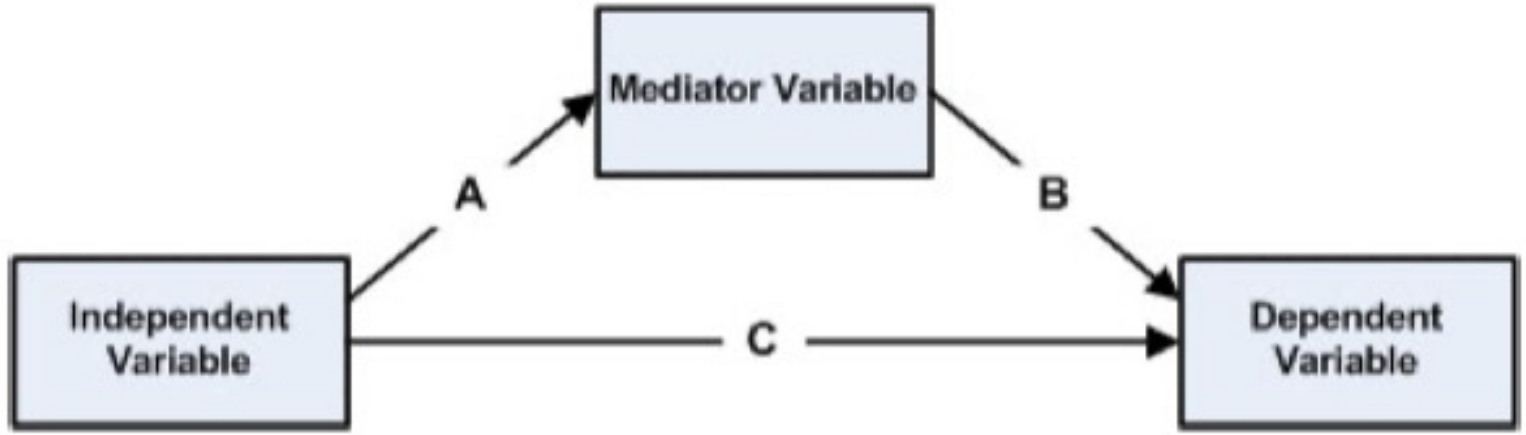
Transcriptome-wide Association Studies
Lecture 3

Outline

- Connection between TWAS and Mendelian Randomization
- PMR-Egger
- Apply TWAS Framework to Integrate Proteomics and Epigenetics Data with GWAS data

Mediation Analysis

Mediation Analysis seeks to identify the mechanism that the Independent Variable (Instruments) affects the Dependent Variable (Response) through the Mediator Variable



Mendelian Randomization (MR)

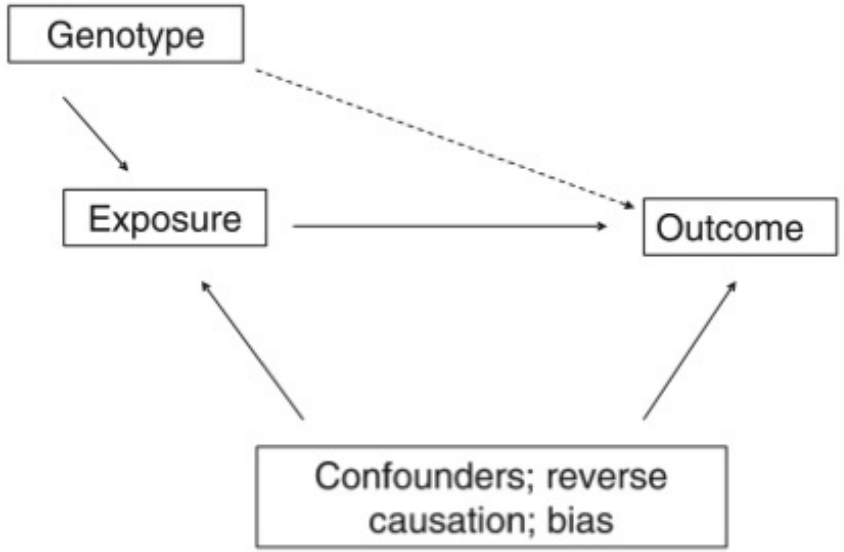
Mendelian Randomization uses genetic variants as instruments for Mediation Analysis

- Under the Mendel's Second Law, genotypes are assigned randomly when passed from parents to offspring during meiosis (independent assortment)
- Alternative of traditional randomized trials to estimate a putative causal effect of the mediator on phenotype
- Instruments are proxies of the exposure that are free of confounders

Mendelian Randomization (MR)

Assumptions for instruments,
[1] Didelez and Sheehan, 2007.

- Associated with the Exposure (Mediator)
- Not associated with any confounder of the Exposure-Outcome association
- Conditionally independent of the outcome given the Exposure and confounders



[1] Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330. PMID: 17715159.

MR Model

Let X denote genotype, M denote mediator, Y denote outcome

- Genetic model for M: $M = \beta_0 + \beta_{XM}X + \varepsilon$
- Genetic model for Y: $Y = \widetilde{\beta}_0 + \beta_{XY}X + \varepsilon$
- Joint model for Y

$$Y = \beta_0 + \beta_{direct}X + \beta_{causal}M + \varepsilon$$

$$Y = \widetilde{\beta}_0 + \beta_{direct}X + \beta_{causal}\beta_{XM}X + \varepsilon$$

$$Y = \widetilde{\beta}_0 + (\beta_{direct} + \beta_{causal}\beta_{XM})X + \varepsilon$$

- If $\beta_{direct} = 0$, $\beta_{XY} = \beta_{causal}\beta_{XM}$, equivalently $\beta_{causal} = \beta_{XY}/\beta_{XM}$
- If $\beta_{direct} \neq 0$, $\beta_{XY} = \beta_{direct} + \beta_{causal}\beta_{XM}$

Goal: test if β_{causal} significantly different from 0.

Two-Stage TWAS & MR

- Common Features
  - Uses both eQTL (transcriptomic) and GWAS data that might be profiled for two independent cohorts
  - Test association/causality between multiple-SNPs-per-Gene and phenotype of interest
  - Two-stage TWAS is “equivalent” to a Two-Stage MR inference procedure, which fails to account for the uncertainty of estimating eQTL weights and separate horizonal pleiotropy

- Different Features
  - MR methods such as Inverse-Variance-Weighted (IVW) Regression, MR-Egger, SMR, GSMR uses Single SNP Instrument or Multiple Independent SNP Instruments
  - Stage I in TWAS (e.g., PrediXcan and TIGAR) models LD among all SNPs with non-zero eQTL weights per gene by a multiple linear regression model. All SNPs with non-zero eQTL weights are taken as Instrument Variables.

MR-Egger Regression (Bowden et al., 2015)

- Weighted linear regression of the SNP-outcome coefficients on the SNP-exposure coefficients
- MR-Egger slope estimate will equal to the Inverse-Variance Weighted (IVW) Two-Sample MR estimate, with 0 intercept term

Let X denote genotype, M denote mediator, Y denote outcome

- Genetic model for M: $M = \beta_0 + \beta_{XM}X + \varepsilon$
- Genetic model for Y: $Y = \widetilde{\beta}_0 + \beta_{XY}X + \varepsilon$
- Joint model for Y

$$Y = \beta_0 + \beta_{direct}X + \beta_{causal}M + \varepsilon$$

$$Y = \widetilde{\beta}_0 + \beta_{direct}X + \beta_{causal}\beta_{XM}X + \varepsilon$$

$$Y = \widetilde{\beta}_0 + (\beta_{direct} + \beta_{causal}\beta_{XM})X + \varepsilon$$

- If $\beta_{direct} = 0$, $\beta_{XY} = \beta_{causal}\beta_{XM}$, equivalently $\beta_{causal} = \beta_{XY}/\beta_{XM}$
- If $\beta_{direct} \neq 0$, $\beta_{XY} = \beta_{direct} + \beta_{causal}\beta_{XM}$

Goal: test if β_{causal} significantly different from 0.

