TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits

Jingjing Yang, PhD

CCQG
CENTER FOR COMPUTATIONAL
AND QUANTITATIVE GENETICS

EMORY
UNIVERSITY
SCHOOL OF
MEDICINE

Outline

Introduction

Methods

Transcriptome-wide Association Study

Gene Expression (GReX) Imputation Models

Gene-based Association Test

Results

Simulation Studies

Mapping Alzheimer’s Dementia Related Phenotypes

TWAS Based on SKAT

Application Results with ROS/MAP Data

Summary

Introduction

Methods

Results

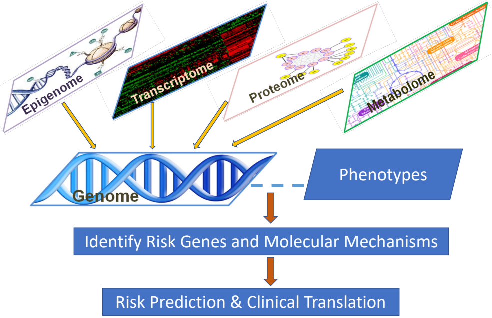
TWAS Based on SKAT

Summary

Etiology of Complex Diseases

Examples complex diseases
Type II Diabetes, Cardiovascular Diseases, Alzheimer’s Dementia

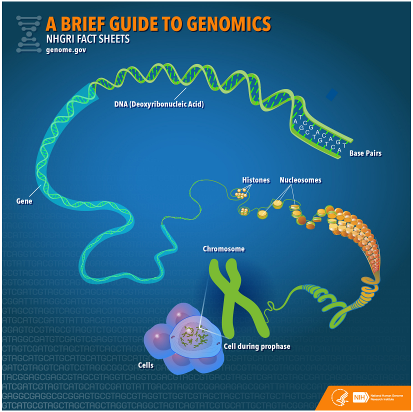
- Polygenic with low penetrance by individual genes
- Largely unknown genomic etiology
- Integrate multi-layers of Omics data

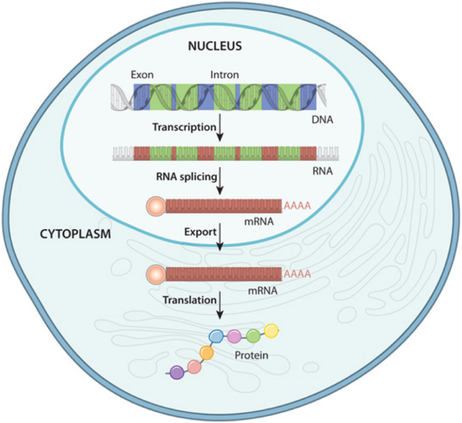


Identify Risk Genes and Molecular Mechanisms

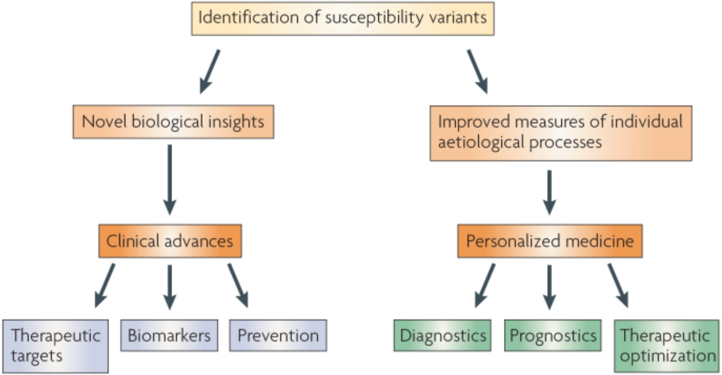
Risk Prediction & Clinical Translation

Overview of Genomics Data


A BRIEF GUIDE TO GENOMICS
NHGRI FACT SHEETS
genome.gov
DNA (Deoxyribonucleic Acid)
Base Pairs
Gene
Histones
Nucleosomes
Chromosome
Cell during prophase
Cells
NIH National Human Genome Research Institute

