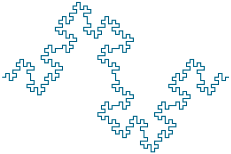
Bayesian TWAS: a causal inference approach

Xiaoquan (William) Wen

Biostatistics, University of Michigan



Background and Introduction

- How eQTL data can aid us to interpret GWAS results

- TWAS (PrediXcan/MetaXcan, Fusion, SMR/GSMR) as an effective integrative analysis approach

- Is TWAS causal inference?

Philosophical discussion of causal inference with observational data

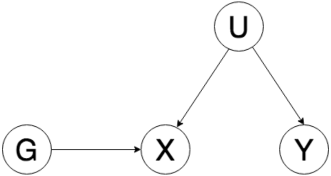
- No inference approaches can eliminate the effects of confounding

- ”Shoe leather” methodology (Freedman, 1991)

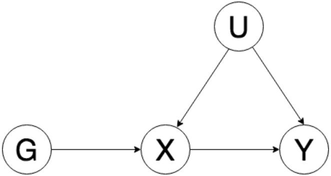
  ”exploits natural variation to mitigate confounding and relies on intimate knowledge of the subject matter to develop meticulous research designs and eliminate rival explanations”

Instrumental Variable Analysis

- The null model



- The causal model



Assumptions for IV analysis

- I: Inclusion assumption (link between G and X)
- R: Randomization assumption (no link between G and U)
- E: Exclusion assumption (no link between G and Y)

Note that R and E are theoretically un-testable.

Testing causal link from gene to trait

- Information is fully encoded in the DAGs

$$H_0 : \; G \perp\!\!\!\perp Y \qquad \text{vs.} \quad H_1 : \; G \not\perp\!\!\!\perp Y$$

- Sufficient to establish the causal link by testing association between eQTLs and traits*, **if IV assumptions hold*** (Katan, 1986)

- Implication: colocalization $\implies$ causality*, **if IV assumptions hold***.

- Why not estimation?

Statistical Issues in TWAS

From the perspective of IV analysis

- Strength of individual eQTLs (weak vs. strong)
  - Individual eQTLs are typically not strong instruments
- Linkage disequilibrium
- Number of independent eQTLs per gene
- Study designs: one-sample (G, X, Y) vs. two-sample $(G, X) \cup (G', Y')$
- Can we check exclusion assumption?

Key Idea: use of multiple instruments

- Wide-spread allelic heterogeneity suggests multiple independent IVs are available for a single gene
- Composite IV/allele score $(\sum_i w_i G_i)$ has better power over a single IV (Pierce et al, 2011; Burgess, 2013)
  - Is there an optimal weight? What is a principled way to construct weights
- Multiple IVs enable checking severe departure from exclusion assumption
  - Heterogeneity between estimated effects by independent IVs should be constrained

Method Outline

- The ability to distinguish SNPs in LD vs. independent association signals is critical
- Construct composite IV via Bayesian model averaging (BMA) and two-stage least squares (2SLS)
- Examine heterogeneity between estimates from independent association signals (eQTLs)

Probabilistic representation of Genetic Association Discovery/eQTLs

- Motivated by Bayesian credible sets by Maller et al. 2012
- Each association model is also assessed with a model-level probability, P_M
- Simultaneous construction of Bayesian credible sets for multiple association signals
  - Each eQTL/association signal is represented by a group of SNPs in LD
  - Strength of a signal is quantified by a probability, q
  - Strength of a member SNP is quantified by a probability, p

$$q = \sum_i p_i$$

- Software implementation DAP-G: (Lee et al., bioRxiv doi:10.1101/316471, `https://github.com/xqwen/dap`)

Construction of composite IV by BMA

1. Fit each noteworthy (sparse) candidate model $i = 1, ..., K$ by least squares, and obtain $\{\hat{\beta}_{M_i,j}\}$

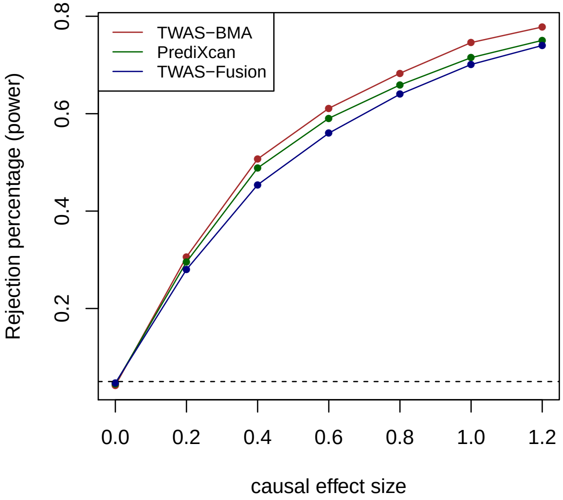
2. For each SNP j, compute the weight by averaging its estimated effects across K models

