Bayesian Genome-wide TWAS method integrating both cis- and trans- eQTL with GWAS summary statistics

Jingjing Yang, PhD

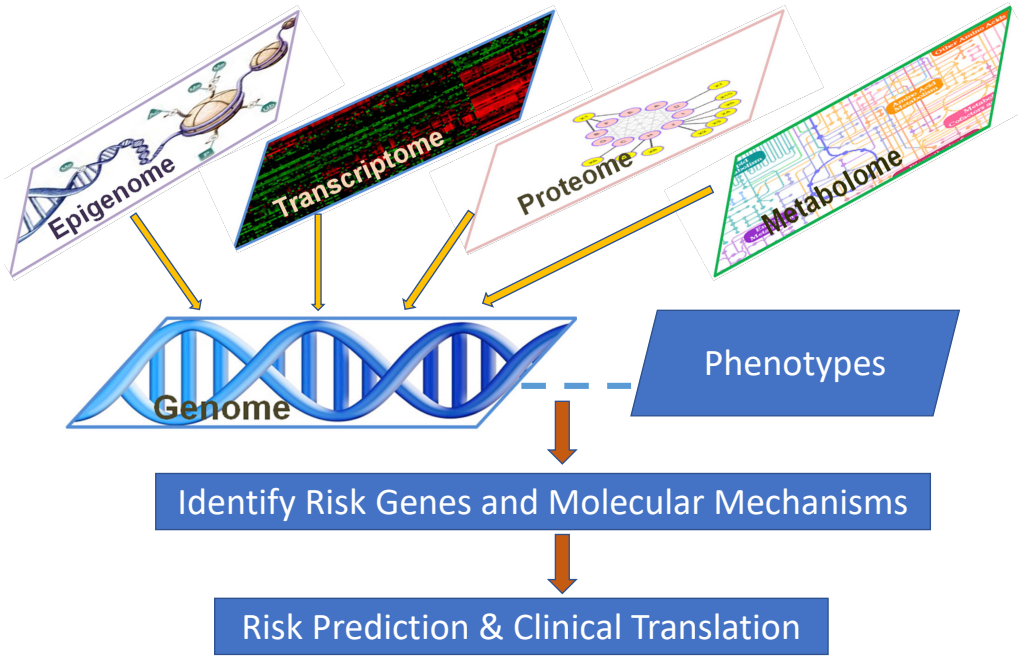
EMORY
UNIVERSITY
SCHOOL OF
MEDICINE

CCQG
CENTER FOR COMPUTATIONAL
AND QUANTITATIVE GENETICS

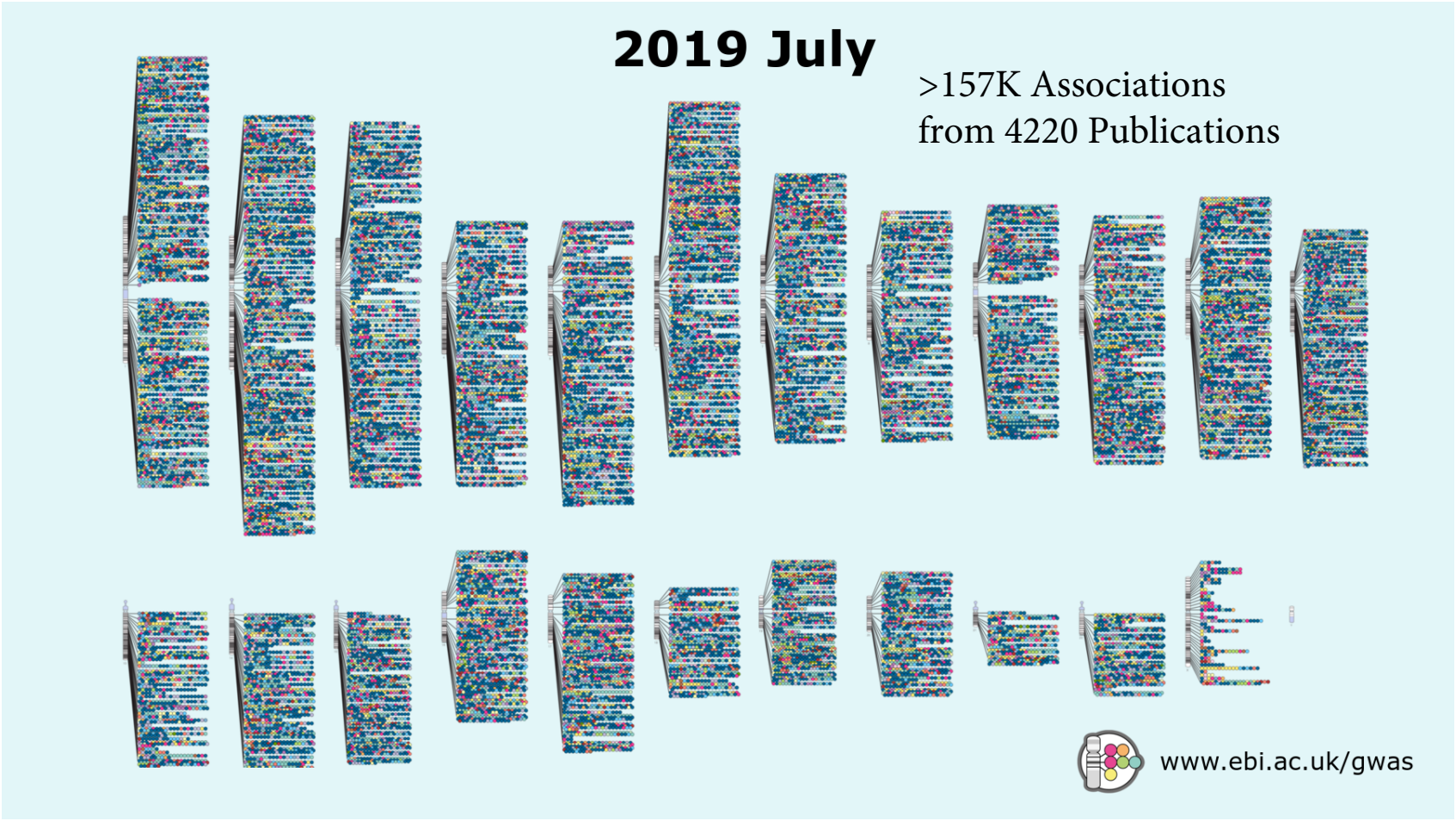
Outline

Genetic Etiology of Complex Diseases

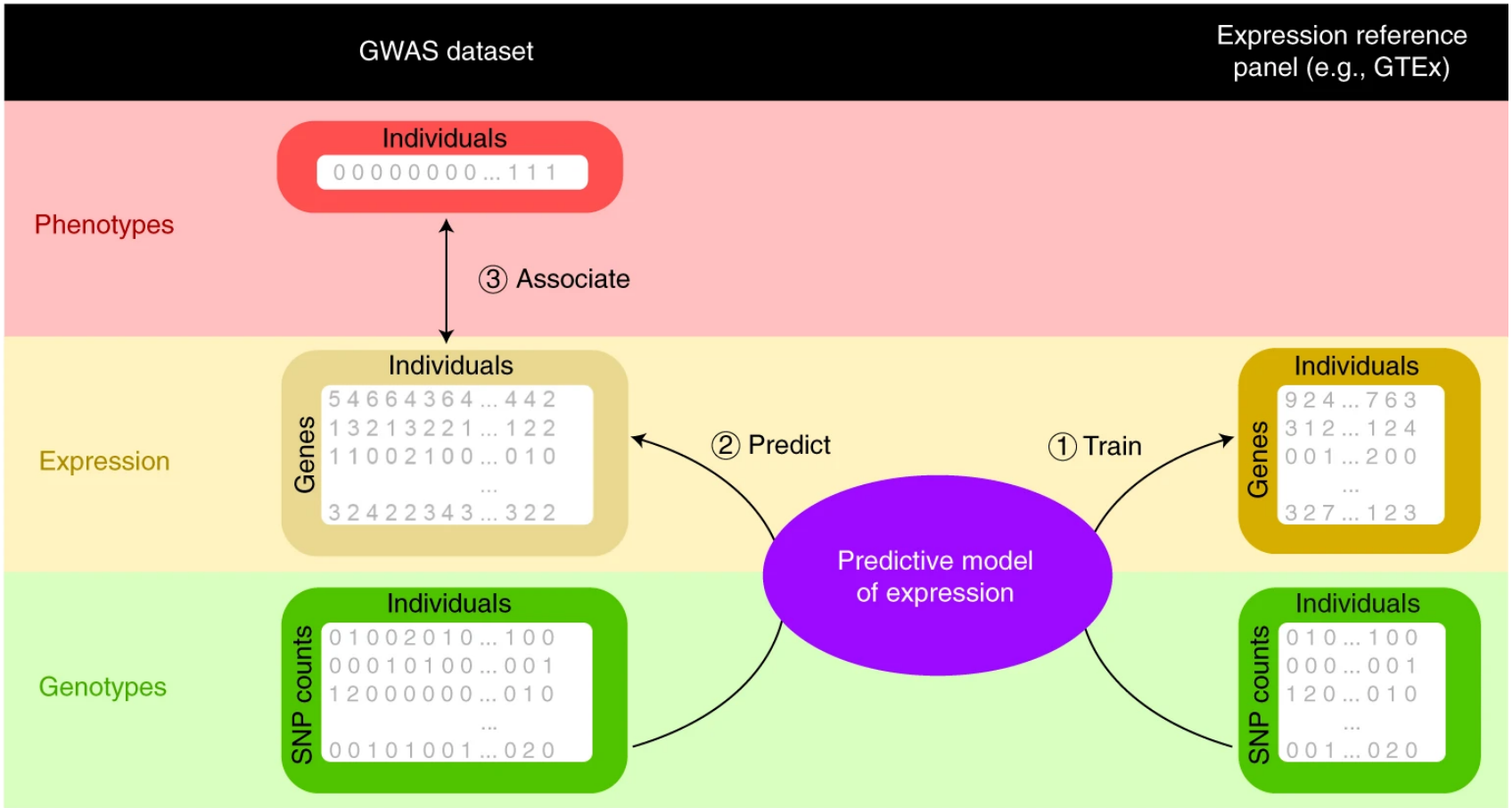
- Polygenic with low penetrance by individual genes
- Composed of multiple omics layers
- Biological mechanisms are largely unknown