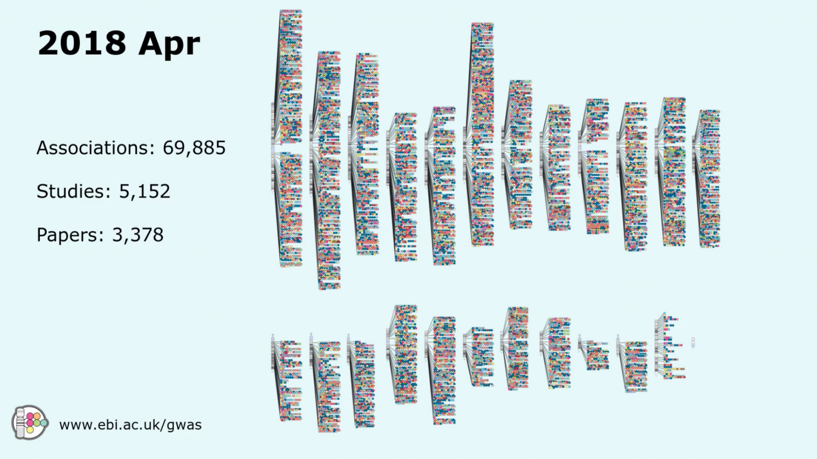


GOAL of Mapping Complex Human Diseases



McCarthy I.M. et. al. Nature Reviews. 2008.

GWAS Findings



GWAS: Genome-wide Association Study

Introduction

Methods

- Transcriptome-wide Association Study
- Gene Expression (GReX) Imputation Models
- Gene-based Association Test

Results

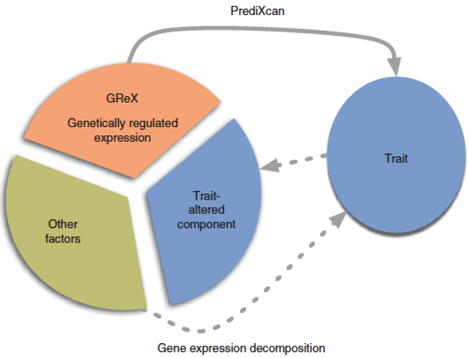
TWAS Based on SKAT

Summary

Integrate Transcriptomic Data in GWAS

Transcriptome-wide Association Study (TWAS)

- Leverage existing public transcriptomic data resources (e.g., GTEx, GEUVADIS, DGN)
- Conduct “Functional” gene-based association test
- Improve biological interpretation
- Identify novel risk genes

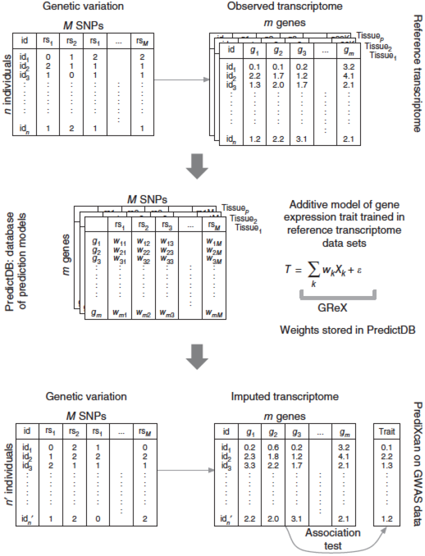

PrediXcan
GReX
Genetically regulated expression
Trait
Trait-altered component
Other factors
Gene expression decomposition

Gamazon ER et. al., Nat Genetics, 2015.

Existing Tools

- PrediXcan: based on the Elastic-Net penalized linear regression model (EN).
  Gamazon et. al., Nat Genetics, 2015.

- FUSION: based on the Bayesian Sparse Linear Mixed Model (BSLMM).
  Gusev et. al. Nat Genetics, 2016.



