data

Thank you!