$$w_j = \sum_{i=1}^{K} P_{M_i} \hat{\beta}_{M_i,j}$$

3. The resulting composite IV is given by

$$\sum_{j=1}^{p} w_j G_j = \sum_{i=1}^{K} P_{M_i} \left(\sum_{j=1}^{p} \hat{\beta}_{M_i,j} G_j \right) = \sum_{i=1}^{K} P_{M_i} \, \hat{x}_{M_i}$$

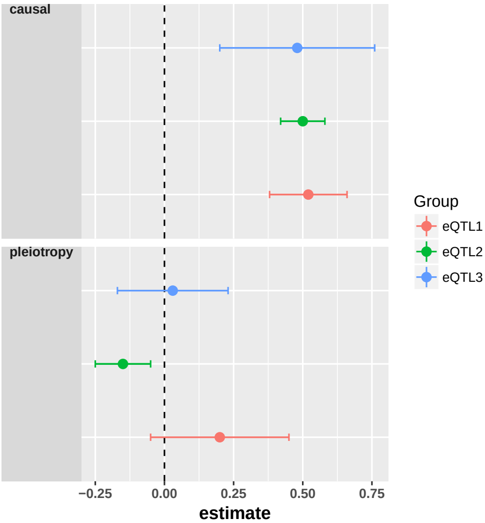
Simulation: Power of composite IV by BMA



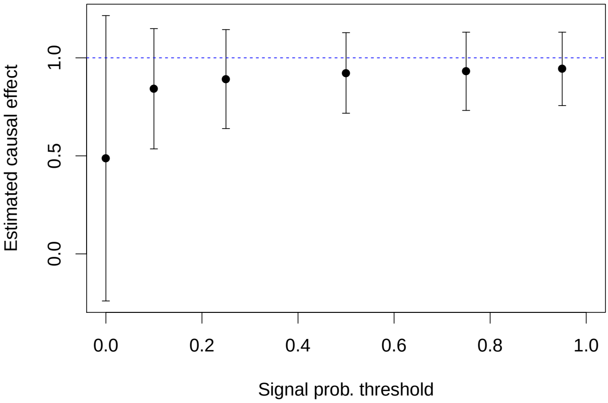
Consistency of causal effects by multiple IVs

1. For a signal cluster with signal-level prob $q = \sum_{i=1}^{m} p_i$, re-normalize $\tilde{p}_i = p_i/q$ for each member SNP i
2. Compute single-SNP 2SLS/Wald ratio estimate $\hat{\beta}_{xy}$
3. Signal-level estimate $\hat{\beta}_{xy} = \sum_{i=1}^{m} \tilde{p}_i \hat{\beta}_{xy,i}$
4. Assess heterogeneity of $\hat{\beta}_{xy}$ from multiple signals by computing Cochran's Q and I^2

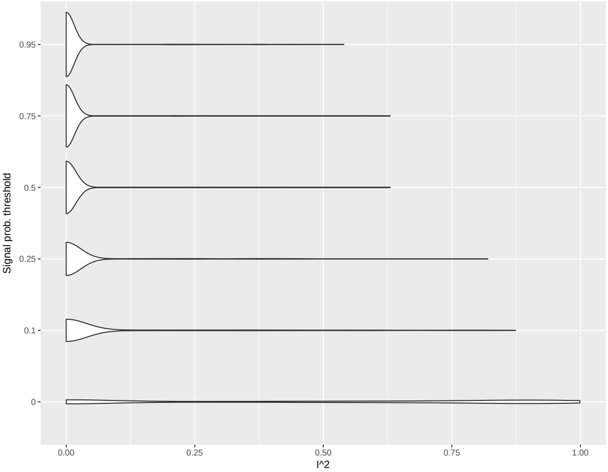
Simulation: Identifying severe violation of exclusion assumption