Genetic variation

M SNPs (*n* individuals)

id	rs$_1$	rs$_2$	rs$_1$	...	rs$_M$
id$_1$	0	1	2		2
id$_2$	2	1	1		1
id$_3$	1	0	1		1
⋮	⋮	⋮	⋮	⋮	⋮
id$_n$	1	2	1		1

Observed transcriptome

m genes

id	g_1	g_2	g_3	...	g_m
id$_1$	0.1	0.1	0.2		3.2
id$_2$	2.2	1.7	1.2		4.1
id$_3$	1.3	2.0	1.7		2.1
⋮	⋮	⋮	⋮	⋮	⋮
id$_n$	1.2	2.2	3.1		2.1

M SNPs (*m* genes)

	rs$_1$	rs$_2$	rs$_3$	...	rs$_M$
g_1	w_{11}	w_{12}	w_{13}		w_{1M}
g_2	w_{21}	w_{22}	w_{23}		w_{2M}
g_3	w_{31}	w_{32}	w_{33}		w_{3M}
⋮	⋮	⋮	⋮	⋮	⋮
g_m	w_{m1}	w_{m2}	w_{m3}		w_{mM}

Additive model of gene expression trait trained in reference transcriptome data sets

$$T = \underbrace{\sum_k w_k X_k}_{\text{GReX}} + \varepsilon$$

Weights stored in PredictDB

Genetic variation

M SNPs (*n′* individuals)

id	rs$_1$	rs$_2$	rs$_1$	...	rs$_M$
id$_1$	0	2	1		0
id$_2$	1	2	2		2
id$_3$	2	1	1		1
⋮	⋮	⋮	⋮	⋮	⋮
id$_n$′	1	2	0		2

Imputed transcriptome

m genes

id	g_1	g_2	g_3	...	g_m	Trait
id$_1$	0.2	0.6	0.2		3.2	0.1
id$_2$	2.3	1.8	1.2		4.1	2.2
id$_3$	3.3	2.2	1.7		2.1	1.3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
id$_n$′	2.2	2.0	3.1		2.1	1.2

Nonparametric Bayesian Model

Advantages

- Include parametric models (e.g., Elastic-Net, BSLMM) as special cases
- Better modeling the underlying complex genetic architecture of transcriptomic profiles
- Improve GReX imputation accuracy
- Improve TWAS power

Nonparametric Bayesian Model

- Considering gene expression levels $\mathbf{E_g}$ of gene g
  genotype data matrix $\mathbf{X_{n\times p}}$ of all cis-SNPs
- $\mathbf{E_g}$ are normalized and adjusted for confounding covariates such as age, sex, top genotype PCs, PEER factors of transcriptomic data
- The nonparametric Bayesian Dirichlet process regression (DPR) model (Zeng & Zhou, Nat. Comm., 2017) is setup as:

$$\mathbf{E_g} = \mathbf{X_{n\times p}} \mathbf{w_{p\times 1}} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2 \mathbf{I}),\ \sigma_\varepsilon^2 \sim IG(a_\varepsilon, b_\varepsilon)$$

$$w_i \sim N(0, \sigma_\varepsilon^2 \sigma_w^2),\ \sigma_w^2 \sim D,\ D \sim DP(IG(a,b), \xi),\ i = 1, \cdots, p$$

- Estimate cis-eQTL effect-sizes $\mathbf{w_{p\times 1}}$ by MCMC or Variational Bayesian Approximation

Nonparametric Bayesian Model

Another intuitive way of viewing this nonparametric model

- σ_w^2 can be viewed as a Latent variable
- Integrating out σ_w^2 will induce a Nonparametric prior distribution on w_i
- Equivalent to a normal mixture model for w_i

$$\begin{aligned} w_i &\sim \pi_0 N(0, \sigma_\varepsilon^2 \sigma_0^2) + \sum_{k=1}^{+\infty} \pi_k N(0, \sigma_\varepsilon^2 (\sigma_k^2 + \sigma_0^2)); \\ \pi_k &= v_k \prod_{l=0}^{k-1} (1 - v_l),\ v_k \sim Beta(1, \xi),\ \xi \sim Gamma(a_\xi, b_\xi); \\ \sigma_k^2 &\sim IG(a_k, b_k),\ k = 0, 1, \cdots, +\infty. \end{aligned}$$

Gene-based Association Test by Existing TWAS Tools

General framework with phenotype $\boldsymbol{Y}$, genotype matrix $\boldsymbol{X}$, and covariate matrix $\boldsymbol{Z}$

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

$$\widehat{\boldsymbol{GReX}} = \boldsymbol{X}\hat{\boldsymbol{w}}$$

$$H_0 : \beta = 0$$

Equivalent to a gene-based burden test taking cis-eQTL effect size estimates $\hat{\boldsymbol{w}}$ as variant weights

