Integrating multimodal neuroimaging with GWAS for identifying modality-level causal pathways to Alzheimer's disease

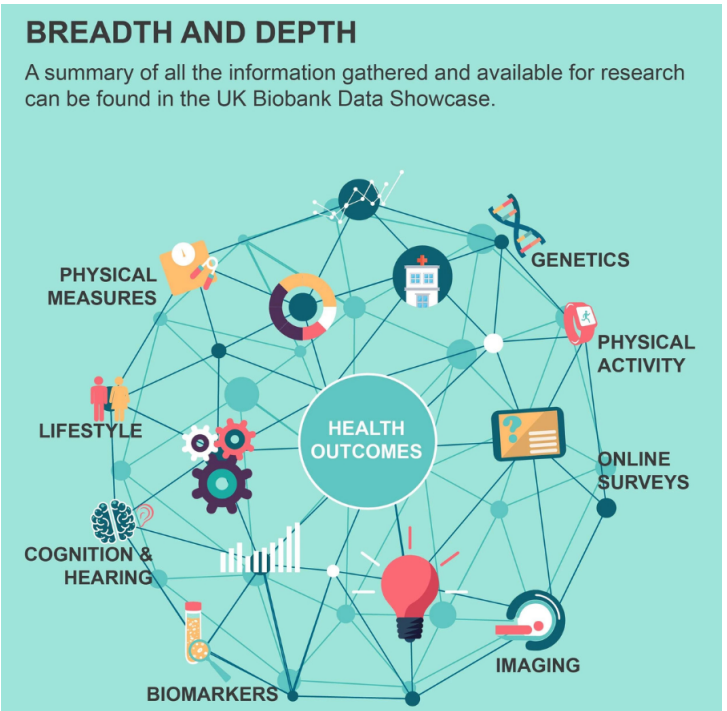
Jun Young Park

Department of Statistical Sciences
Department of Psychology
University of Toronto

Statistical Methods in Imaging 2024
May 30, 2024

GENETICS - IMAGING - TRAIT

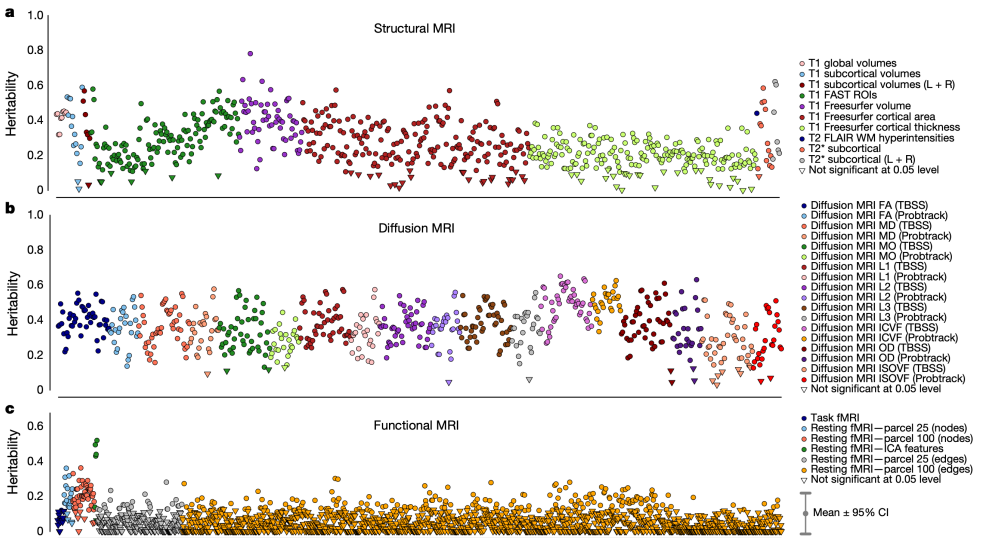
- **UK Biobank**
  - contains more than 40K subjects' brain MRI data
  - summary-level imaging data available ('imaging-driven phenotypes (IDPs)')
  - Other data types:
    - Genotypes (i.e., SNPs): $> 500K$ subjects
    - electronic health records data (e.g., ICD-10 codes)
    - and more
- **Research question**: How to integrate genetics-imaging-trait effectively to understand pathways to Alzheimer's disease (AD)?