Weaker assumption IV3 (exclusion restriction assumption), Bowden et al., 2015

- Direct causal effect $SNP_j \rightarrow Y$ (Assumption IV3) is not 0 for all IVs
- Assume Instrument Strength Independent of Direct Effect (InSIDE)
  - The distributions of $\beta_{direct,j}$ and $\beta_{causal,j}$ are independent
- <u>Egger regression:</u> Causal effect of $\mathrm{SNP}_{\mathrm{j}=1,\ldots,J} \rightarrow M \rightarrow Y$ can be estimated by the linear regression slope of $\hat{\beta}_{X_jY} \sim \hat{\beta}_{X_jM}, j = 1, \ldots, J$

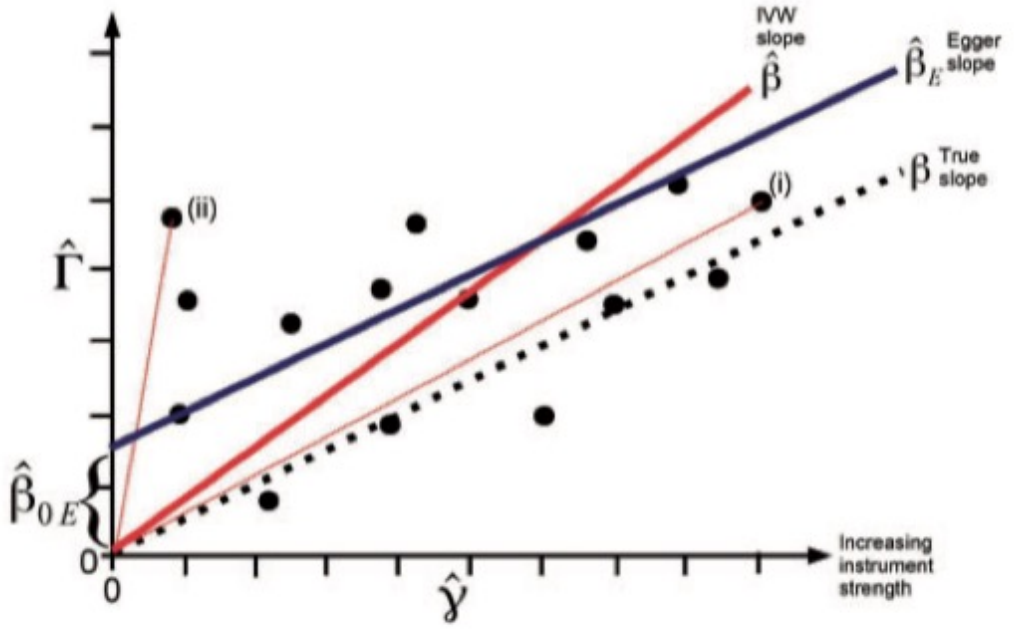


Figure 2. Plot of the gene–outcome ($\hat{\Gamma}$) vs gene–exposure ($\hat{\gamma}$) regression coefficients for a fictional Mendelian randomization analysis with 15 genetic variants. The true slope is shown by a dotted line, the inverse-variance weighted (IVW) estimate by a red line, and the MR-Egger regression estimate by a blue line. Refer to text for explanation of points (i) and (ii).

PMR-Egger : Probabilistic Two-sample Mendelian Randomization

- Consider reference cohort (E_g, X_e) and test cohort $(\widetilde{E_g}, X_y, y)$ in the following model

$$E_g = \mu_e + X_e\beta + \epsilon_e \qquad (1)$$
$$\widetilde{E_g} = \mu_e + X_y\beta + \epsilon_{\tilde{e}}. \qquad (2)$$
$$y = \mu_y + \widetilde{E_g}\alpha + X_y\gamma + \epsilon \qquad (3)$$

  - Gene expression $E_g, \widetilde{E_g}$(unobserved); genotype data X_e, X_y; phenotype data y
  - β: eQTL weights, shared by two cohorts
  - γ: horizontal pleiotropic effect, H_0: $\gamma = 0$
  - α: causal effect, H_0: $\alpha = 0$

- Replacing $\widetilde{E_g}$ in Equation (3) by Equations (2):

$$y = \widetilde{\mu_y} + X_y\beta\alpha + X_y\gamma + \epsilon_y \qquad (4)$$
$$\widetilde{\mu_y} = \mu_e\alpha + \mu_y, \qquad \epsilon_y = \epsilon_{\tilde{e}}\alpha + \epsilon$$

PMR-Egger

- Incorporate multiple correlated SNP instruments in a likelihood inference framework

$$E_g = \mu_e + X_e\beta + \epsilon_e \quad (1)$$

- Unifies many existing TWAS (e.g., PrediXcan and TIGAR) and MR methods (e.g., IVW Regression, MR-Egger, SMR, GSMR)

$$\widetilde{E_g} = \mu_e + X_y\beta + \epsilon_{\tilde{e}}. \quad (2)$$

$$y = \mu_y + \widetilde{E_g}\alpha + X_y\gamma + \epsilon \quad (3)$$

- Test for causality of multiple-SNP-per-gene -> gene expression -> phenotype

$$\mathrm{H}_0: \alpha = 0$$

Yuan Z. et al. Nature Communications, 2020.

PMR-Egger

<u>Test and control for horizontal pleiotropy</u>

- Horizontal pleiotropy:
  - SNP affects the outcome (test phenotype) through pathways other than or in addition to the exposure variable (target test gene expression)
- H_0: $\gamma = 0$

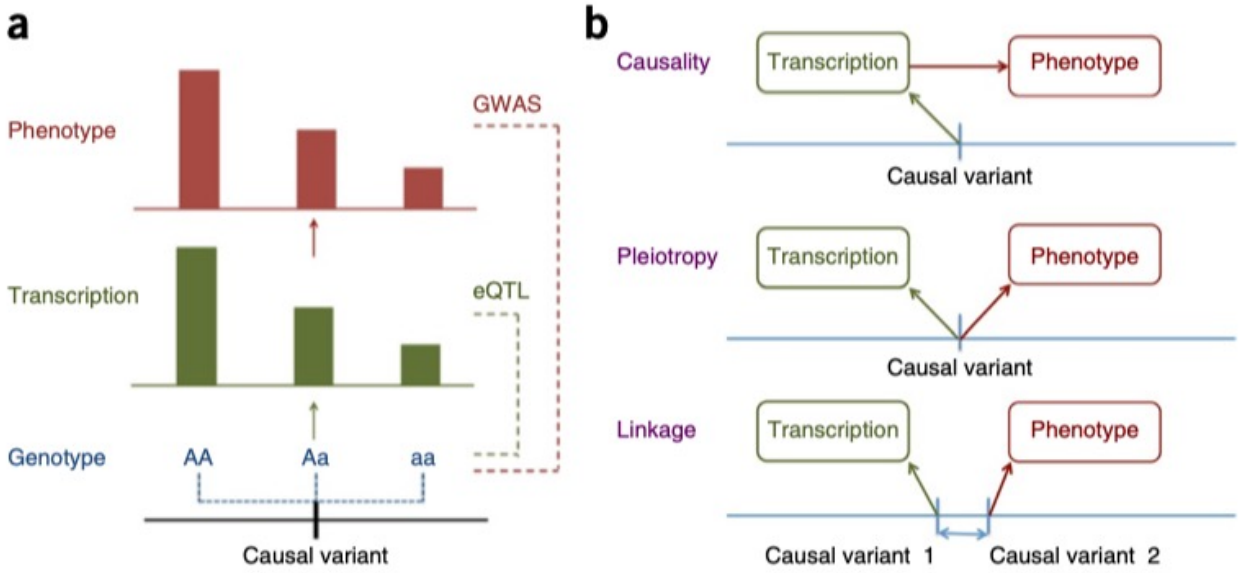


Figure 1 Association between gene expression and phenotype through genotypes. (**a**) A model of causality where a difference in phenotype is caused by a difference in genotype mediated by gene expression (transcription). (**b**) Three possible explanations for an observed association between a trait and gene expression through genotypes.

Zhu et. al., Nature Genetics, 2016.