Introduction

Methods

Results

- Simulation Studies
- Mapping Alzheimer’s Dementia Related Phenotypes

TWAS Based on SKAT

Summary

Simulation Study Design

- Use the real genotype data of gene *ABCA7* with $2,799$ cis-SNPs with MAF > 5% and HWP $> 10^{-5}$
- Training sample size $(100, 300, 499)$, test sample size $1,200$
- Consider scenarios with various proportion of causal SNPs for gene expression, $p_{causal} = (0.01, 0.05, 0.1, 0.2)$
- Consider scenarios with various gene expression heritability and phenotype heritability, $(p_e^2, p_h^2) = ((0.05, 0.8), (0.1, 0.5), (0.2, 0.25), (0.5, 0.1))$
- Compare PrediXcan and DPR methods with respect to gene expression prediction R^2 and TWAS power

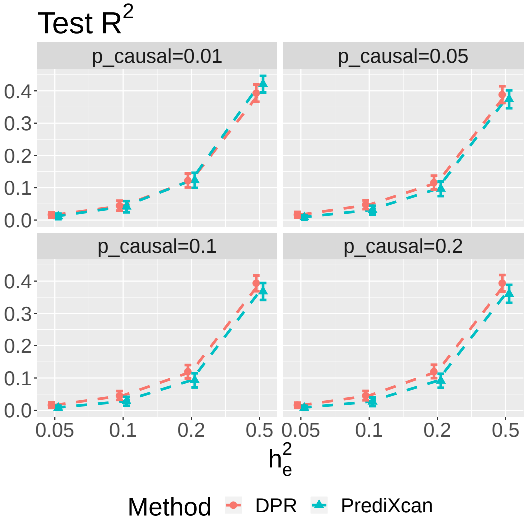


Figure 1: Gene expression prediction R^2 on test data.

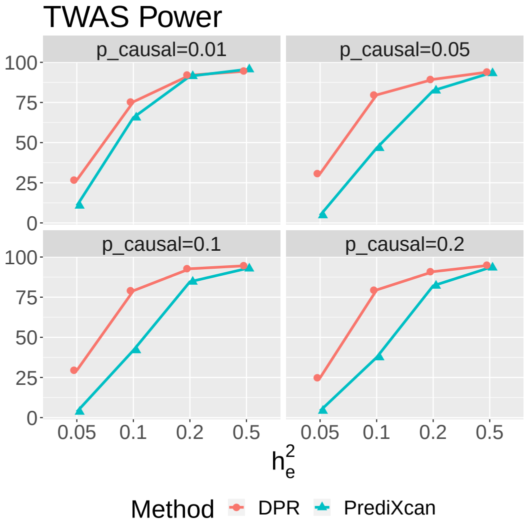


Figure 2: TWAS power with test data.

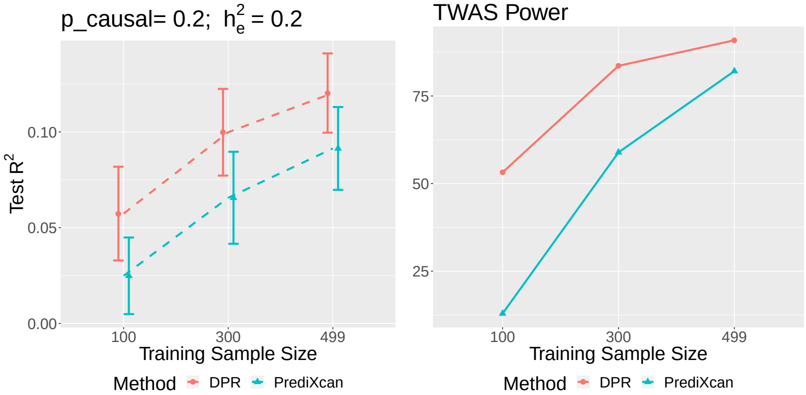
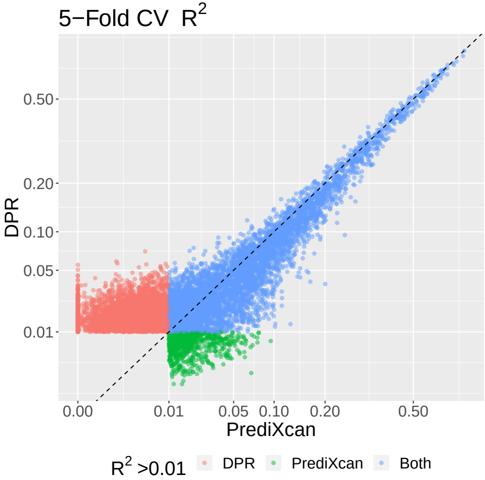