Heritability of IDP

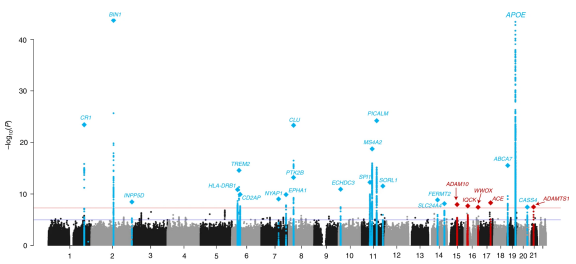
across different imaging modalities (Elliott et al., 2018)

- Thousands of IDPs available
- Brain IDPs are heritable, but the the effect sizes (h^2) vary by imaging modalities (sMRI, dMRI, fMRI)



IMAGING GENETICS FOR BRAIN DISORDER

GWAS (GENOME-WIDE ASSOCIATION STUDY)



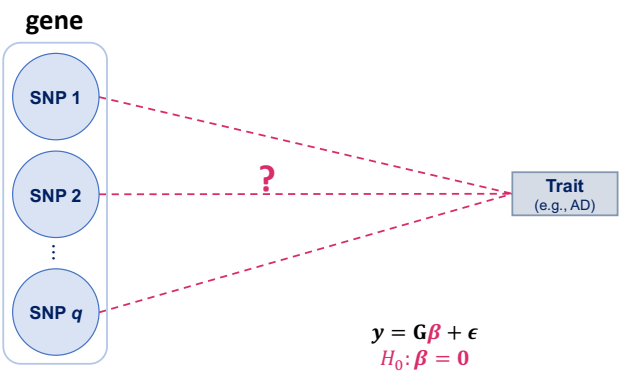
$$\boldsymbol{y} = \mathbf{g}\beta + \boldsymbol{\epsilon}$$
$$H_0: \beta = 0$$

- Notes:
  - **GWAS**: use a linear model to test and identify loci associated with a phenotype
  - Millions of SNPs to be tested: $p < 5 \times 10^{-8}$ to account for multiple testing
  - Most genetic effects are tiny, so using linear regression is valid for binary traits.

IMAGING GENETICS FOR BRAIN DISORDER

GENE-BASED ASSOCIATION TESTING

- **Gene-based testing**: borrow information across SNPs in a gene (out of 20K genes) to improve statistical power
  - (+) a relaxed threshold of $p < 0.05/20000 \approx 2.5 \times 10^{-6}$ (compared to $p < 5 \times 10^{-8}$)

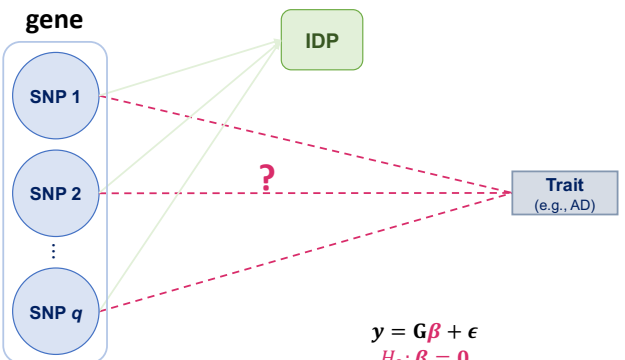


- No uniformly most powerful test!
  - Many tests are developed $\Rightarrow$ burden test, SKAT, ACAT, etc
  - Often formulated as finding the 'best weight' for the marginal correlation vector: $T = f(z_1, \ldots, z_q) = \sum_{j=1}^{q} w_j z_j$