Genome-wide Association Study (GWAS) Findings



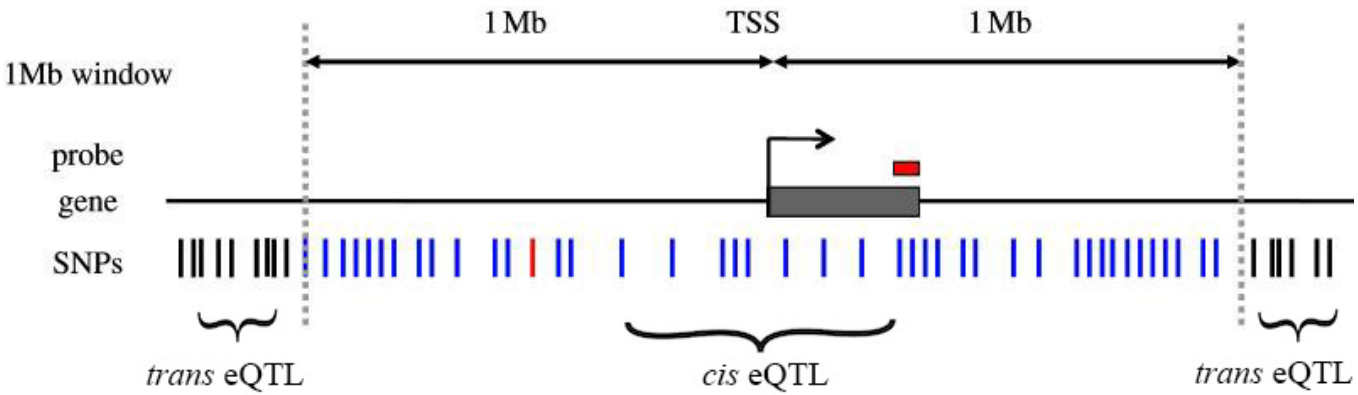
Transcriptome-wide Association Study (TWAS)



[Wainberg M. et. al. Nat. Genetics. 2019.]

Existing TWAS Tools

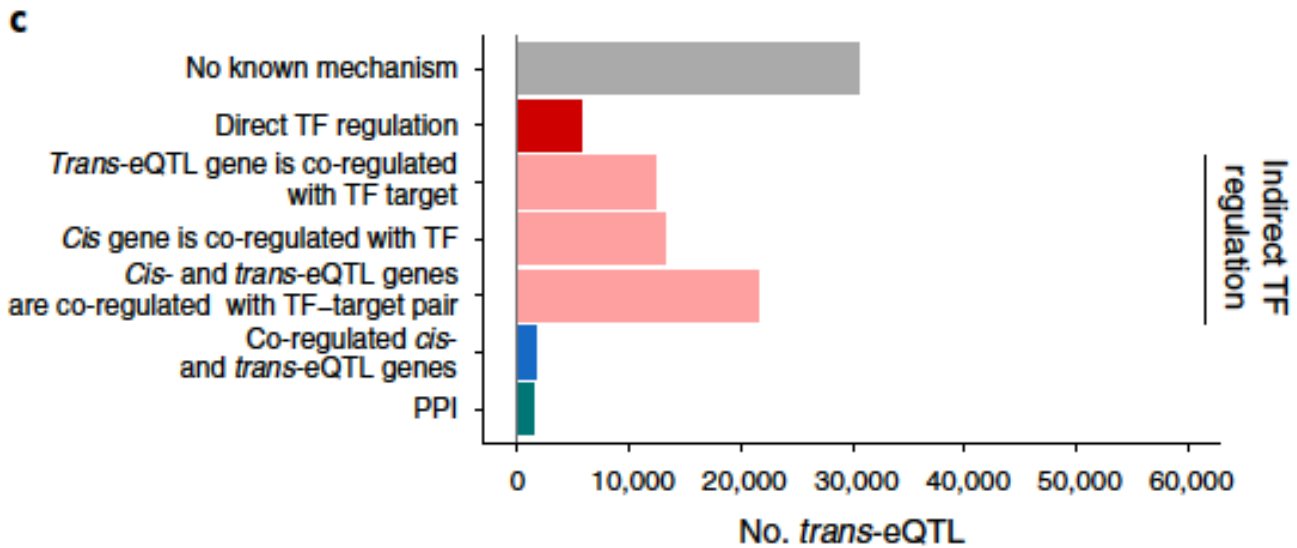
- Tools for TWAS:
  - PrediXcan. *[Gamazon et al., Nat. Genetics. 2015]*
  - FUSION. *[Gusev et al., Nat. Genetics. 2016]*
  - TIGAR. *[Nagpal et al., AJHG. 2019]*
- Caveat: utilize only *cis*-eQTL, defined by proximity to gene



Variants around a transcription starting site, *cis* or *trans* acting. *[Nica & Dermitzakis, Philos Trans R Soc Lond B Biol Sci. 2013.]*

Importance of *trans*-eQTL

- Gene expression levels are affected by both *cis* and *trans*-eQTL. *[Gusev et al., Nat. Genetics. 2016]*
- In whole blood tissue, $> 30\%$ genes have significant *trans*-eQTL. *[Lloyd-Jones et al., AJHG, 2017]*
- In eQTLGen Consortium studies of $31,684$ blood samples, *trans*-eQTL were detected for 37% of $10,317$ trait-associated GWAS signals, which primarily working through regulations by transcription factors. *[Vosa U. et al., Nat. Genetics. 2021]*



Bayesian Genome-Wide TWAS (BGW-TWAS)

Bayesian Variable Selection Regression (BVSR) Model

1. Consider quantitative gene expression trait $\boldsymbol{T}_g$ and "spike-and-slab" priors for “eQTL” effect size w_i

$$\boldsymbol{T}_g = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$$
$$w_i \sim \pi_q N(0, \sigma_\varepsilon^2 \sigma_q^2) + (1 - \pi_q)\delta_0(w_i)$$
$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

2. Consider an indicator variable γ_i per SNP i, cis or $trans$ as denoted by q

$$\gamma_i \sim Bernoulli(\pi_q) \text{ such that } w_i \sim \begin{cases} N(0, \sigma_\varepsilon^2 \sigma_q^2) & \text{if } \gamma_i = 1 \\ 0 & \text{if } \gamma_i = 0 \end{cases}$$

Allow respective "spike-and-slab" prior for effect sizes of cis and $trans$ “eQTL”.