Figure 3: Gene expression prediction R^2 and TWAS power with various sample sizes.

ROS/MAP Data

- Prospective cohort studies of aging and dementia with participants of Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP)
- GWAS data of $2,093$ European samples
- RNAseq data (transcriptomic profiles) of 499 post-mortem brain samples that also have GWAS genotype data (after QC)
- Considered two important indices of Alzheimer's dementia pathology as quantitative complex traits
    - β-amyloid (Amyloid)
    - Neurofibrillary tangle density (Tangles)

PrediXcan vs. DPR



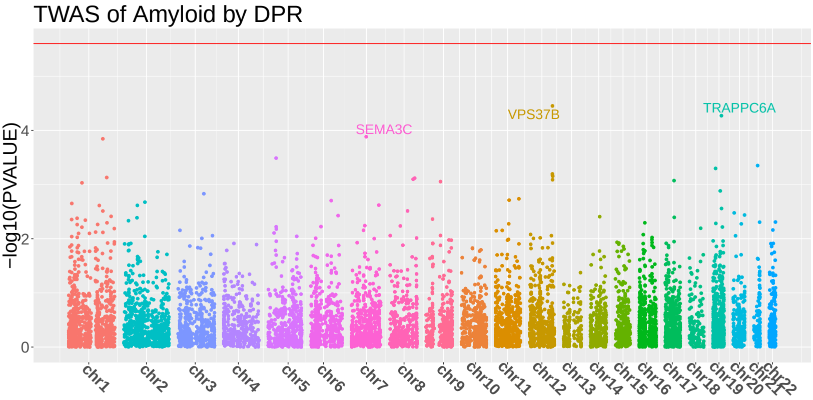


Figure 4: TWAS of β-Amyloid using DPR weights.

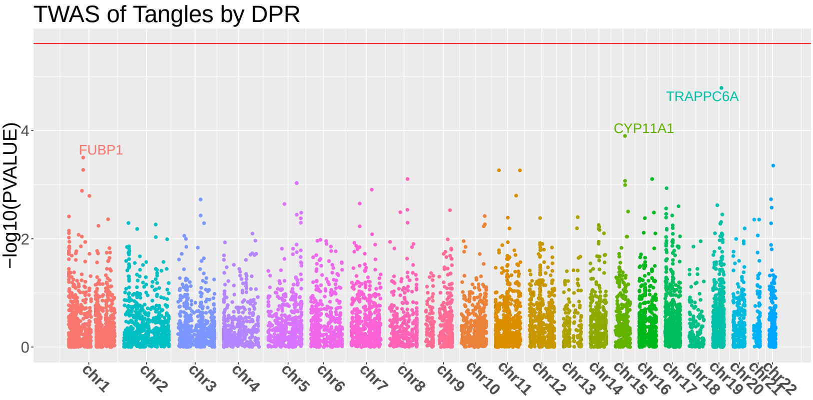


Figure 5: TWAS of Tangles using DPR weights.