IMAGING GENETICS FOR BRAIN DISORDER

TWAS (GUSEV ET AL., 2016) OR IWAS (XU ET AL., 2017)

- '*Can we use biologically-informed weights?*'
  - Obtain weights $\mathbf{w} = (\hat{w}_1, \ldots, \hat{w}_q)'$ by fitting $\mathbf{m} \sim \mathbf{G}$ (e.g., LASSO)
- **Insight of IWAS**: weights derived from brain MRI would give higher power for AD than using gene expression

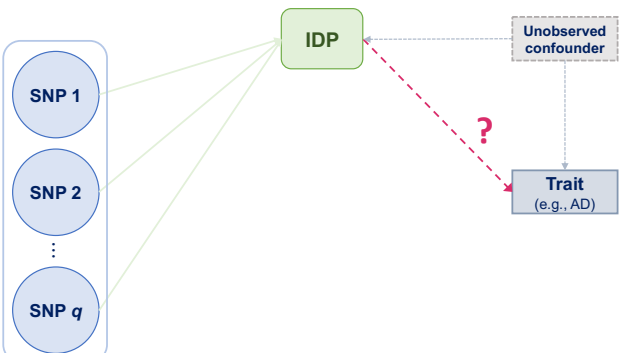


- IWAS is equivalent to testing for the correlation between $\mathbf{y}$ and $\hat{\mathbf{m}} = \mathbf{G}\hat{\mathbf{w}}$
  - $\hat{\mathbf{m}}$= 'genetically imputed' IDP
  - Shown to be powerful in ADNI study (Xu et al., 2017), when weights are derived from gray-matter volume from DMN-related ROIs.

IMAGING GENETICS FOR BRAIN DISORDER

'CAUSAL' INTERPRETATION OF IWAS (BRAINXCAN, LIANG ET AL., 2021)

- '*Wait, this looks like causal inference!*'
  - **IDP** $\xrightarrow{?}$ **AD**
  - Instrumental variable (IV) approach, implemented in 2-stage least squares (2SLS)
  - SNPs are used as candidate IVs.
- Directed Acyclic Graph (DAG) provides a good data-generating model



- (Standard) IV assumptions:

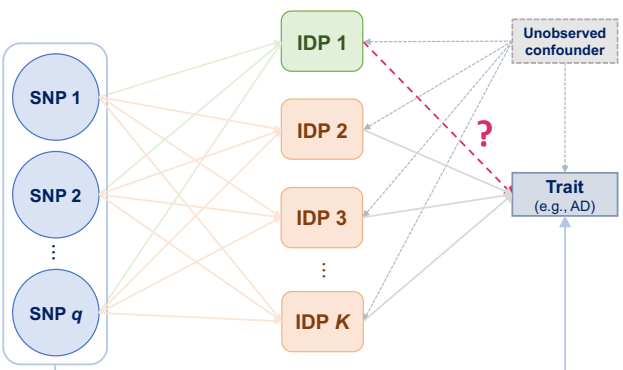
  A1. SNPs directly affect the IDP
  A2. SNPs do not affect unobserved confounders
  A3. SNPs do not affect the trait, except through the IDP.

CHALLENGES IN ACHIEVING CAUSALITY

A MORE REALISTIC DAG IN MULTIVARIATE IWAS (KNUTSON ET AL., 2020)

'*Is the previous DAG a realistic model?*'

1. **(Horizontal) pleiotropy**
   - thousands of IDPs (or even more) collected in neuroimaging
     $\Rightarrow$ observed confounders
   - SNPs affect other IDPs
   - SNPs affect the trait through other IDPs
2. **Direct effects**
   - SNP effects to AD, *not* regulated by imaging



$\Rightarrow$ Ignoring these effects leads to invalid inference (inflated T1E)