Bayesian Genome-Wide TWAS (BGW-TWAS)

3. Estimate "eQTL" effect size $\widehat{w}_i$ and *Posterior Causal Probability* (PP), $\hat{\gamma}_i = E[\gamma_i] = Prob(\gamma_i = 1)$, by MCMC.

4. With GWAS data of additional test samples, predict Genetically Regulated gene eXpression ($GReX$) by

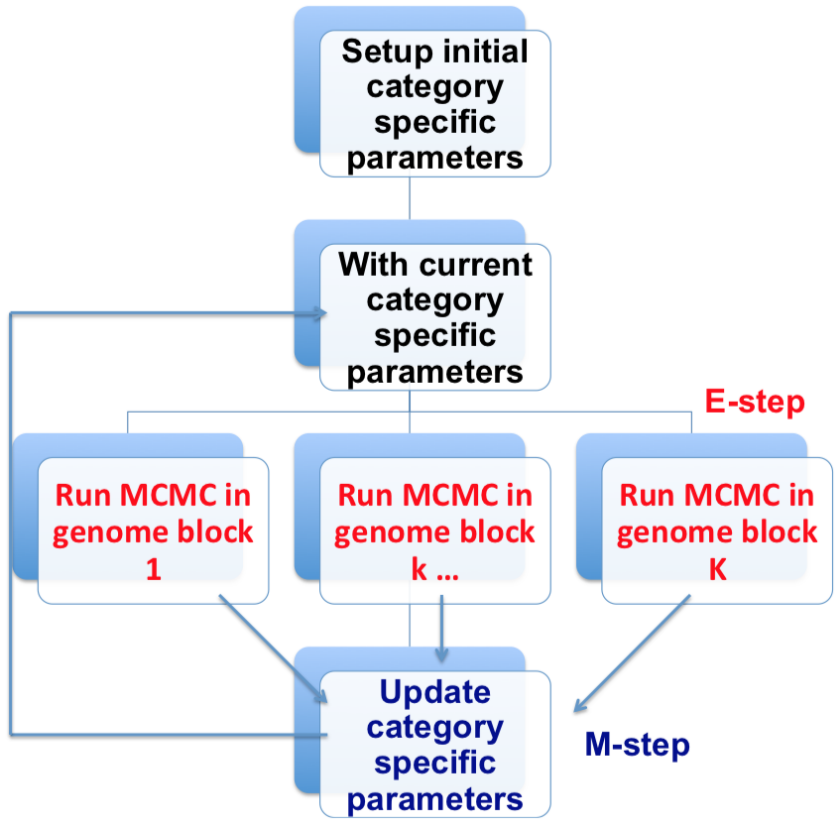
$$\widehat{GReX}_g = \sum_{i=1}^{m} \widehat{\gamma}_i \widehat{w}_i x_i^*$$

$$E[g(\mathbf{Y}_{pheno}|\mathbf{X}, \widehat{\mathbf{w}}, \widehat{\gamma})] = \beta \widehat{GReX}_g = \beta \left(\sum_{i=1}^{m} \widehat{\gamma}_i \widehat{w}_i x_i^* \right)$$

5. TWAS is to test $H_0 : \beta = 0$

Estimate $\mathbf{w}$ and $\mathbf{E}[\gamma]$

1. Employ EM-MCMC algorithm *[Yang et al., AJHG 2017]*
2. Use pre-calculated summary statistics from single variant model, $T_g = \boldsymbol{x}_i w_i + \boldsymbol{\varepsilon}$
3. Pre-calculate LD correlation coefficients
4. Parallelize over segmented genome blocks



[*Yang et al., AJHG 2017*]

Segment and Prune Genome Blocks

- Genome-wide SNPs segmented into blocks with $\sim 3,000$ - $10,000$ variants based on block-wise LD structure
- Prune to genome blocks that:
  - have variants in cis
  - have potential marginally significant (p-value $< 10^{-5}$) variant by single variant tests
  - up to 50 blocks, ranked by top significant p-values by single variant tests

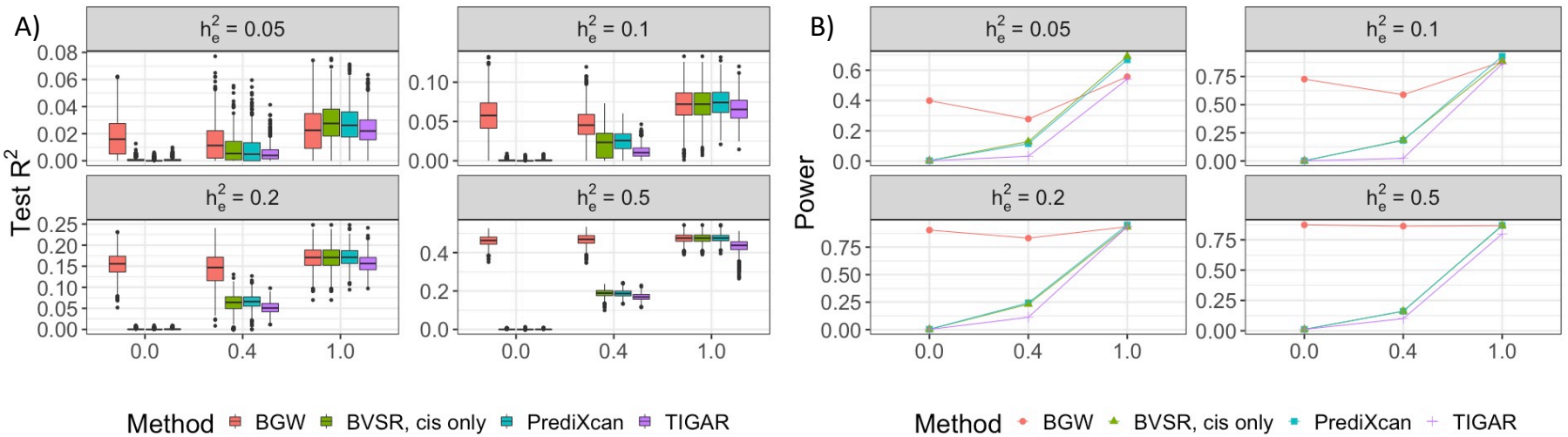
Simulation Study Design

- Use real genotype data of $22,641$ variants - $1,269$ *cis* and $21,372$ *trans* of $1,708$ samples
- Simulate quantitative gene expression traits from selected true causal eQTL
- Apply **BGW** (BVSR), **PrediXcan** (Elastic-Net), and **TIGAR** (non-parametric Bayesian Dirichlet process regression) to train gene expression prediction models with 499 training samples
- Predict $GReX$ values and conduct TWAS tests using $1,209$ test samples

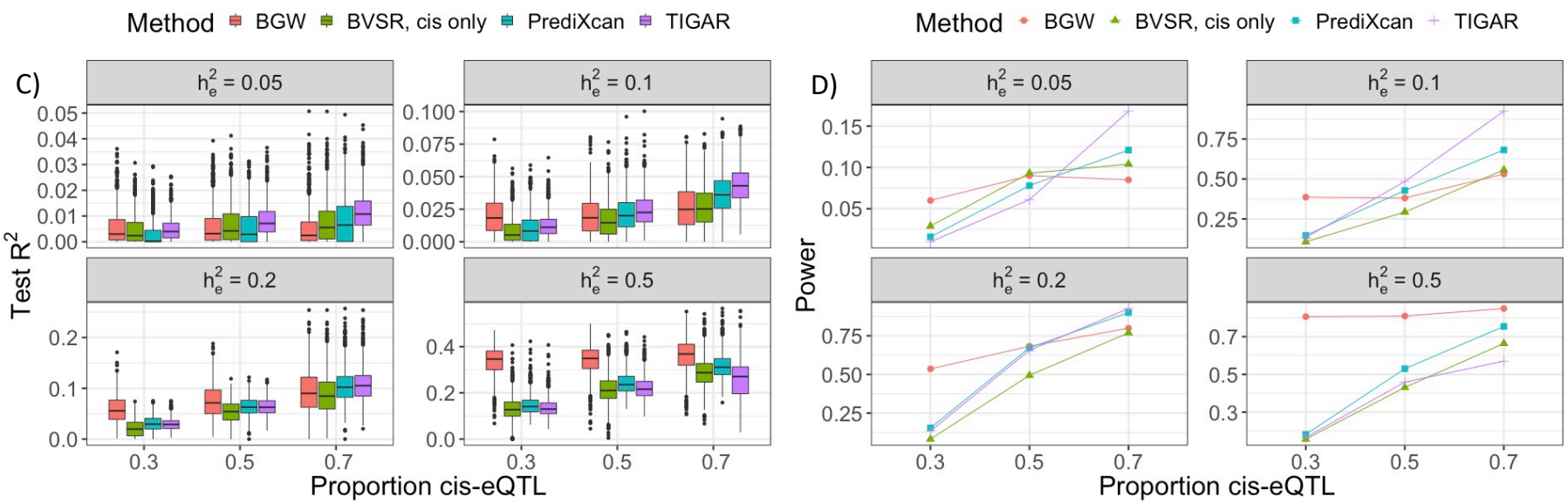
Simulation Study Design

- Consider the following scenarios:
  - 5 true causal eQTL and various proportions of cis variants, $(0\%, 40\%, 100\%)$
  - 22 true causal eQTL and various proportions of cis variants, $(30\%, 50\%, 70\%)$
  - Various heritability for quantitative gene expression traits $h_e^2 = (0.05, 0.1, 0.2, 0.5)$
- Repeat simulation for $1,000$ times to compare both prediction R^2 and TWAS power