PMR-Egger : Inference

- Maximum Likelihood Inference Framework
- EM algorithm is used
  - MLE of γ , α are obtained from the joint likelihood based on Equations (1) and (4)
  - Apply EM algorithm to two reduced models, one without α and the other without γ, to obtain the corresponding maximum likelihoods
- Likelihood ratio test is used for testing H_0: $\gamma = 0$ and H_0: $\alpha = 0$
  - Likelihood from the joint model vs. reduced model of γ with $\gamma = 0$ or α with $\alpha = 0$
- Probabilistic as estimating and testing in a joint maximum likelihood framework of two models for the reference and test cohorts

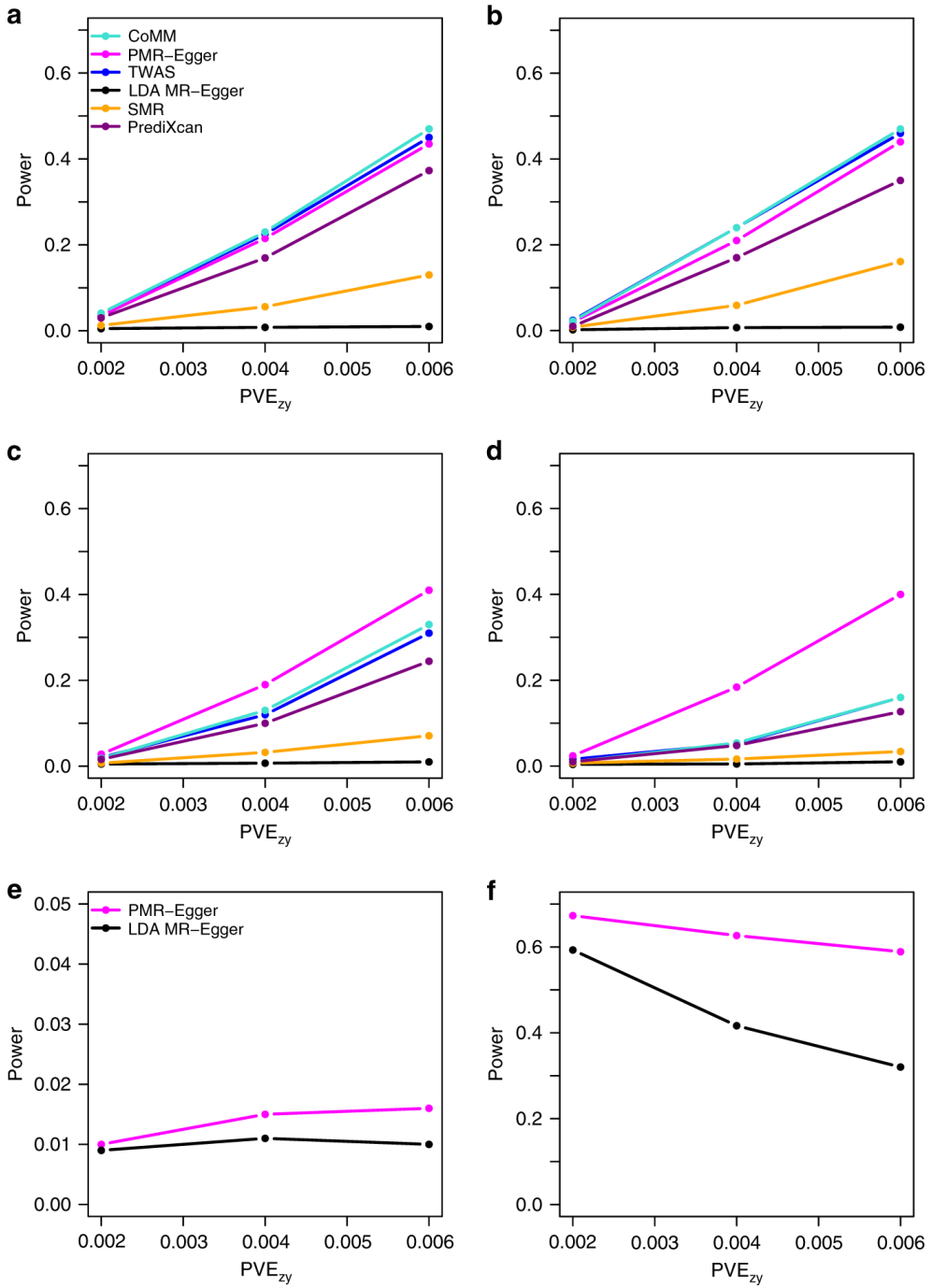


Fig. 2: Power of different methods under various simulation scenarios.
Power (*y* axis) at a false discovery rate of 0.1 to detect the causal effect (**a**–**d**) or the horizontal pleiotropic effect (**e**, **f**), plotted against different causal effect size characterized by PVE_{zy} (*x* axis).

Compared methods include: **CoMM (turquoise), PMR-Egger (magenta), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple)**. Simulations are performed under different horizontal pleiotropic effect sizes: **a** $\gamma=0$; **b** $\gamma=0.0001$; **c**, **e** $\gamma=0.0005$; **d**, **f** $\gamma=0.001$.

Yuan Z. et al. Nature Communications, 2020.

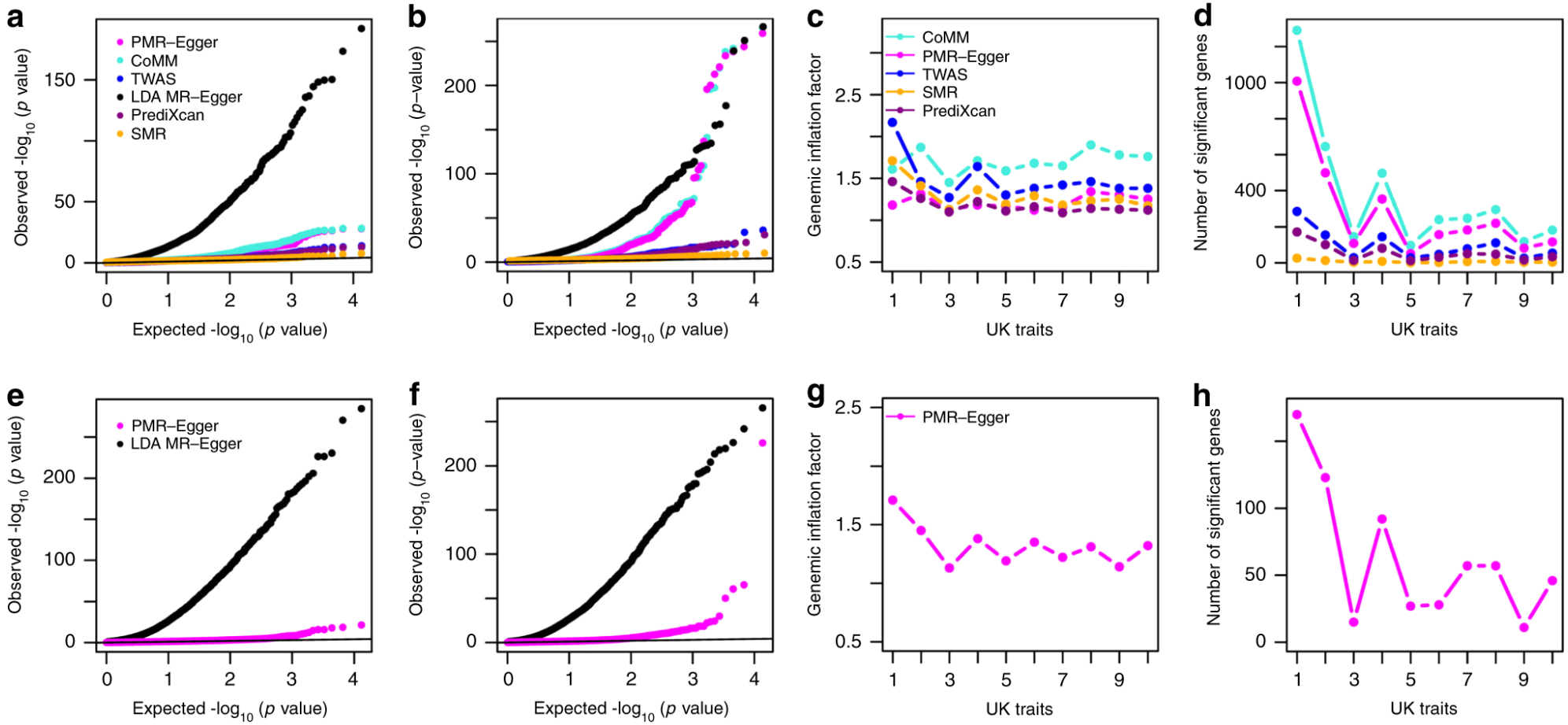


Fig. 6: TWAS analysis results by different methods for UK Biobank traits. **a** QQ plot for testing the causal effect for BMI. **b** QQ plot for testing the causal effect for platelet count. **c** Genomic inflation factor for testing the causal effect for each of the 10 traits by different methods. **d** Number of causal genes identified for each of the 10 traits. **e** QQ plot for testing the horizontal pleiotropic effect for BMI. **f** QQ plot for testing the horizontal pleiotropic effect for platelet count. **g** Genomic inflation factor for testing the horizontal pleiotropic effect for each of the 10 traits. **h** Number of genes identified to have significant horizontal pleiotropic effect for each of the 10 traits. For **c**, **d**, **g**, **h**, the number on the *x* axis represents 10 traits in order: Height, platelet count, bone mineral density, red blood cell count, FEV1–FVC ratio, BMI, RDW, eosinophils count, forced vital capacity, white blood cell count.