CHALLENGES IN ACHIEVING CAUSALITY

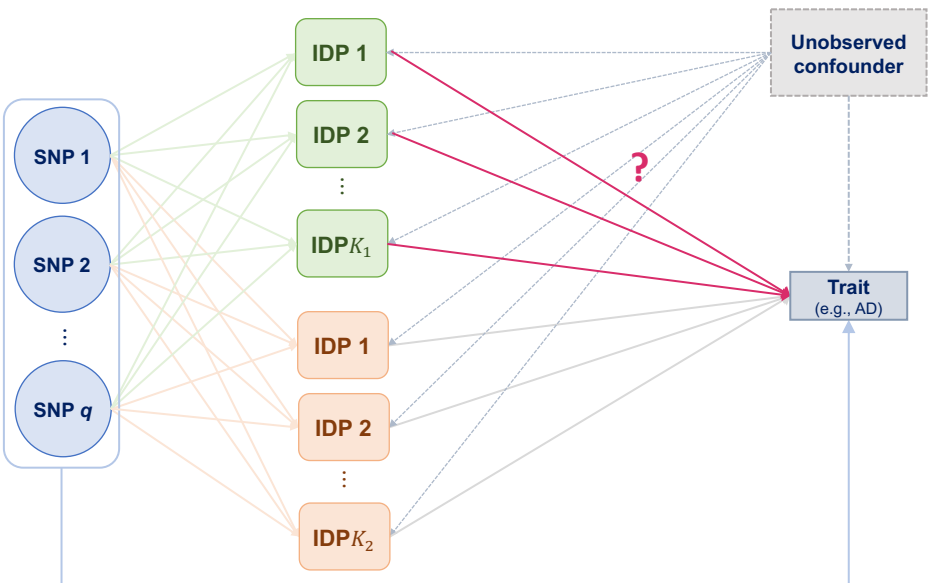
MULTIVARITE IWAS (KNUTSON ET AL., 2020)

$$\mathrm{E}[\mathbf{y}] = \underbrace{\widehat{\mathbf{m}}_1\beta_1}_{\text{our interest}} + \underbrace{\sum_{k=2}^{K}\widehat{\mathbf{m}}_k\beta_k}_{\text{horizontal pleiotropy}} + \underbrace{\sum_{j=1}^{q}\mathbf{g}_j\cdot\mu}_{\text{direct effect}}$$

- $H_0 : \beta_1 = 0$ (‘*test one IDP given all the others*’)
- Direct effect adjusted via Mendallian Randomization(MR)-Egger regression (Burgess et al., 2017).
- **Challenges**:
  1. **Multiple comparisons**
     Tests each pair of [gene-IDP] separately: 20000 genes $\times$ 1000 IDPs?
     $\Rightarrow$ MV-IWAS avoids gene-level analysis and focus on **IDP** $\overset{?}{\rightarrow}$ **AD**
  2. **Multicolinearity** between $\widehat{\mathbf{m}}_1$ and $\widehat{\mathbf{m}}_k$ (or $\sum_{j=1}^{q}\mathbf{g}_j$) affects power

PROPOSED METHOD

MV-VC-IWAS: TESTING **MODALITY-LEVEL** CAUSALITY



PROPOSED METHOD

MV-VC-IWAS

When $\mathbf{m}_1^{(1)}, \ldots, \mathbf{m}_{K_1}^{(1)}$ are IDPs in an imaging modality (e.g., sMRI) and $\mathbf{m}_1^{(2)}, \ldots, \mathbf{m}_{K_2}^{(2)}$ are IDPs in all the other modalities (e.g., dMRI, fMRI),

$$\mathrm{E}[\mathbf{y}] = \underbrace{\sum_{k=1}^{K_1} \widehat{\mathbf{m}}_k^{(1)} \beta_k^{(1)}}_{\text{our interest}} + \underbrace{\sum_{k=1}^{K_2} \widehat{\mathbf{m}}_k^{(2)} \beta_k^{(2)}}_{\text{horizontal pleiotropy}} + \underbrace{\sum_{j=1}^{J} \mathbf{g}_j \cdot \mu}_{\text{direct effect}}$$