With 5 True Causal eQTL



With 22 True Causal eQTL



Sum of $\widehat{\gamma}_i$

Simulation scenarios with 2/5 and 11/22 true *cis*-eQTL:

Gene Expression Heritability		Sum of Posterior Probabilities		
		Whole Genome	Cis-Region	Trans-Region
5 True Causal eQTL	0.05	0.79	0.46	0.33
	0.1	2.28	1.13	1.15
	0.2	3.72	1.44	2.28
	0.5	4.91	1.56	3.35
22 True Causal eQTL	0.05	0.05	0.02	0.03
	0.1	0.21	0.11	0.10
	0.2	1.43	0.87	0.56
	0.5	6.46	3.89	2.57

Application Studies of Alzheimer's Dementia (AD)

ROS/MAP

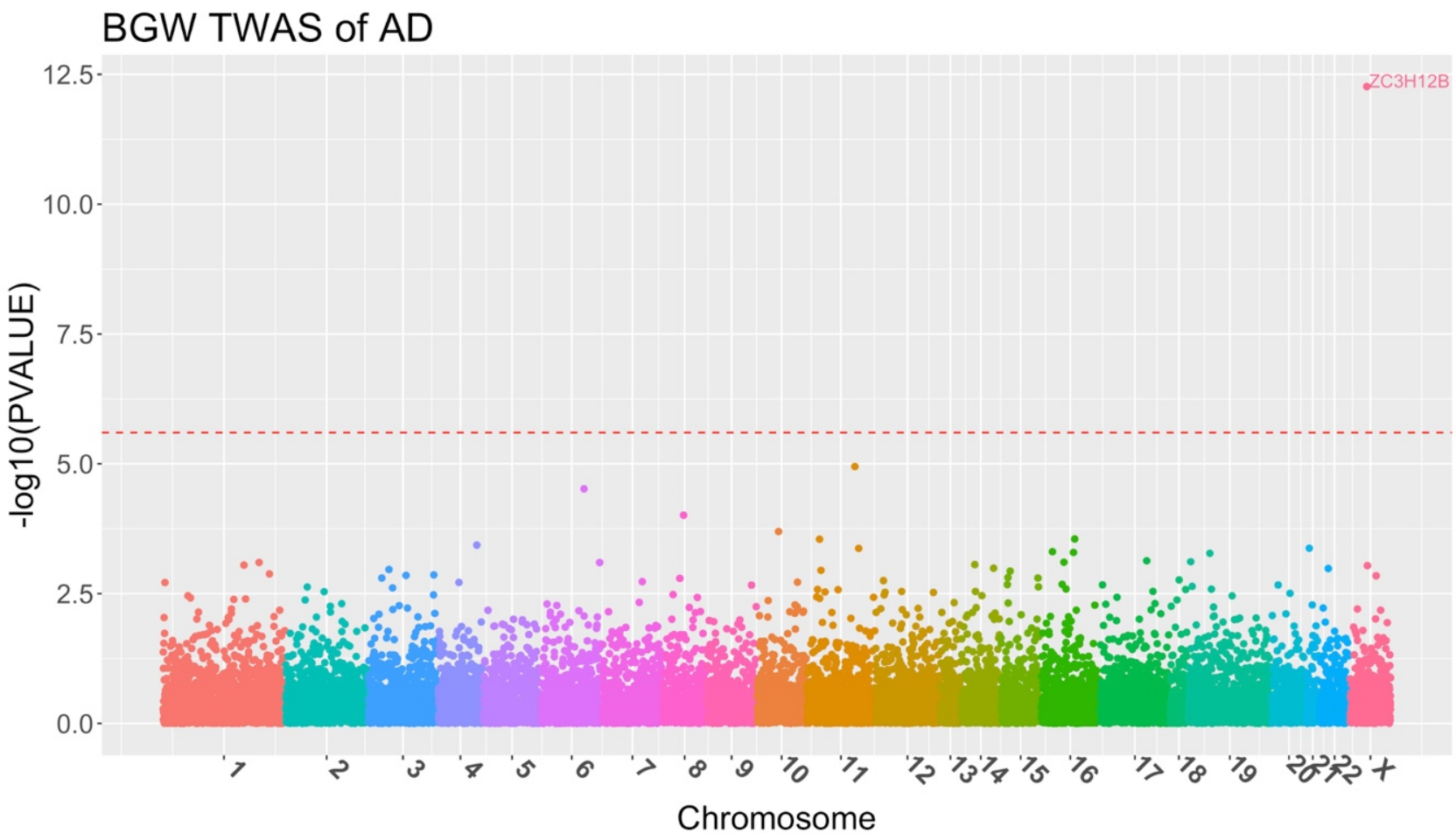
- Training data: 499 subjects with both genotype and transcriptomic data ($14,156$ genes)
- Test GWAS data of $2,093$ individuals
- Considered phenotypes: AD clinical diagnosis, β-Amyloid, Tangles, Global AD pathology
- TWAS adjusted for covariates: Age at death, Sex, Smoking, ROS or MAP study, Education level, Top 3 genotype PCs

Mayo Clinic LOAD GWAS Data

- GWAS data of $2,099$ individuals
- Considered phenotypes of AD clinical diagnosis
- TWAS adjusted for covariates: Age, Sex, Top 3 genotype PCs