SIMULATION: ACCURACY OF CAUSAL EFFECT ESTIMATION VS. STRENGTH OF IVS



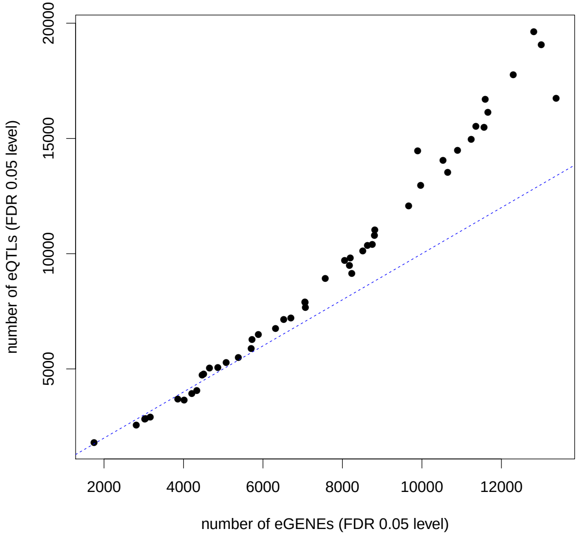
Simulation: I^2 distribution vs. strength of IVs



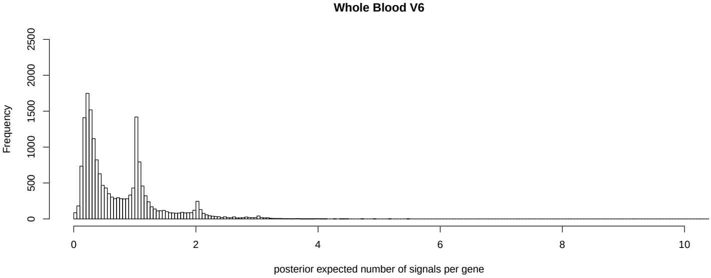
Analysis of GTEx and Complex Traits Data

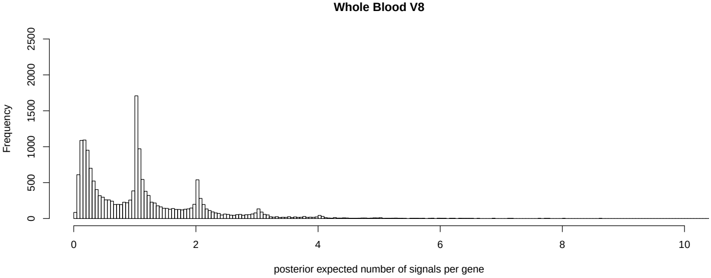
- GTEx v8 fine mapped *cis*-eQTLs across 49 tissues by DAP-G
- Construct composite IVs for all the eGenes in each tissue
- Perform TWAS using GAMBIT and MetaXcan in each tissue using summary statistics from UK Biobank
- Combine *p*-values across tissues using ACAT (Liu et al, 2018)
  - Software packge *GAMBIT* `https://github.com/corbinq/GAMBIT`

Allelic Heterogeneity Shown in GTEx data



Allelic Heterogeneity Shown in GTEx data

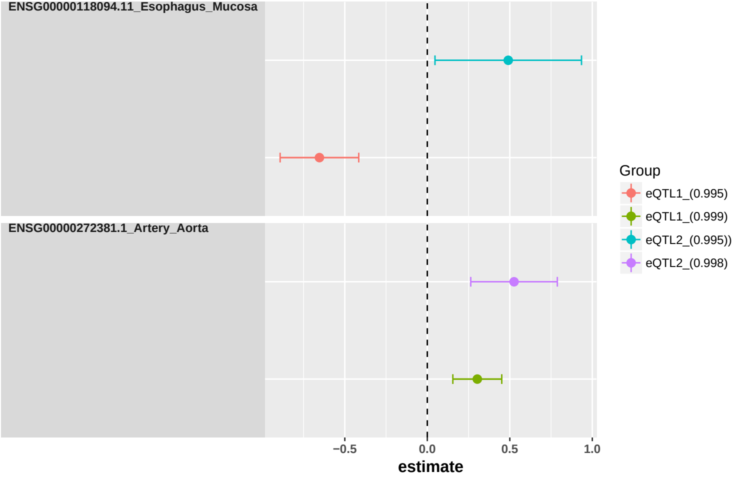




Preliminary results

- Seemingly finding more TWAS signals than PrediXcan
- Vast majority of top signals overlaps with PrediXcan
- Able to perform heterogeneity checking for 10% of top signals
    - Most top signals have $I^2 = 0$ with exceptions

Preliminary results



Challenges and opportunities

- We need more eQTL data.
- Two-sample design: a blessing and a curse
- What about one-sample design? Be aware of weak instruments
- Beyond **T**WAS

Acknowledgment

- Yuhua Zhang (Michigan)
- Corbin Quick (Michigan)
- Alvaro Barbeira (Chicago)
- Hae Kyung Im (Chicago)
- Xiang Zhou (Michigan)
- Roger Pique-Regi (Wayne State)
- Francesca Luca (Wayne State)