PMR-Egger : Advantages

- Account for horizontal pleiotropy in the gene expression and phenotype of interest

- Powerful with individual-level eQTL reference and GWAS test data

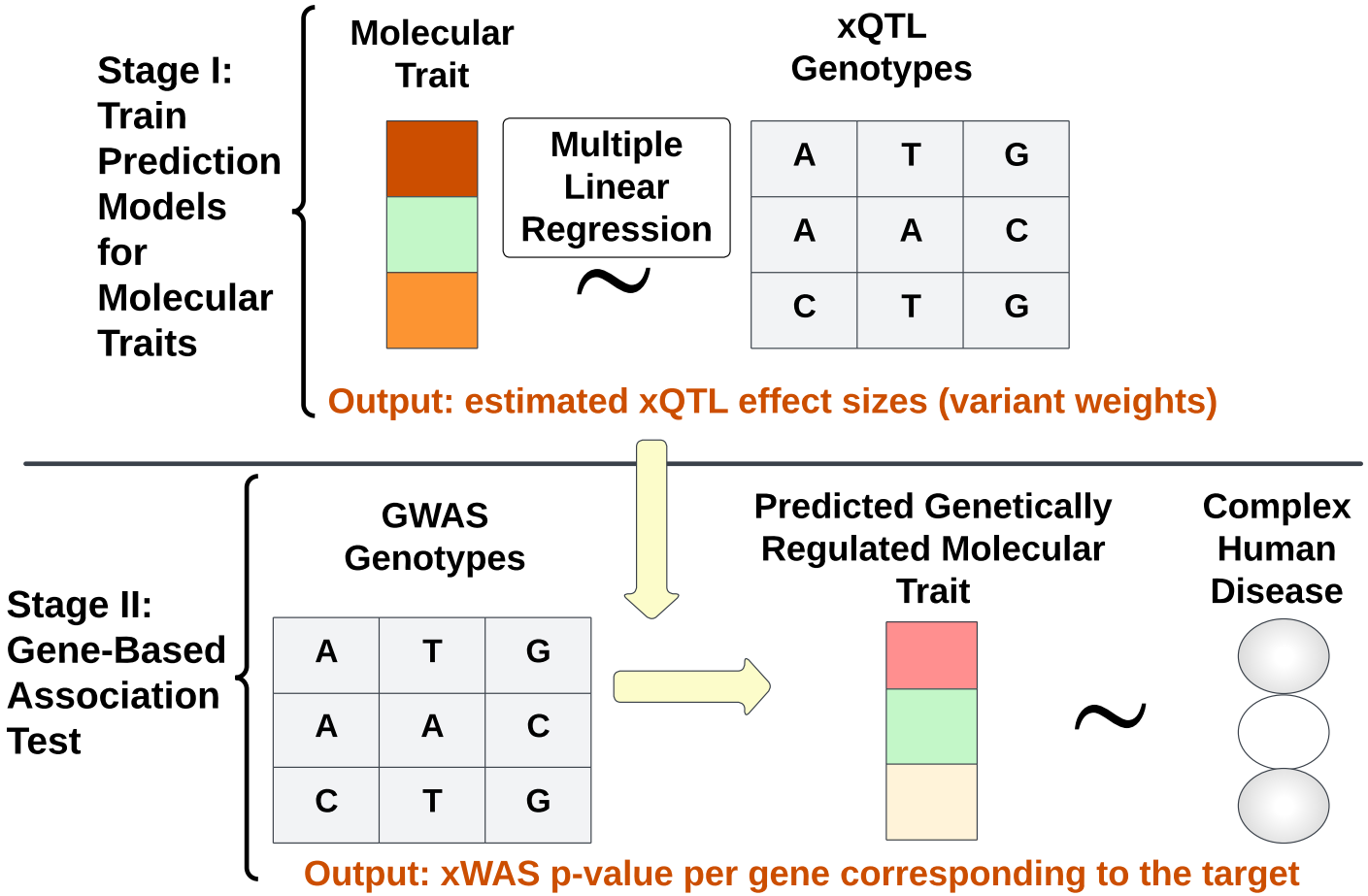
PMR-Egger : Limitations

- Assume equal horizontal pleiotropic effect for all test SNPs
- Developed for continuous traits
- Computationally more complicated and expensive than two-stage TWAS methods
- Possible computation error and reduced power with summary-level eQTL and GWAS data

TWAS and MR

- Uses both eQTL (transcriptomic) and GWAS data that might be profiled for two independent cohorts
- Standard Two-stage TWAS: Test **association (combined causality and the horizonal pleiotropy)** between multiple-SNPs-per-Gene and phenotype of interest, taking non-zero eQTL weights as variant weights for gene-based association tests
- MR: Test **causality** between multiple-SNPs-per-Gene and phenotype of interest, mediated through transcriptome (gene expression)
- PMR-Egger: **Accounts for horizontal pleiotropy** during the the **causality test** between multiple-SNPs-per-Gene and phenotype of interest, mediated through transcriptome (gene expression)

Apply TWAS Framework to Integrate Proteomics Data with GWAS data



Take quantified protein abundances as the response variable of the considered molecular trait.

Proteome-wide Association Study (PWAS)

Limited Reference Proteomics Data?

- OTTERS can be easily applied to leverage other summary-level pQTL data for similar gene-based association study

Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders

Chengran Yang, Fabiana H. G. Farias, Laura Ibanez, Adam Suhy, Brooke Sadler, Maria Victoria Fernandez, Fengxian Wang, Joseph L. Bradley, Brett Eiffert, Jorge A. Bahena, John P. Budde, Zeran Li, Umber Dube, Yun Ju Sung, Kathie A. Mihindukulasuriya, John C. Morris, Anne M. Fagan, Richard J. Perrin, Bruno A. Benitez, Herve Rhinn, Oscar Harari & Carlos Cruchaga

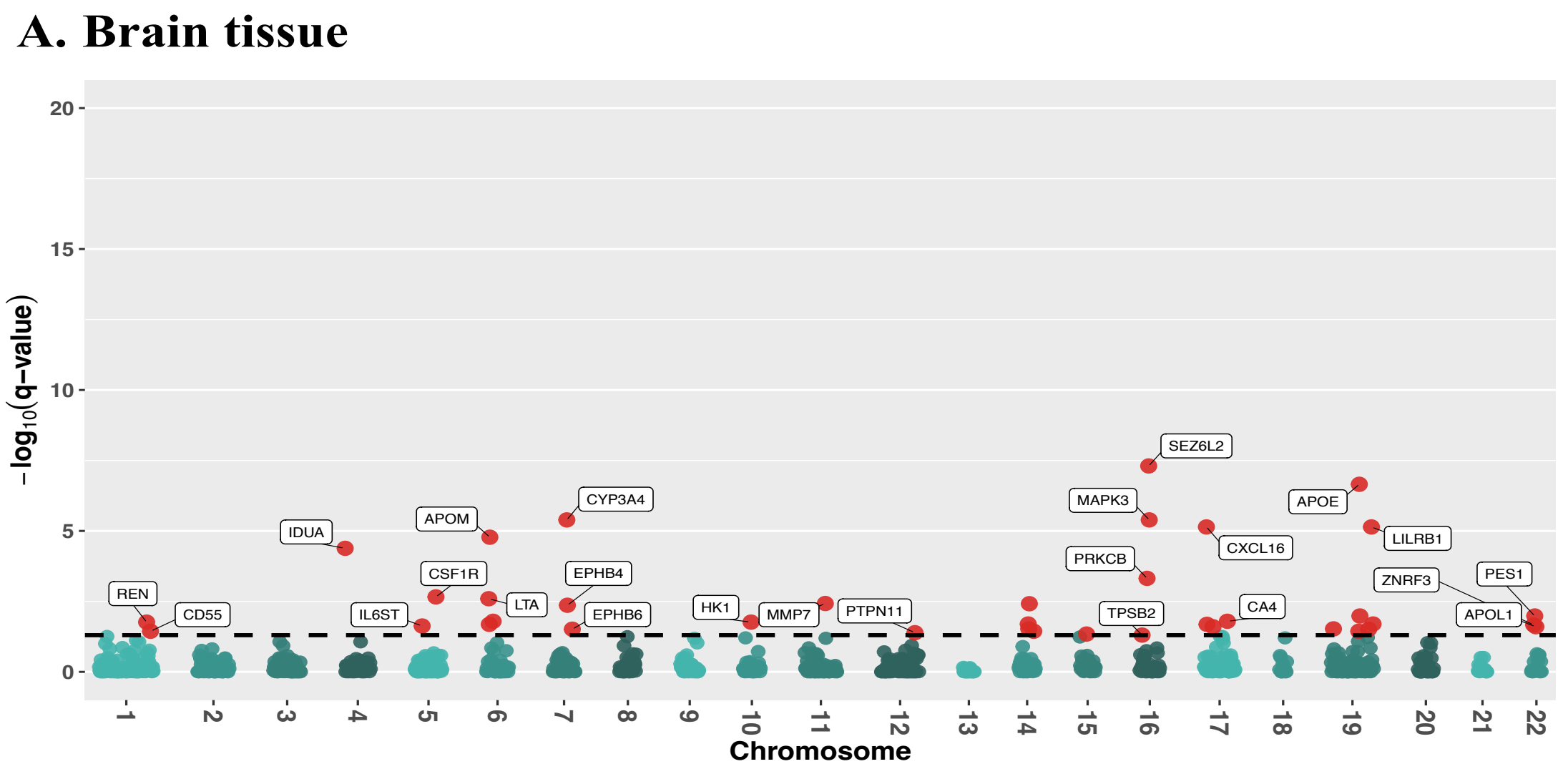
Nature Neuroscience **24**, 1302–1312 (2021) | Cite this article