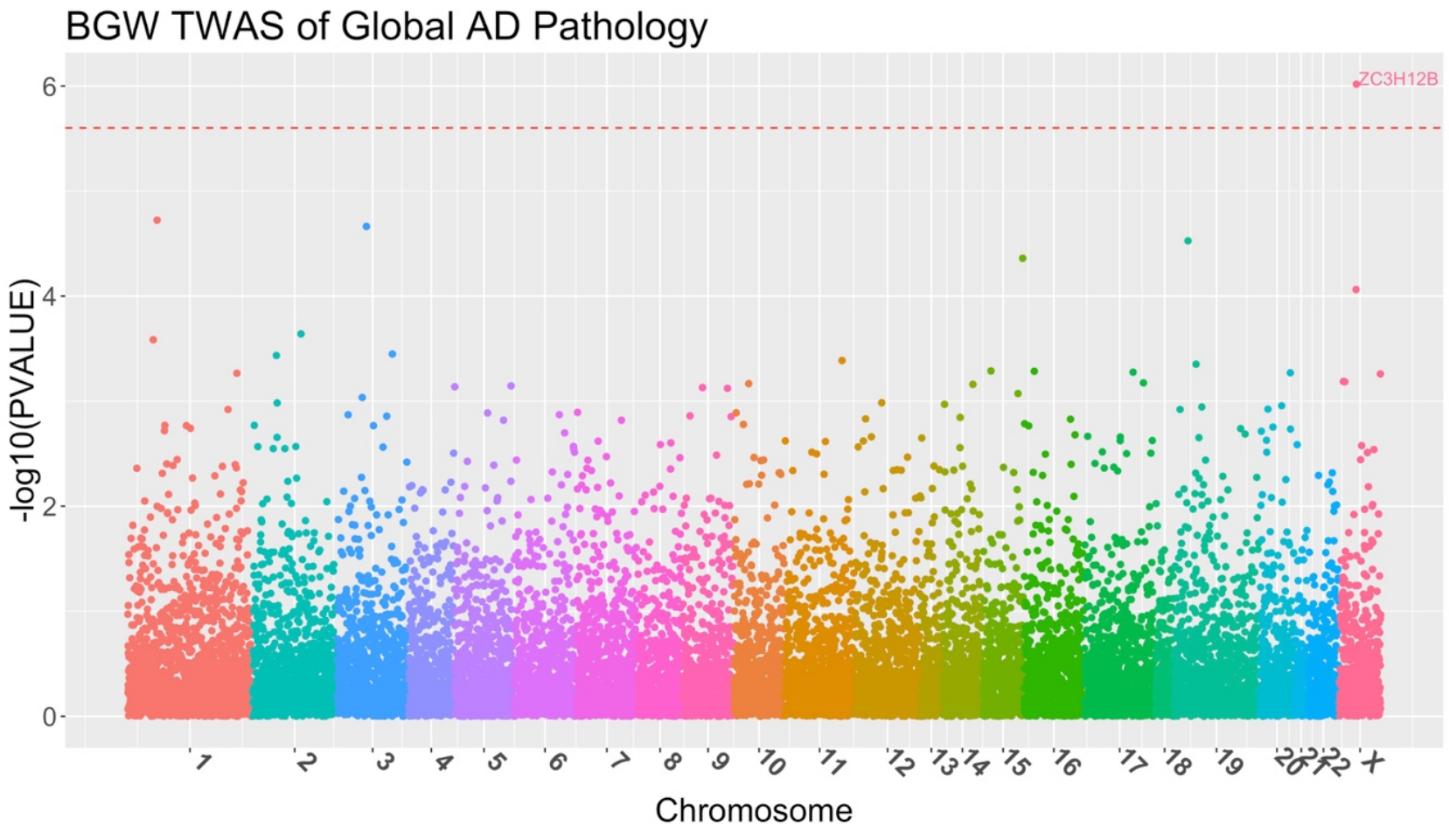
BGW TWAS of AD Clinical Diagnosis

A)



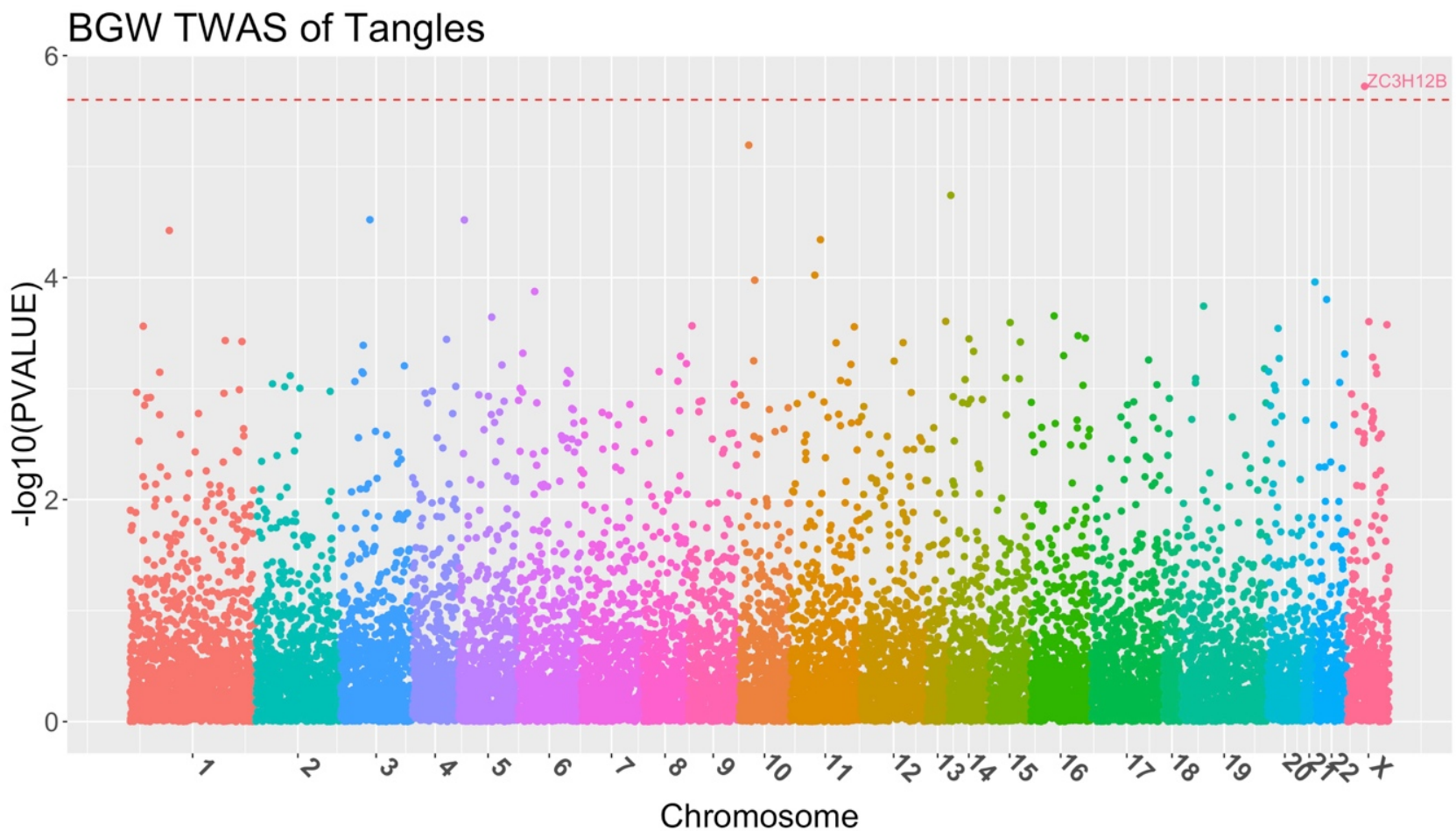
BGW TWAS of Global Pathology

B)

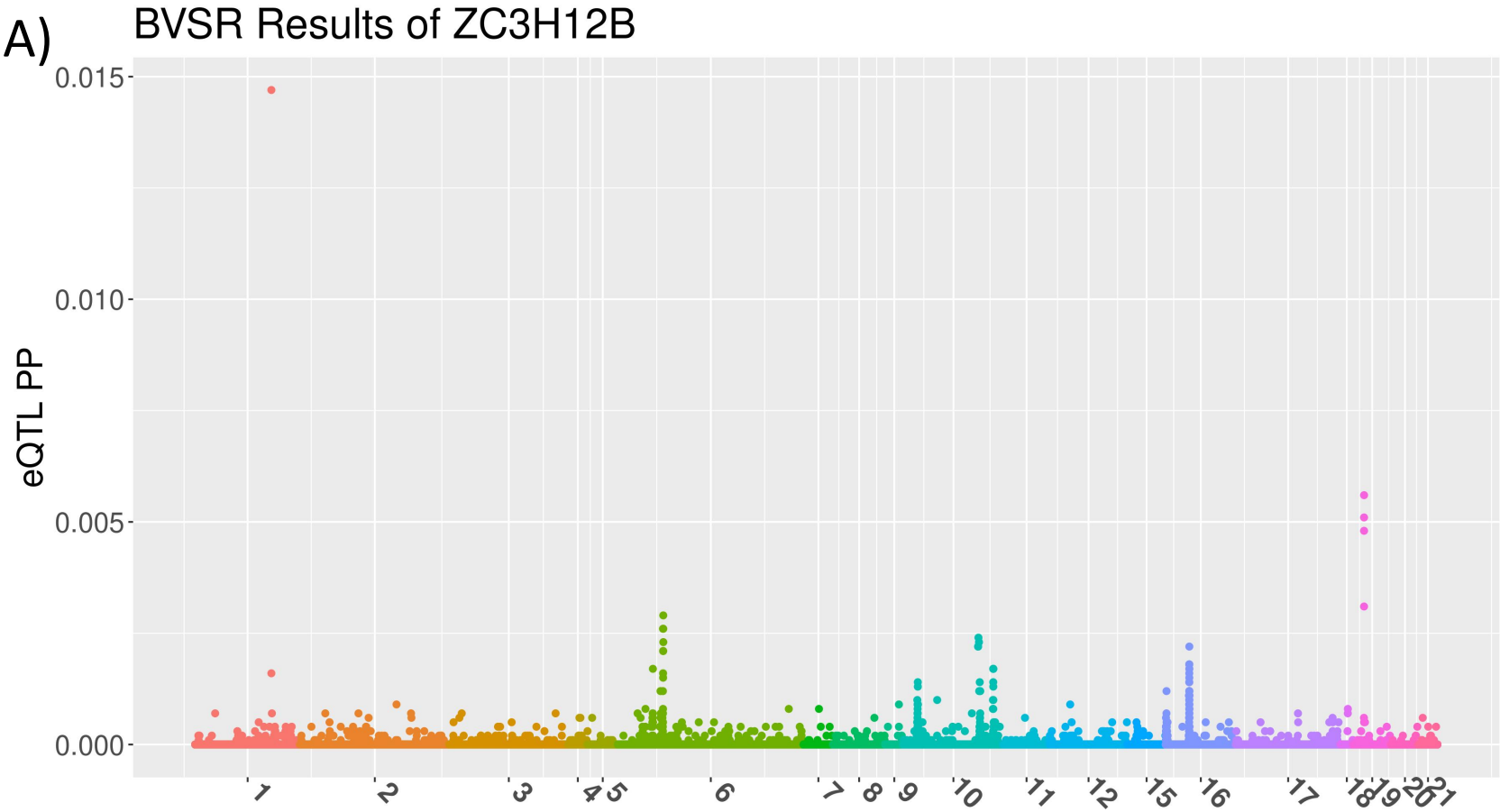


BGW TWAS of Tangles

A)



BVSR Results for Gene *ZC3H12B*



Top Five *trans*-eQTL for Gene *ZC3H12B*

Table 2. Trans-eQTL with top five PP > 0.003 for gene ZC3H12B.