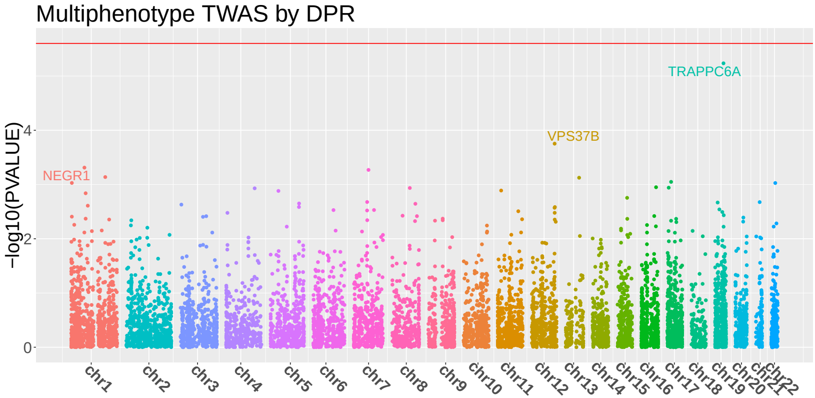
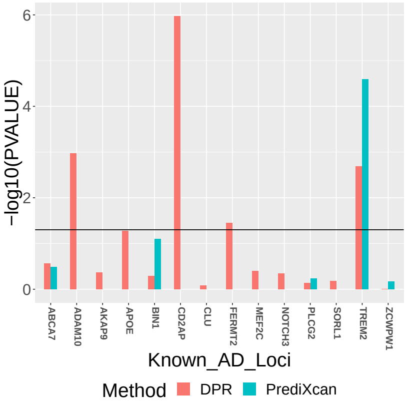


Figure 6: Multiphenotype TWAS with β-Amyloid and Tangles using DPR weights.

TWAS Results with GWAS Summary Statistics

TWAS of known AD loci using DPR weights estimated from ROS/MAP data and public GWAS summary statistics by IGAP



Introduction

Methods

Results

TWAS Based on SKAT

Application Results with ROS/MAP Data

Summary

SKAT TWAS

Sequencial Kernel Association Test (SKAT) (Wu et. al. AJHG, 2011)

- General framework with phenotype $\boldsymbol{Y}$, genotype matrix $\boldsymbol{X}$, and covariate matrix $\boldsymbol{Z}$
$$g(E[\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}]) = \boldsymbol{\beta}'\boldsymbol{X} + \boldsymbol{\alpha}'\boldsymbol{Z},\ \beta_i \sim N(0, w_i^2\tau)$$

- $H_0 : \tau = 0$

- Variance-component score statistic with a diagonal weight matrix $\boldsymbol{W}$ and phenotype mean $\hat{\boldsymbol{\mu}}$ estimated under H_0
$$\boldsymbol{Q} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})'\boldsymbol{K}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}),\ \boldsymbol{K} = \boldsymbol{XWX}'$$

- TWAS: use cis-eQTL effect size estimates $\widehat{w}_i$ by DPR method as variant weights, $W_{i,i} = \widehat{w}_i{}^2$

- Q follows a mixture chi-square distribution under H_0

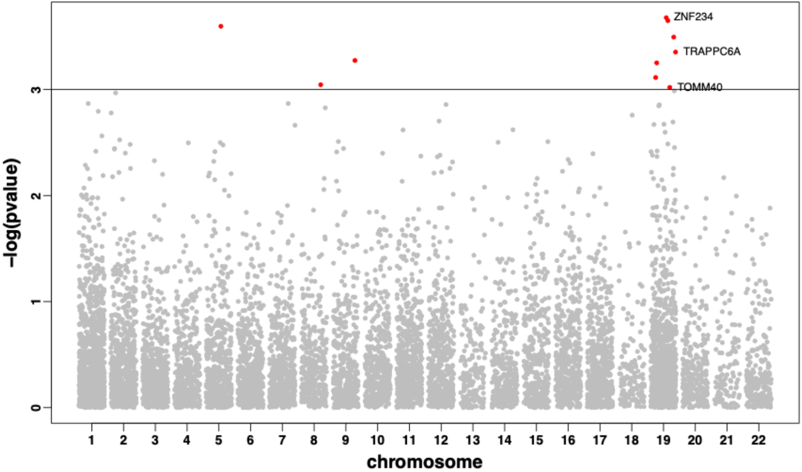


Figure 7: SKAT TWAS with β-Amyloid.