Apply OTTERS to Summary-level pQTL data

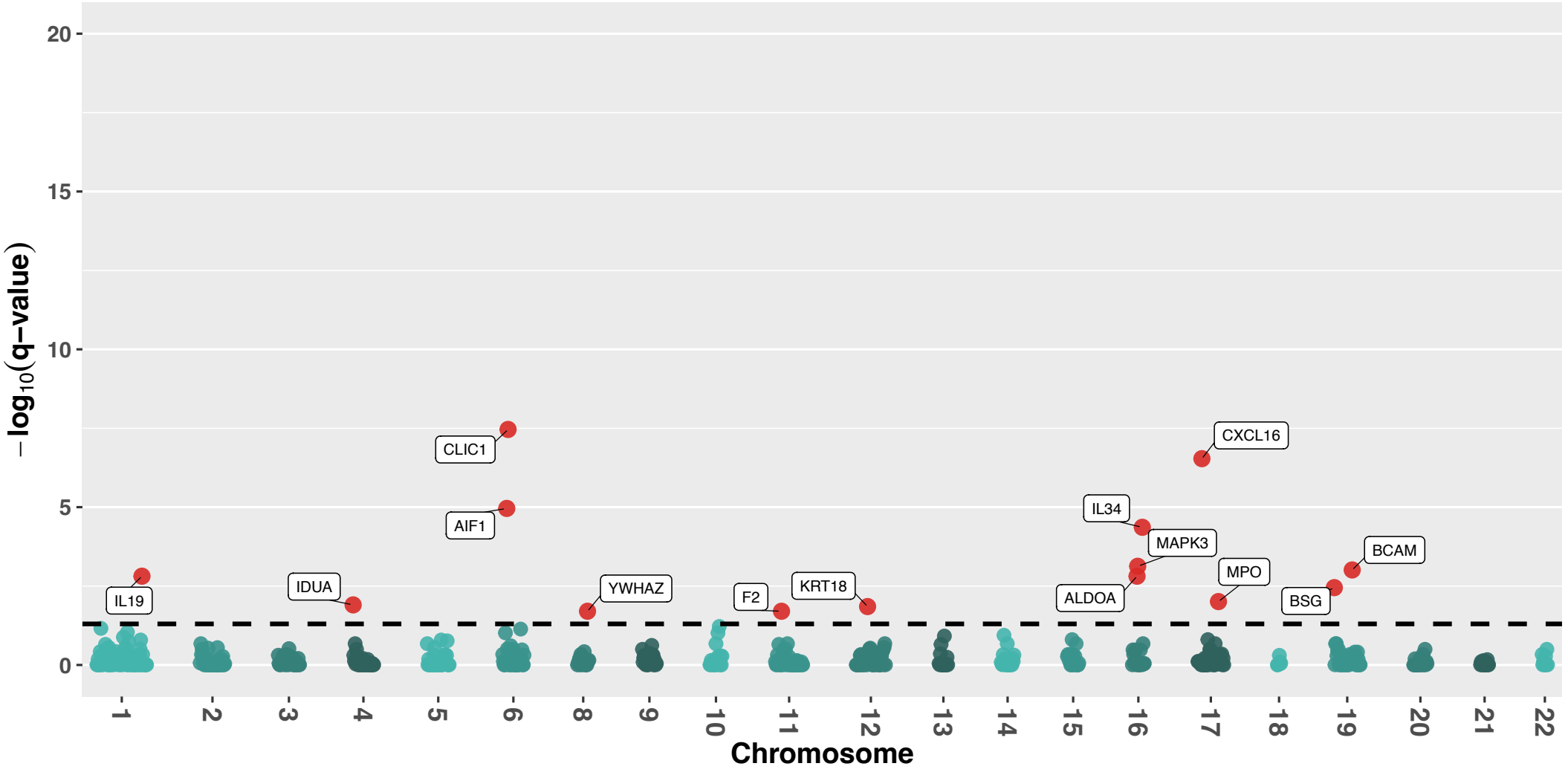
	Brain	CSF (cerebrospinal fluid)	Plasma
Sample size	380	835	529
Number of test proteins/genes	943	591	821

Summary-level pQTL data of brain, CSF, and Plasma.

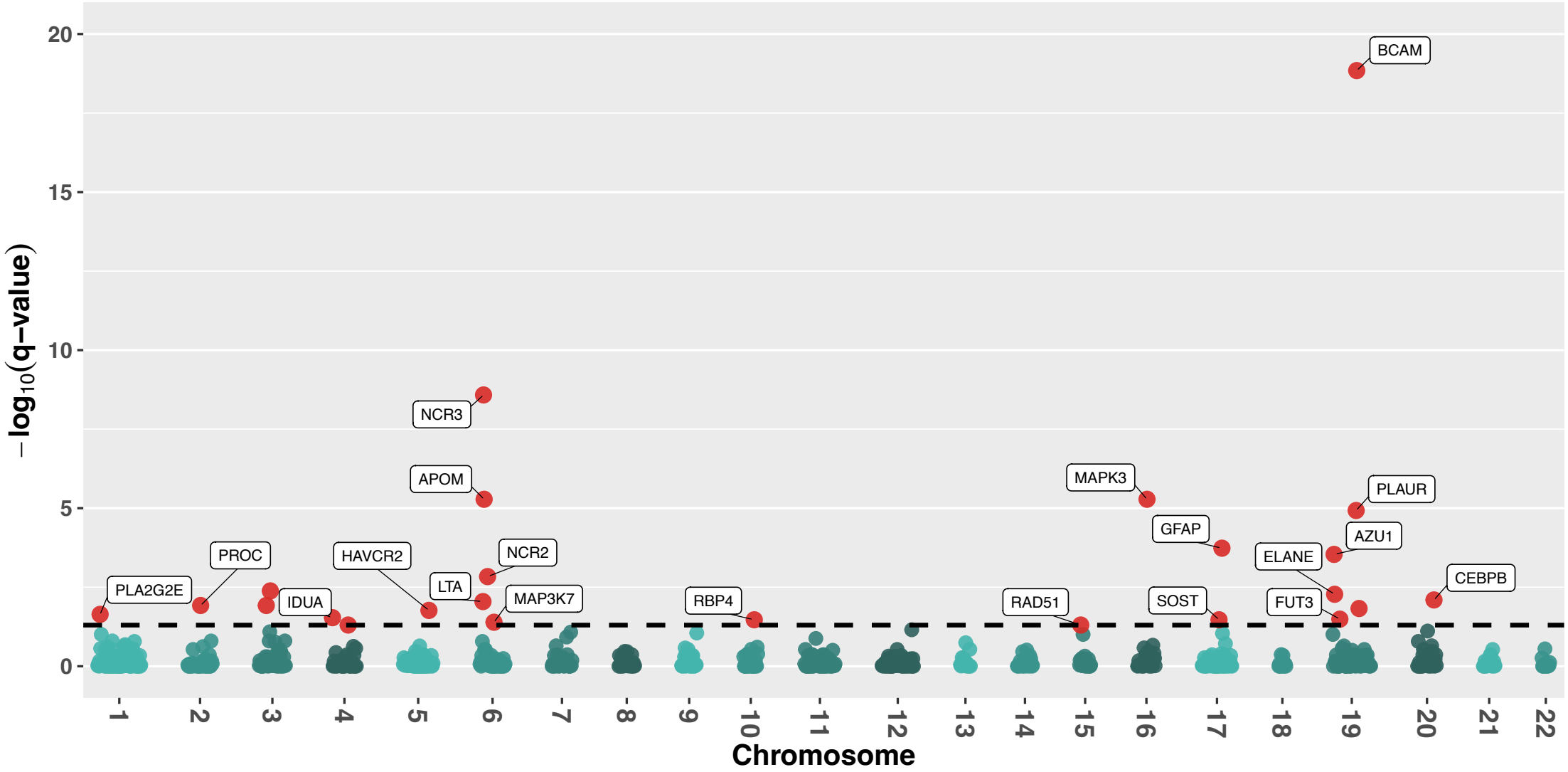
PWAS Results with GWAS Summary Data of AD Dementia (Bellenguez C. et al. Nat. Genet. 2022; n=~789K)