CHR	POS	rsID	Function	MAF	PP	w	p-value
1	159,135,282	rs3026946	Intergenic	0.213	0.0147	-0.071	6.25×10^{-7}
19	45,422,160	rs12721051	3' UTR (APOC1)	0.161	0.0031	0.071	3.94×10^{-6}
19	45,422,846	rs56131196	Downstream (APOC1)	0.173	0.0048	0.069	1.75×10^{-6}
19	45,422,946	rs4420638	Downstream (APOC1)	0.173	0.0051	0.068	1.77×10^{-6}
19	45,424,514	rs157592	Regulatory Region (APOC1)	0.181	0.0056	0.075	1.43×10^{-6}

- *rs12721051* was identified as a GWAS signal of total cholesterol levels
- *rs4420638* is in LD with the *APOE E4* allele (*rs429358*) and was identified to be a GWAS signal of blood lipids
- *rs56131196* and *rs157592* were identified as GWAS signals of AD and independent of *APOE E4*

Sum of $\widehat{\gamma}_i$ in real ROSMAP studies.

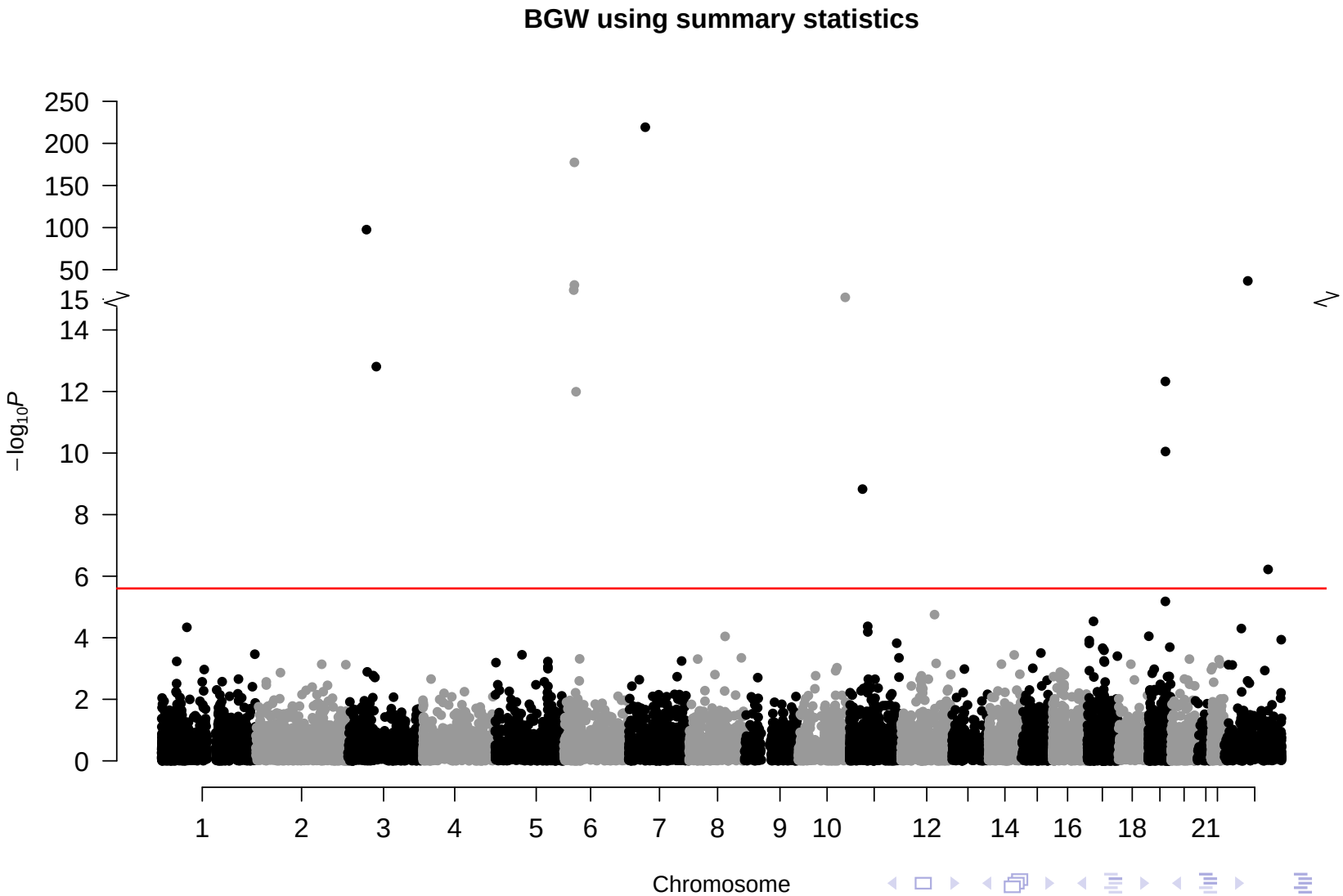
Train R^2	Sum of Posterior Inclusion Probabilities			Number of Genes
	Whole Genome	Cis- Region	Trans-Region	
(0, 0.05)	6.63	0.60	6.23	1,504
(0.05, 0.1)	1.45	0.13	1.32	1,964
(0.1, 0.25)	2.00	0.17	1.83	6,617
(0.25, 0.5)	2.66	0.22	2.44	3,224
(0.5, 1)	3.04	0.31	2.73	474