- $H_0 : \beta_1^{(1)} = \cdots = \beta_{K_1}^{(1)} = 0$
  - ('*test one set given all the other sets*')
- Implementation via a variance component test (i.e., SKAT)

PROPOSED METHOD

MV-VC-IWAS

$$T_{\text{MV-VC-IWAS}} = \underbrace{\frac{1}{N}(\mathbf{y} - \mathbf{G}\widehat{\mathbf{W}}^{(2)}\widetilde{\mathbf{B}}^{(2)})'\mathbf{G}\widehat{\mathbf{W}}^{(1)}}_{\mathbf{s}'} \cdot \underbrace{\frac{1}{N}\widehat{\mathbf{W}}^{(1)\prime}\mathbf{G}'(\mathbf{y} - \mathbf{G}\widehat{\mathbf{W}}^{(2)}\widetilde{\mathbf{B}}^{(2)})}_{\mathbf{s}}.$$

where

$$\mathbf{s} \approx \widehat{\mathbf{W}}^{(1)\prime}\mathbf{z} - \widehat{\mathbf{W}}^{(1)\prime}\mathbf{R}\widehat{\mathbf{W}}^{(2)}(\widehat{\mathbf{W}}^{(2)\prime}\mathbf{R}\widehat{\mathbf{W}}^{(2)})^{-1}\widehat{\mathbf{W}}^{(2)\prime}\mathbf{z}$$

$$\widehat{\text{Cov}}(\mathbf{s}) \approx \widehat{\mathbf{W}}^{(1)\prime}\mathbf{R}\widehat{\mathbf{W}}^{(1)} - \widehat{\mathbf{W}}^{(1)\prime}\mathbf{R}\widehat{\mathbf{W}}^{(2)}(\widehat{\mathbf{W}}^{(2)\prime}\mathbf{R}\widehat{\mathbf{W}}^{(2)})^{-1}\widehat{\mathbf{W}}^{(2)\prime}\mathbf{R}\widehat{\mathbf{W}}^{(1)}.$$

- Implementation via summary statistics only!
  - $\mathbf{z}$: SNP-AD z statistics
  - $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(12}$: obtained by LD clumping or LassoSum (Mak et al., 2017)
  - $\mathbf{R}$: SNP correlation matrix obtained by a reference panel

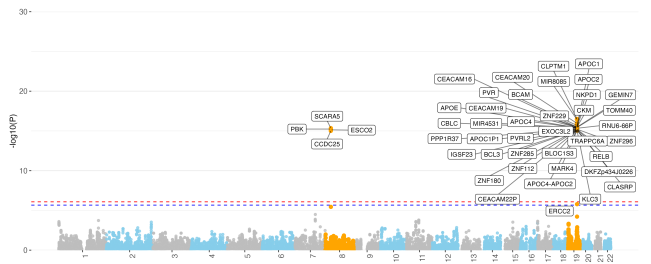
DATA DESCRIPTION

1. IDP-GWAS summary statistics
   - UK Biobank - obtained from the Oxford Brain Imaging Genetics Server - BIG40
   - obtained from 40K subjects whose genetic and imaging data are collected.
   - obtained for 1415 sMRI, 675 dMRI, and 210 fMRI (FC) features
   - adjusted for behavioral measures, artifacts (motion, site), and 40 genetic principal components (PCs)
2. AD-GWAS summary statistics
   - 2.1 International Genomics of Alzheimer's Project (IGAP)
     - 17K AD and 37K non-AD subjects
     - Adjusted for behavioral measures and (at least) 4 genetic principal components (PCs)
   - 2.2 UK Biobank - obtained from the Neal lab (`https://www.nealelab.is/uk-biobank`)
     - 4K AD (from ICD-10) and 357K non-AD subjects
     - Adjusted for behavioral measures and 20 genetic PCs
3. SNP correlation matrix
   - obtained from the 1000 Genome Project (ancestry: EUR)