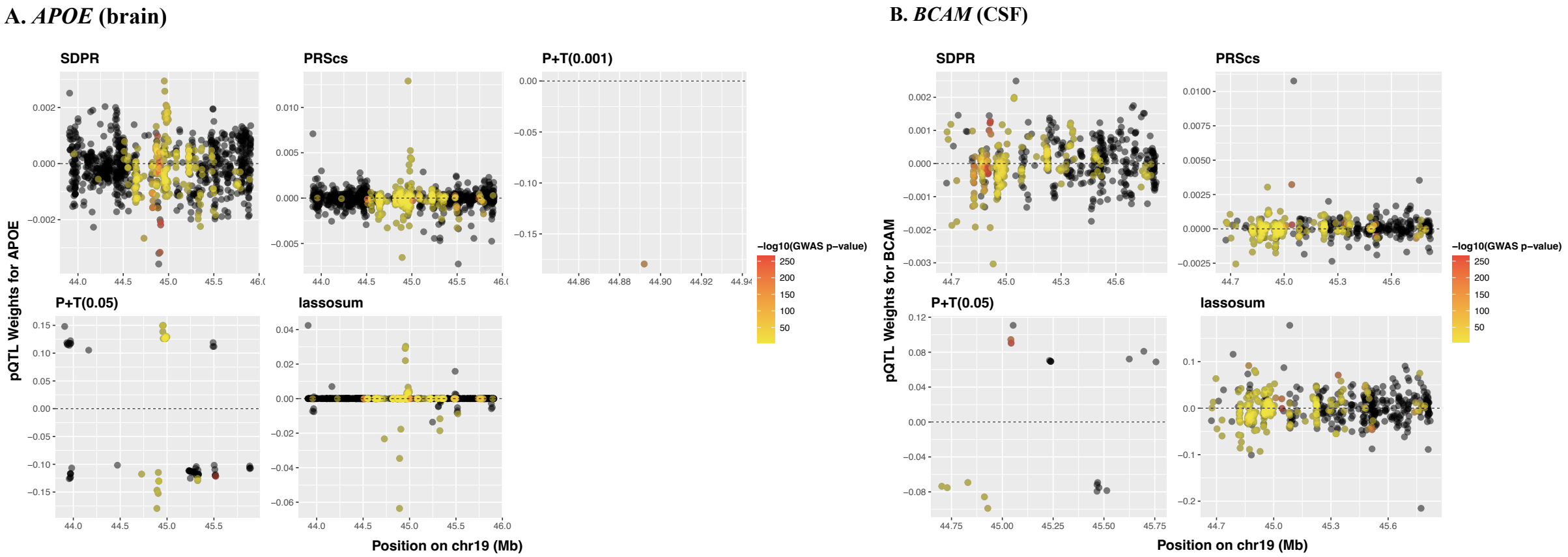
B. CSF tissue


20
15
10
5
0
−log10(q−value)
IL19
IDUA
CLIC1
AIF1
YWHAZ
F2
KRT18
ALDOA
IL34
MAPK3
CXCL16
MPO
BSG
BCAM
1
2
3
4
5
6
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

C. Plasma tissue



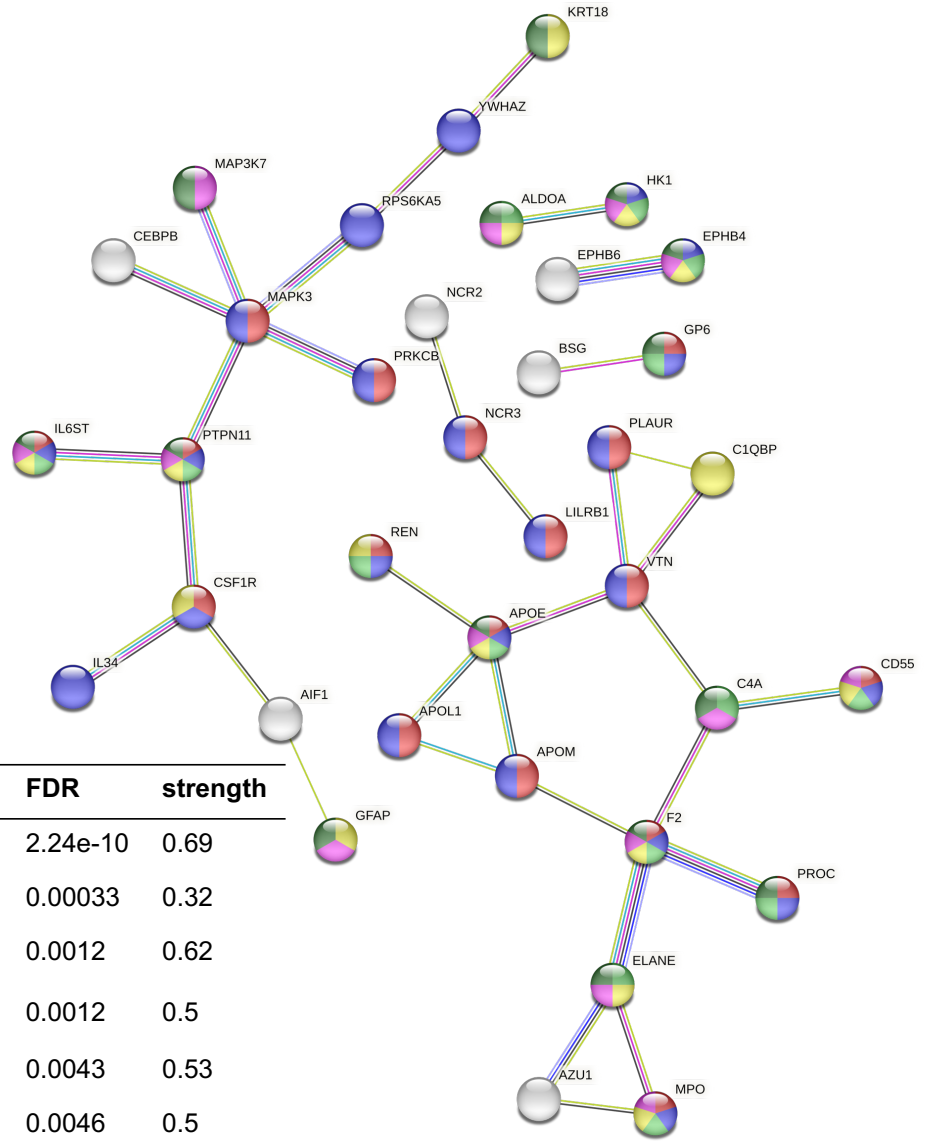
Example Scatter Plots of pQTL Weights



Protein-Protein-Interaction Network by STRING for Genes that are Significant in at least One Tissue

Connections:

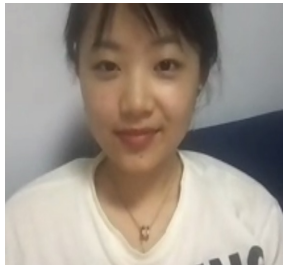
- *From curated databases*
- *Experimentally determined*
- *Gene neighborhood*
- *Gene fusions*
- *Gene co-occurrence*
- *Textmining*
- *Co-expression*
- *Protein homology*



Monarch pathways

	Human Phenotype	phenotype	FDR	strength
	Blood protein measurement	EFO:0007937	2.24e-10	0.69
	Protein measurement	EFO:0004747	0.00033	0.32
	Abnormality of blood and blood-forming tissues	HP:0001871	0.0012	0.62
	Abnormality of metabolism/homeostasis	HP:0001939	0.0012	0.5
	Abnormality of the immune system	HP:0002715	0.0043	0.53
	Abnormality of the skin	HP:0000951	0.0046	0.5

Qiang (Leo) Liu



Tingyang Hu

Apply TWAS Framework to Integrate Epigenetics Data with GWAS data

Meta-Analysis > Nat Genet. 2019 Mar;51(3):445-451. doi: 10.1038/s41588-018-0320-8.
Epub 2019 Jan 14.