TWAS using IGAP summary-level GWAS data of AD

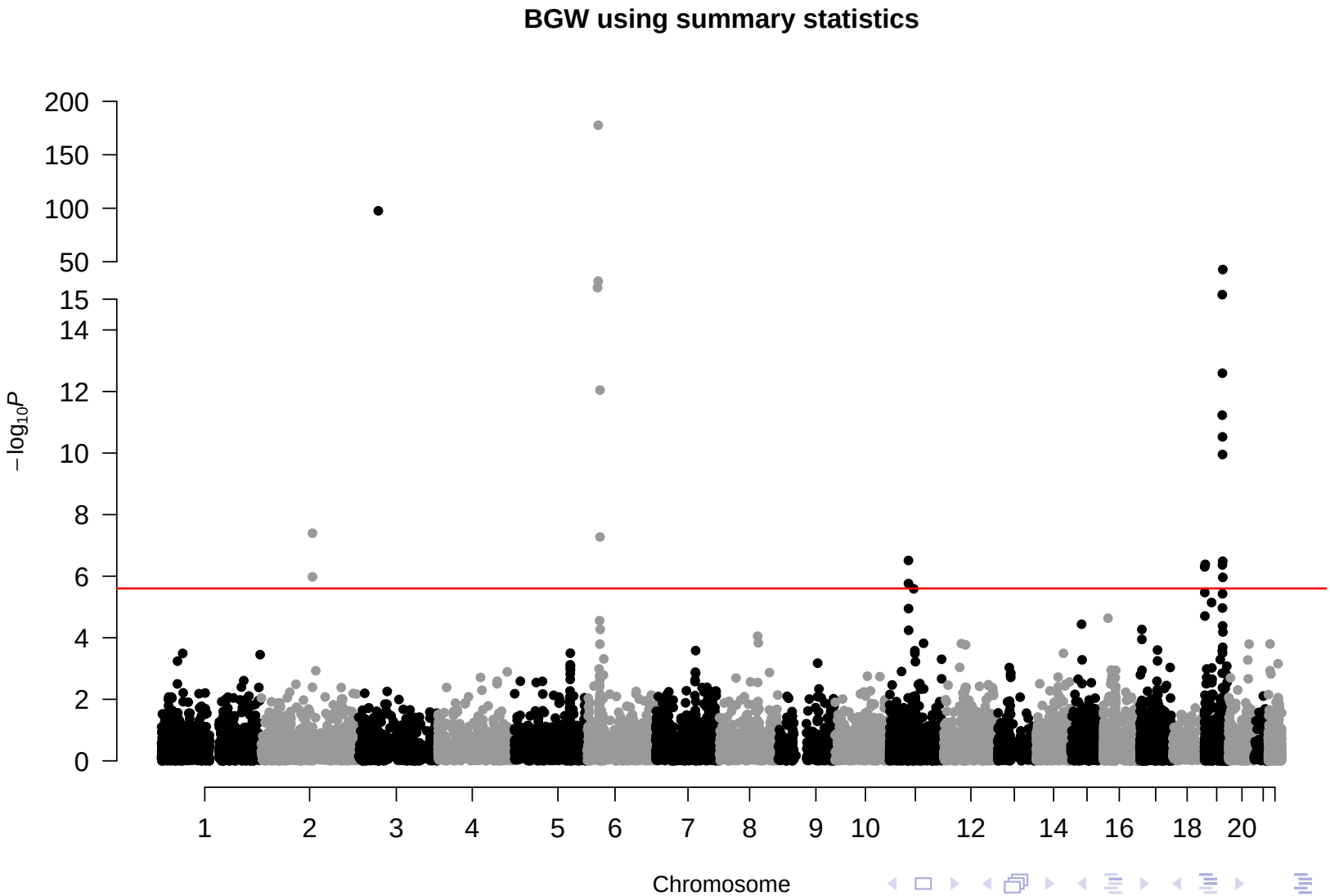
GWAS summary statistics for studying AD by International Genomics of Alzheimer's Project (IGAP):

- Generated by meta-analysis of four consortia ($\sim$ 17K cases and $\sim$ 37K controls; European)
  - Alzheimer's Disease Genetic Consortium (ADGC)
  - Cohorts for Heart and Ageing Research in Genomic Epidemiology (CHARGE) Consortium
  - European Alzheimer's Disease Initiative (EADI)
  - Genetic and Environmental Risk in Alzheimer's Disease (GERAD) Consortium
- Use S-PrediXcan burden test statistic, with variant weights derived by **BGW**, **PrediXcan**, and **TIGAR**.

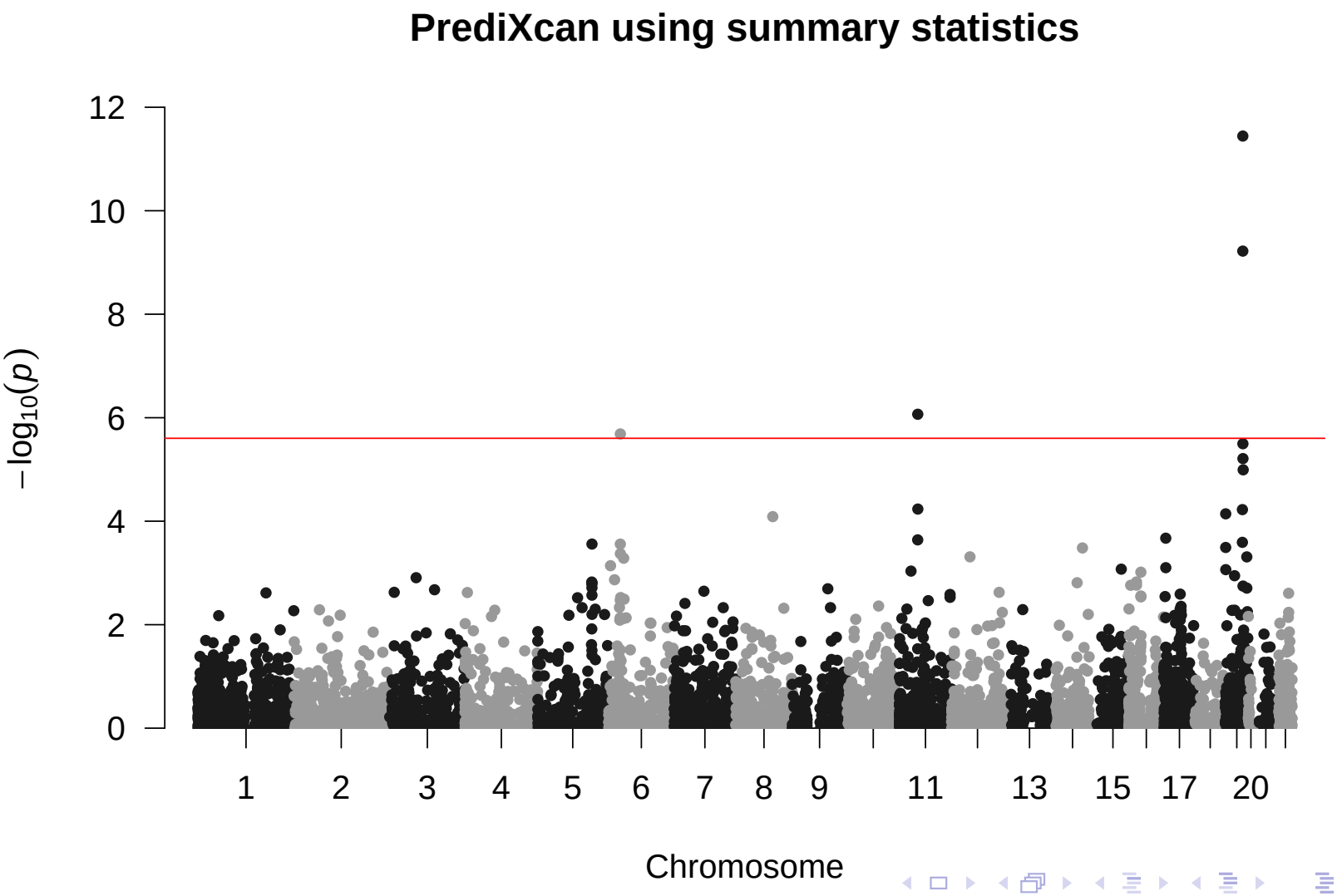
BGW-TWAS considering both *cis*- and *trans*-eQTL