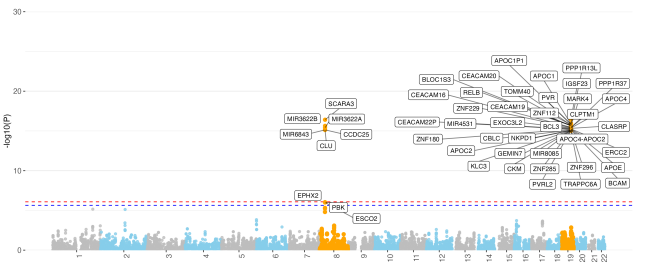
DATA ANALYSIS

1. IGAP

A. **Structural MRI** | other modalities

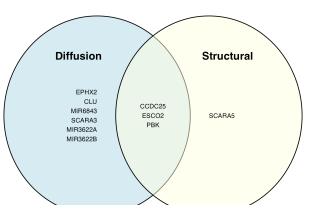


B. **Diffusion MRI** | other modalities

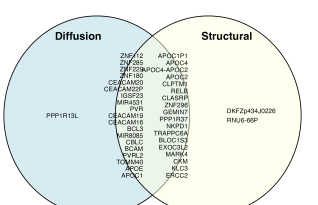


C. **Functional MRI** | other modalities

Chr 8



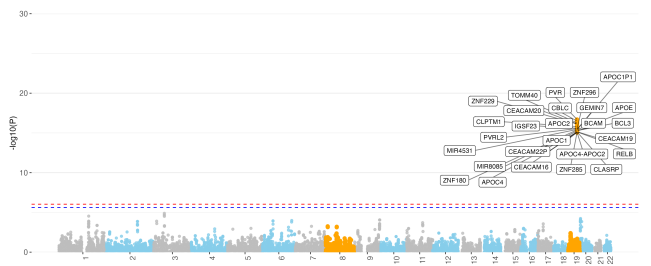
Chr 19



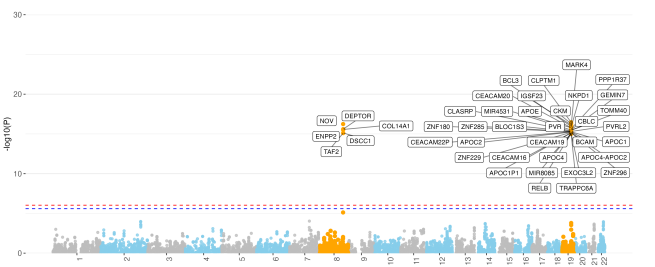
DATA ANALYSIS

2. UK BIOBANK

A. **Structural MRI** | other modalities

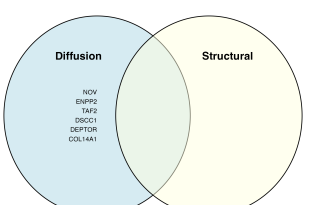


B. **Diffusion MRI** | other modalities

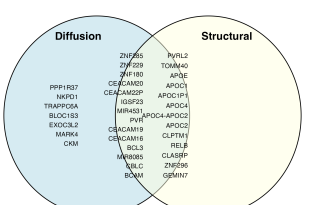


C. **Functional MRI** | other modalities

Chr 8



Chr 19

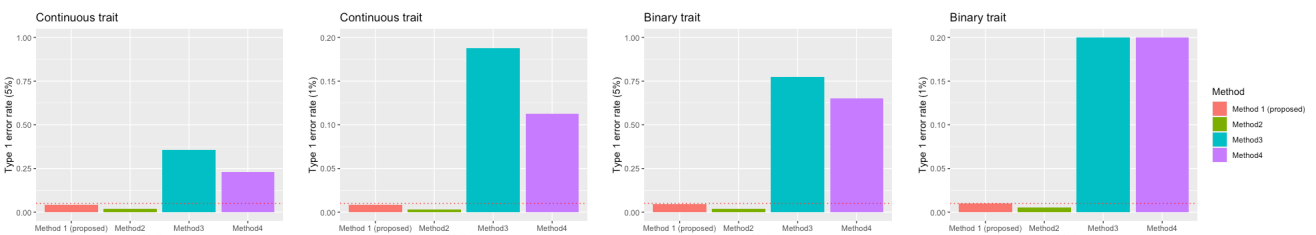


SIMULATION STUDIES

- Used the proposed DAG to generate simulated data
  - $q = 50$ SNPs, $K_1 = K_2 = 10$ IDPs
  - considered both a continuous trait and a binary trait (e.g., AD)
  - Parameters are chosen to ensure that the
    - Average SNP-imaging $R^2 < 0.1$
    - Average SNP-trait $R^2 < 0.1$ (cont)
    - disease prevalence is $\approx 20\%$
- Evaluated
  - Method 1: the proposed method
  - Method 2: the proposed method implemented in minP
  - Method 3: the proposed method without adjusting for pleiotropic effects