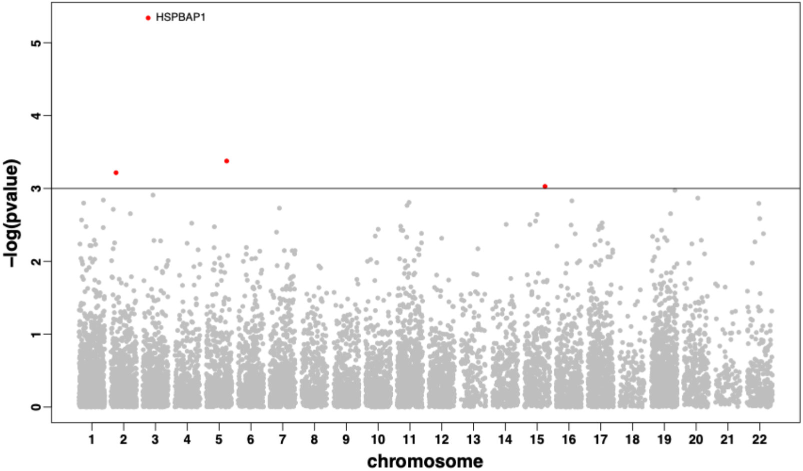


Figure 8: SKAT TWAS with Tangles.

Introduction

Methods

Results

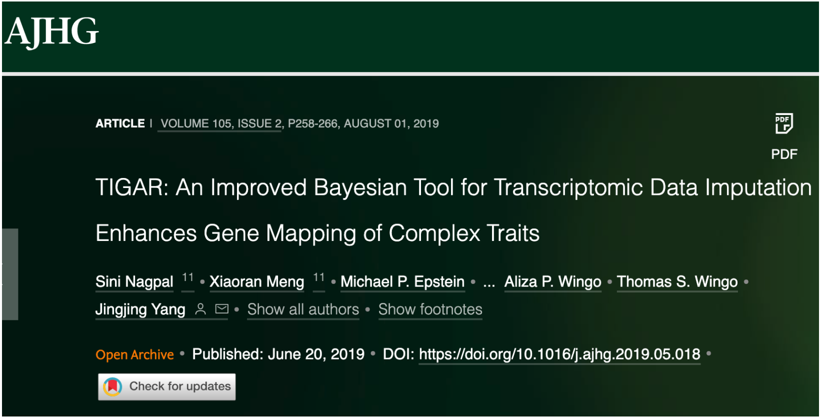
TWAS Based on SKAT

Summary

Summary

- Nonparametric Bayesian method is preferred when the proportion of causal SNPs > 0.01 or expression heritability < 0.2
- TWAS results can help interpret significant risk gene loci
- Promising TWAS results in ROS/MAP application studies by using nonparametric Bayesian method
    - Potentially novel loci *TRAPPC6A, ZNF234, HSPBAP1* for AD pathological indexes
    - Known AD loci *ADAM10, CD2AP, TREM2* identified by TWAS
- Multiple phenotype TWAS can leverage pleiotropy

Published Paper



AJHG

ARTICLE | VOLUME 105, ISSUE 2, P258-266, AUGUST 01, 2019

PDF

TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits

Sini Nagpal [11] • Xiaoran Meng [11] • Michael P. Epstein • ... Aliza P. Wingo • Thomas S. Wingo • Jingjing Yang • Show all authors • Show footnotes

Open Archive • Published: June 20, 2019 • DOI: https://doi.org/10.1016/j.ajhg.2019.05.018 •

Check for updates

Software Resource

Transcriptome-Integrated Genetic Association Resource

https://github.com/yanglab-emory/TIGAR

- Implement both Elastic-Net and DPR models for training GReX imputation models
- Integrate training GReX imputation model, GReX prediction, TWAS in the same tool
- TWAS based on Burden test and SKAT
- TWAS with both individual-level and summary-level GWAS data
- TWAS with multiple phenotypes
- Multi-thread computation
- Load VCF/Dosage genotype input files

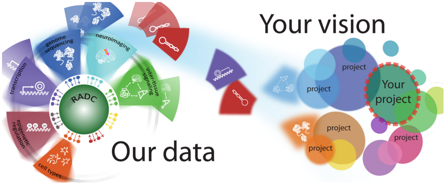
Acknowledgement

Yang Lab

github.com/yanglab-emory

Rush Alzheimer's Disease Center

www.radc.rush.edu



EMORY UNIVERSITY SCHOOL OF MEDICINE

Department of Human Genetics