Multivariate genome-wide analyses of the well-being spectrum

Bart M L Baselmans [1 2], Rick Jansen [3 4], Hill F Ip [1], Jenny van Dongen [1 2], Abdel Abdellaoui [2 5], Margot P van de Weijer [1], Yanchun Bao [6], Melissa Smart [6], Meena Kumari [6], Gonneke Willemsen [1 2 4], Jouke-Jan Hottenga [1 2 4]; BIOS consortium; Social Science Genetic Association Consortium; Dorret I Boomsma [1 2 4], Eco J C de Geus [1 2 4], Michel G Nivard [# 7 8], Meike Bartels [# 9 10 11]

Affiliations + expand

PMID: 30643256 DOI: 10.1038/s41588-018-0320-8

Abstract

We introduce two novel methods for multivariate genome-wide-association meta-analysis (GWAMA) of related traits that correct for sample overlap. A broad range of simulation scenarios supports the added value of our multivariate methods relative to univariate GWAMA. We applied the novel methods to life satisfaction, positive affect, neuroticism, and depressive symptoms, collectively referred to as the well-being spectrum (N_{obs} = 2,370,390), and found 304 significant independent signals. Our multivariate approaches resulted in a 26% increase in the number of independent signals relative to the four univariate GWAMAs and in an ~57% increase in the predictive power of polygenic risk scores. Supporting transcriptome- and methylome-wide analyses (TWAS and MWAS, respectively) uncovered an additional 17 and 75 independent loci, respectively. Bioinformatic analyses, based on gene expression in brain tissues and cells, showed that genes differentially expressed in the subiculum and GABAergic interneurons are enriched in their effect on the well-being spectrum.

- Taking methylation of CpG sites (151,729) that have at lease one significant mQTL with FDR < 5%
- Used whole-blood methylation data from the BIOS Consortium (Bonder, M.J. et al. Nat. Genet. 2017) as reference data, n=4,008
- Report significant loci

Apply TWAS Framework to Integrate Epigenetics Data with GWAS data

Meta-Analysis > Nat Genet. 2019 Mar;51(3):445-451. doi: 10.1038/s41588-018-0320-8.
Epub 2019 Jan 14.

Multivariate genome-wide analyses of the well-being spectrum

Bart M L Baselmans [1 2], Rick Jansen [3 4], Hill F Ip [1], Jenny van Dongen [1 2], Abdel Abdellaoui [2 5], Margot P van de Weijer [1], Yanchun Bao [6], Melissa Smart [6], Meena Kumari [6], Gonneke Willemsen [1 2 4], Jouke-Jan Hottenga [1 2 4]; BIOS consortium; Social Science Genetic Association Consortium; Dorret I Boomsma [1 2 4], Eco J C de Geus [1 2 4], Michel G Nivard [# 7 8], Meike Bartels [# 9 10 11]

Affiliations + expand

PMID: 30643256 DOI: 10.1038/s41588-018-0320-8

Abstract

We introduce two novel methods for multivariate genome-wide-association meta-analysis (GWAMA) of related traits that correct for sample overlap. A broad range of simulation scenarios supports the added value of our multivariate methods relative to univariate GWAMA. We applied the novel methods to life satisfaction, positive affect, neuroticism, and depressive symptoms, collectively referred to as the well-being spectrum (N_{obs} = 2,370,390), and found 304 significant independent signals. Our multivariate approaches resulted in a 26% increase in the number of independent signals relative to the four univariate GWAMAs and in an ~57% increase in the predictive power of polygenic risk scores. Supporting transcriptome- and methylome-wide analyses (TWAS and MWAS, respectively) uncovered an additional 17 and 75 independent loci, respectively. Bioinformatic analyses, based on gene expression in brain tissues and cells, showed that genes differentially expressed in the subiculum and GABAergic interneurons are enriched in their effect on the well-being spectrum.

- A total of 913 CpG genetically regulated methylation-trait associations, mapped to 141 loci, were detected.
- For 75 out of 913 CpG methylation–trait associations (36 loci), the corresponding locus did not contain a significant GWAS signal.
- For 396 CpG methylation–trait associations (83 loci), the maximum LD between the MWAS model SNPs and the GWAS top SNP was larger than 0.8

Differences of Methylome-wide Association Study (MWAS) from TWAS/PWAS

- A large number of CpG sties
  - Multiple CpG sites are located near one gene
  - Correlated CpG sites near one gene
- How to select CpG sites for the analysis?
- How to combine genetically regulated DNA methylation-trait associations across multiple nearby CpG sites for one target gene?

JOURNAL ARTICLE

A gene-level methylome-wide association analysis identifies novel Alzheimer's disease genes FREE

Chong Wu ✉, Jonathan Bradley, Yanming Li, Lang Wu, Hong-Wen Deng

Bioinformatics, Volume 37, Issue 14, July 2021, Pages 1933–1940,
https://doi.org/10.1093/bioinformatics/btab045
Published: 01 February 2021 **Article history** ▼

Suggested MWAS Steps

(Adapted from the Cross Methylome Omnibus, CMO, method by Wu et. al, Bioinformatics 2021)

- Links CpG sites located in enhancers, promoters, and gene body (exons, introns) to a target gene.
  - These linked CpG sites are included in the MWAS analysis.
- Step 1: Train imputation models for DNAm ~ cis-SNPs
  - For CpG sites with at least one potential significant mQTL with FDR < 5%, or p-value < 1e-5.
  - Cis-SNPs : A narrower window around the CpG sites, +-250Kb?
  - **Apply OTTERS to summary-level mQTL data.**
- Step 2: Test associations between genetically regulated DNAm of each linked CpG and the trait of interest in the test GWAS data.
- Step 3: For each target gene, combine tests of multiple linked CpG sites by the ACAT method.

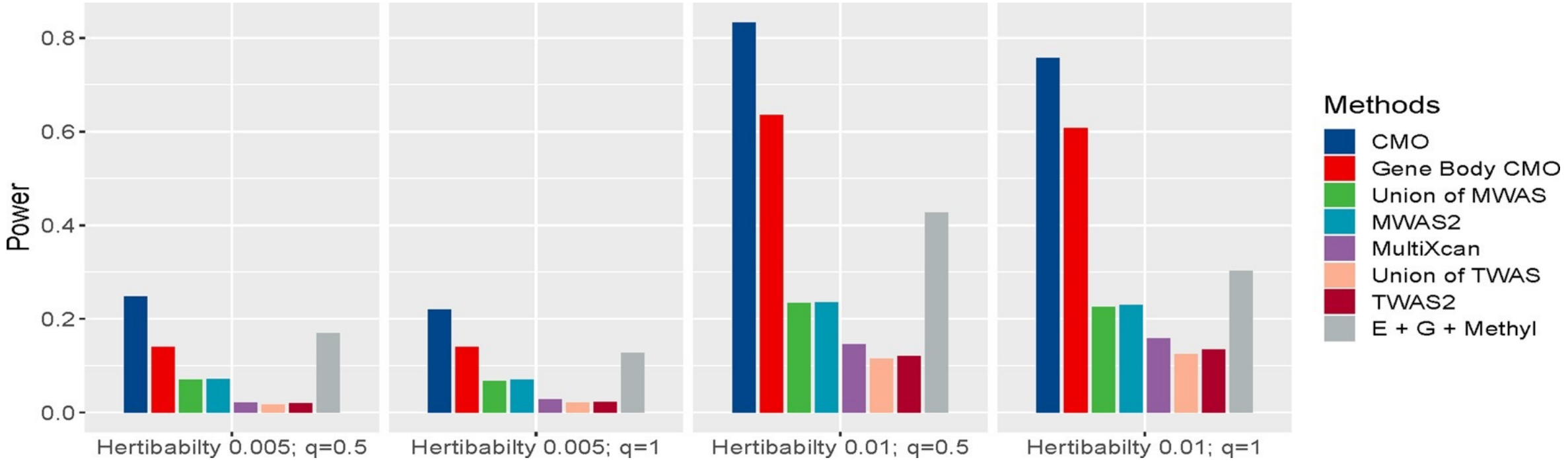
Linking CpG Sites to a Target Gene
(Wu et. al, Bioinformatics 2021)

- Determine CpG sites located in the enhancers, promoters, and gene body to the target gene.
  - A gene body region includes all the introns and exons of a target gene.
  - To include cis-acting regulatory regions and alleviate the burden of determining gene direction, two promoters of a target gene are defined as a 500 bp extension (Andersson et al., 2014) on either side of the gene body region beyond its transcription start site (TSS) and transcription end site (TES), respectively.
  - Used an integrated database called *GeneHancer* (Fishilevich *et al.*, 2017) to determine enhancer regions for each gene, which integrates reported enhancers from four genome-wide databases (the ENCODE, the Ensembl regulatory build, the FANTOM and the VISTA Enhancer browser).