  - Method 4: univariate analysis without adjusting for pleiotropic effects
- Once data is generated, these were converted to summary statistics for implementation

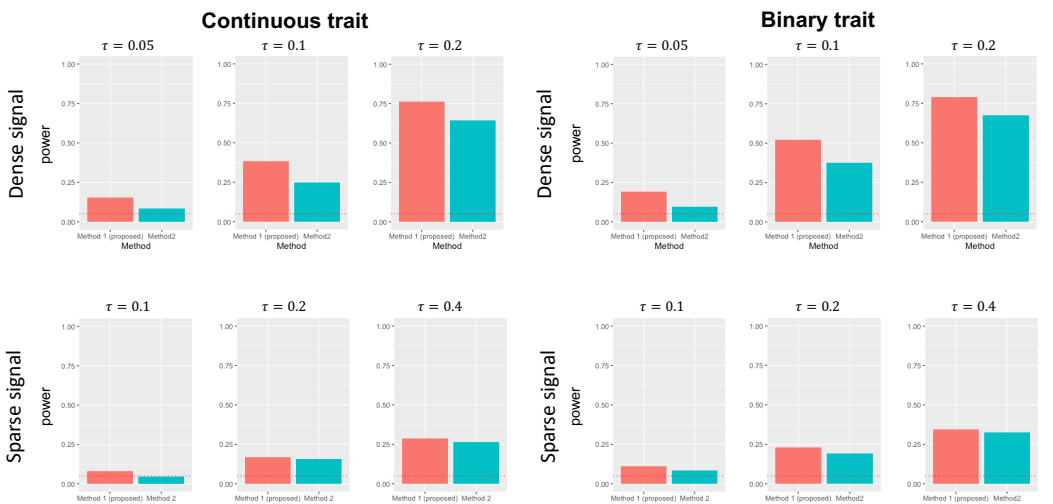
Simulation results

Type 1 error rate



SIMULATION RESULTS

STATISTICAL POWER

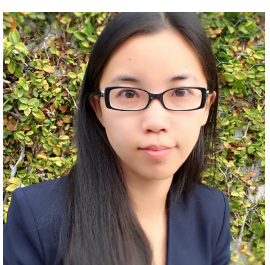


Discussions

- Our proposed method
  - extends MV-IWAS to a set-based testing, enhancing flexibility in research questions.
  - provides a more comprehensive understanding of gene-imaging-AD pathways.
  - is fully implemented via summary statistics
- Extensions beyond brain IDP
  - e.g., spatial-extent causal inference ('test and localize') after adjusting for other pleiotropic effects?
  - e.g., adjusting for other (non-brain) IDPs?
- Good stage 1 model would be more efficient
  - multi-task learning (with summary statistics)?
- Low-rank decomposition of imaging data $\Rightarrow$ JIVE?
  - limited by the lack of summary statistics

ACKNOWLEDGEMENT

- This is a joint work with Yuan Tian, a PhD student at the University of Toronto



- This research is supported by
  - Natural Sciences and Engineering Research Council of Canada Discovery Grant
  - McLaughin Centre Accelerator Grant

Thank you!

Any questions?