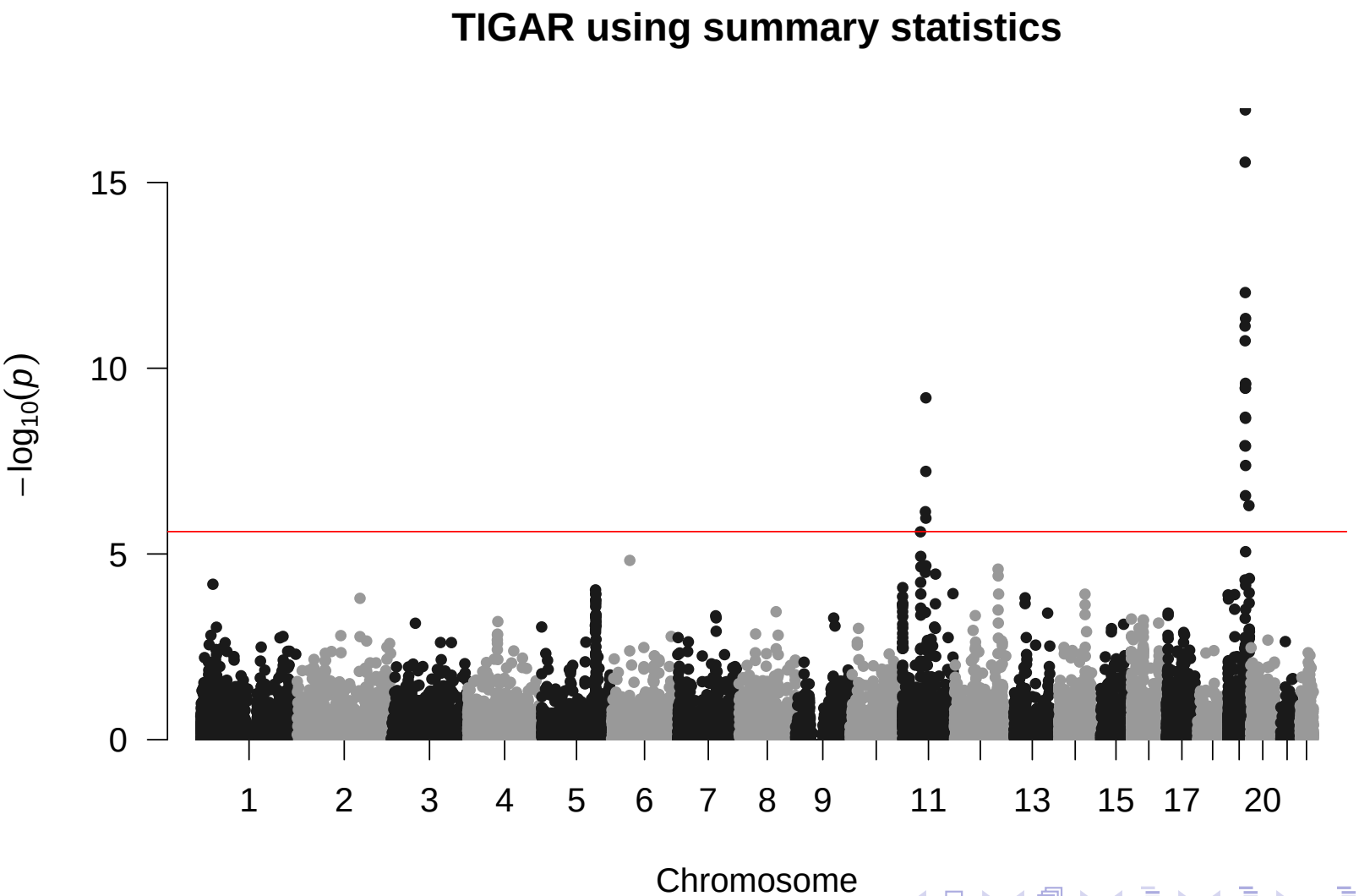
BGW-TWAS considering only *cis*-eQTL



PrediXcan considering only *cis*-eQTL



TIGAR considering only *cis*-eQTL



Significant TWAS genes by BGW-TWAS

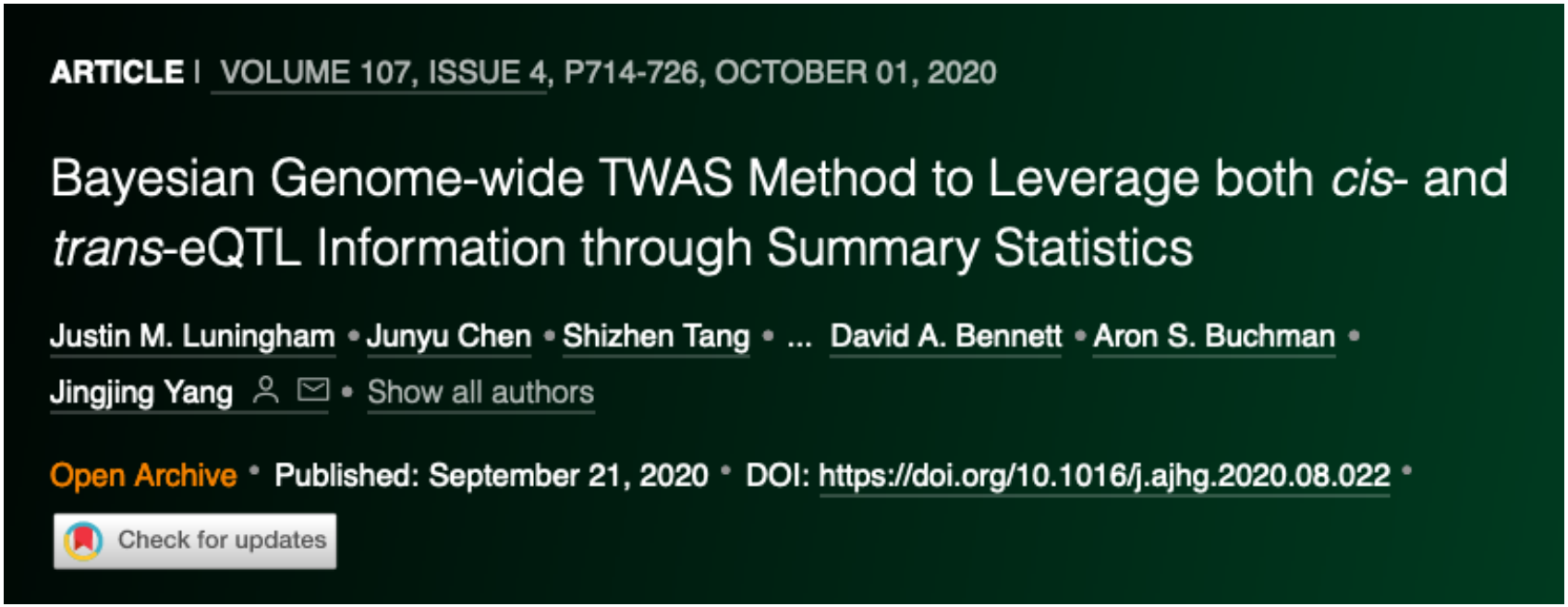
Gene	CHR	Position	TWAS P-VALUE			
			BGW-TWAS	BVSR cis-eQTL	PrediXcan	TIGAR
GPX1[a]	3	49,394,608	2.45×10^{-98}	2.45×10^{-98}	-	3.15×10^{-1}
FAM86DP	3	75,484,261	1.55×10^{-13}	4.81×10^{-1}	5.38×10^{-1}	9.63×10^{-1}
BTN3A2[a]	6	26,378,546	1.59×10^{-26}	1.56×10^{-26}	3.17×10^{-1}	5.04×10^{-1}
ZNF192[a]	6	28,124,089	1.26×10^{-32}	1.25×10^{-32}	8.56×10^{-2}	2.07×10^{-1}
AL022393.7[a]	6	28,144,452	3.25×10^{-178}	2.24×10^{-178}	1.50×10^{-1}	8.36×10^{-2}
HLA-DRB1[ab]	**6**	**32,557,625**	$\mathbf{1.02 \times 10^{-12}}$	$\mathbf{8.99 \times 10^{-13}}$	$\mathbf{2.06 \times 10^{-6}}$	-
AEBP1	7	44,154,161	5.55×10^{-220}	8.62×10^{-1}	6.69×10^{-1}	4.19×10^{-1}
BUB3	10	124,924,886	6.64×10^{-18}	1.05×10^{-2}	-	4.76×10^{-1}
FBXO3	11	33,796,089	1.48×10^{-9}	6.88×10^{-1}	-	1.13×10^{-1}
CEACAM19[abc]	**19**	**45,187,631**	$\mathbf{4.7 \times 10^{-13}}$	$\mathbf{2.54 \times 10^{-13}}$	$\mathbf{3.60 \times 10^{-12}}$	$\mathbf{2.83 \times 10^{-16}}$
APOC1[a]	19	45,422,606	8.9×10^{-11}	1.11×10^{-10}	3.18×10^{-6}	7.2×10^{-3}
ZC3H12B	**X**	**64,727,767**	$\mathbf{2.08 \times 10^{-37}}$	-	-	-
CXorf56	X	118,699,397	6.02×10^{-07}	-	-	-

a. Genes that were also identified as significant by using BVSR cis-eQTL estimates.
b. Genes that were also identified by PrediXcan.
c. Genes that were also identified by TIGAR.

Summary

- Propose a novel **BGW-TWAS** tool for leveraging both *cis*- and *trans*-eQTL in TWAS
- Computationally manageable with a computation cost of $\sim$10 minutes per gene
- Gain power when there are true *trans*-eQTL signals
- Identified that the genetic effects of known GWAS signals (*rs4420638, rs56131196, rs157592, near APOE E4 on Chr 19*) could be mediated through the gene expression levels of *ZC3H12B on Chr X* which is significant for both AD and AD pathology Tangles

Publication



BGW-TWAS Software:
https://github.com/yanglab-emory/BGW-TWAS.git

Acknowledgement

EMORY UNIVERSITY

CCQG CENTER FOR COMPUTATIONAL AND QUANTITATIVE GENETICS

Yang Lab @ Emory

RUSH UNIVERSITY

Rush Alzheimer's Disease Center www.radc.rush.edu

NIH National Institute of General Medical Sciences

Mayo Clinic LOAD GWAS

NIH National Institute on Aging

AMP-AD Knowledge Portal