Simulation Studies (Wu et. al, Bioinformatics 2021)

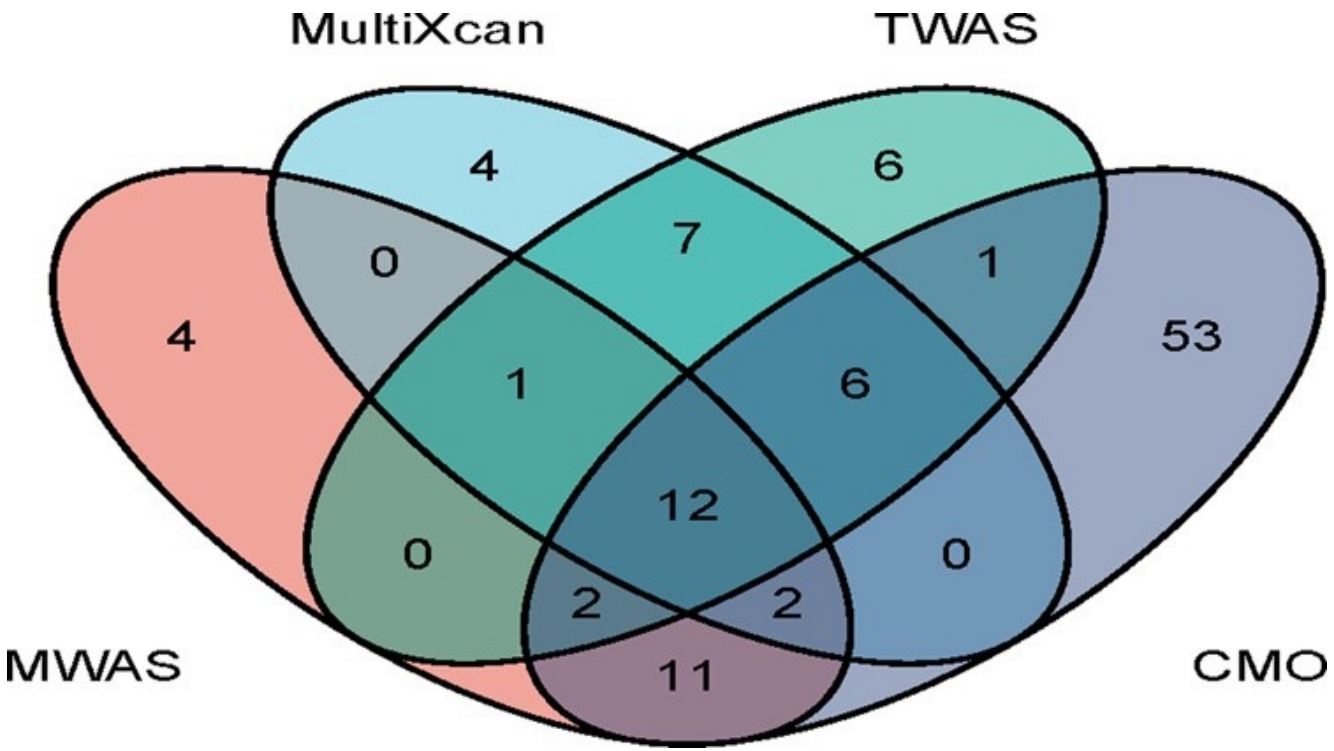
- Select a target gene. Used the real genotype data.
- Randomly select m CpG sites
- Consider expressions in m tissues
- Generate the cis-component of DNAm in CpG i and gene expression values in tissue j by $C_i = Xw_i^c,\ E_j = Xw_j^E$
  - w_i^c and w_j^E are derived from estimates from the real data
- Simulate quantitative trait by $Y = \beta \sum_{i=1}^{m} C_i + \alpha \sum_{j=1}^{m} E_j + \epsilon$, values of (β, α) were determined for the target DNAm and Expression heritability

Example MWAS Results by CMO (Wu C. et al. Bioinformatics, 2021)



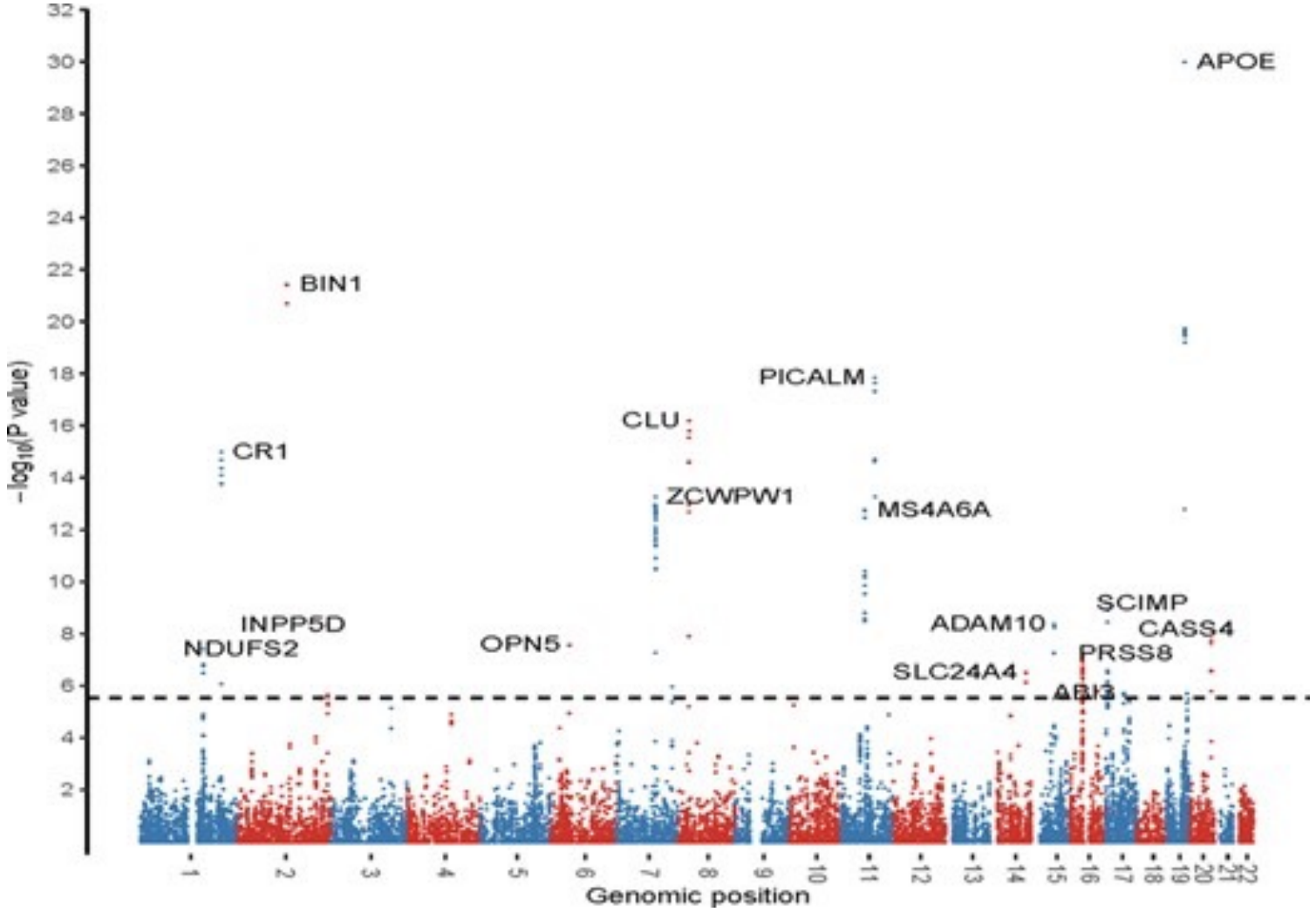
Simulation Studies: Assuming three CpG sites and three tissues were associated with the trait (*m* = 3) and DNAm levels were negatively associated with gene expression.

Example MWAS Results by CMO (Wu C. et al. Bioinformatics, 2021)



CMO identified more significant genes in the IGAP dataset. TWAS stands for TWAS that considers all tissues while applying an additional Bonferroni correction. MWAS is a method that combines results across CpG sites in the gene body and promoter regions while applying an additional Bonferroni correction

Example MWAS Results by CMO (Wu C. et al. Bioinformatics, 2021)



CMO results using GWAS summary data generated by Jansen et al. Nat. Genet. 2019 (n=455,258).

Web Resources

- PMR-Egger
  - https://github.com/yuanzhongshang/PMR
- CMO
  - https://github.com/ChongWuLab/CMO