Transcriptome-wide Association Studies

BIOS 770

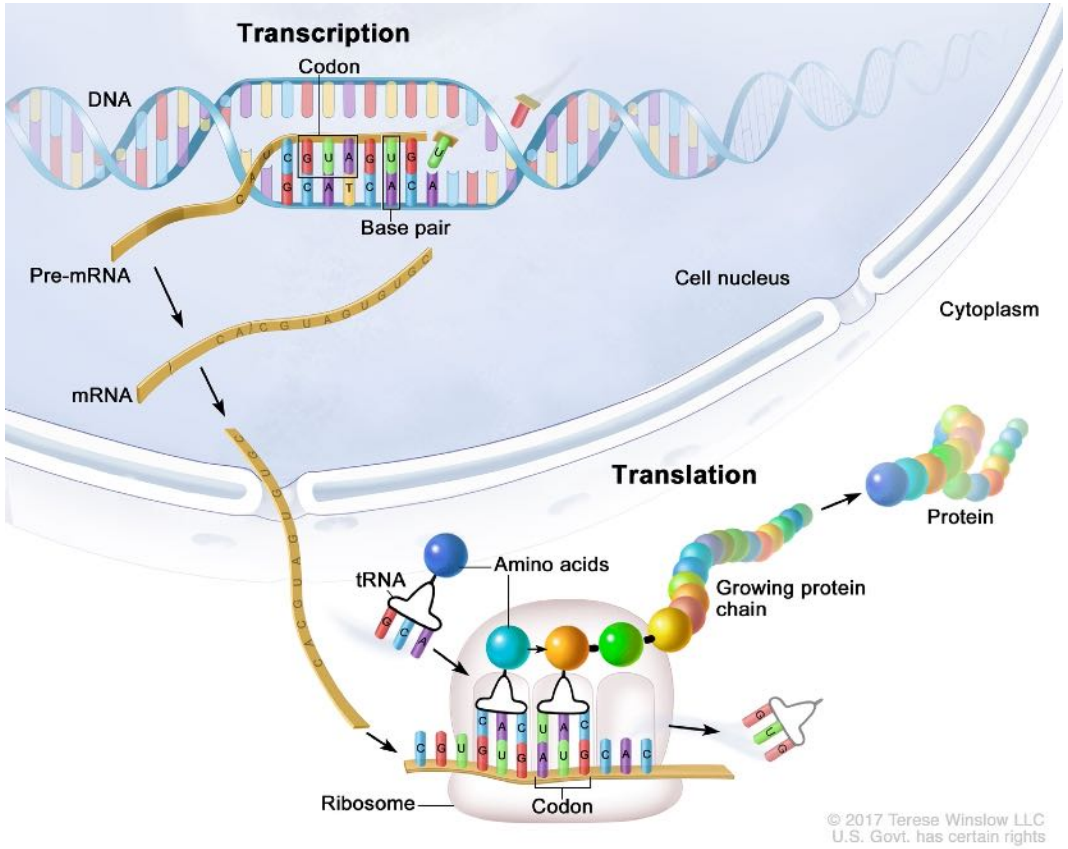
03/01/2022

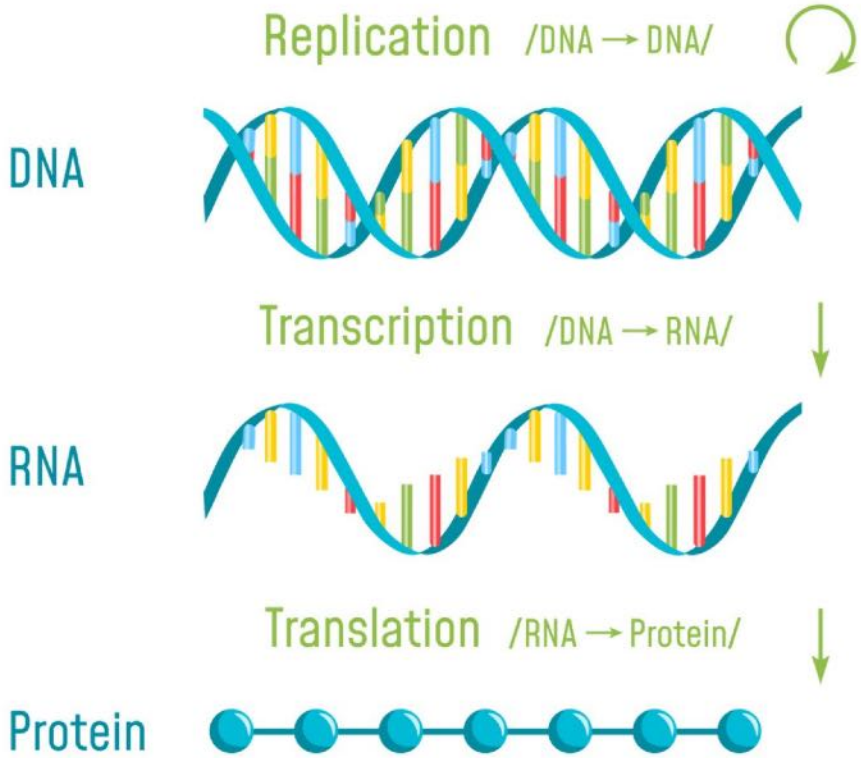
Jingjing Yang (jingjing.yang@emory.edu)

Outline

- Transcriptome-wide Association Studies (TWAS)
  - PrediXcan based on Elastic-Net model (Gamazon et. al. Nature Genetics, 2015)
  - TIGAR based on Bayesian Dirichlet Process Regression model (Nagpal et. al, AJHG 2019)
    - Burden TWAS
    - Variance Component TWAS
  - PMR-Egger based on a Mendelian Randomization (MR) likelihood framework that unifies existing TWAS and MR methods (Yuan et. al, Nature Communications, 2020).

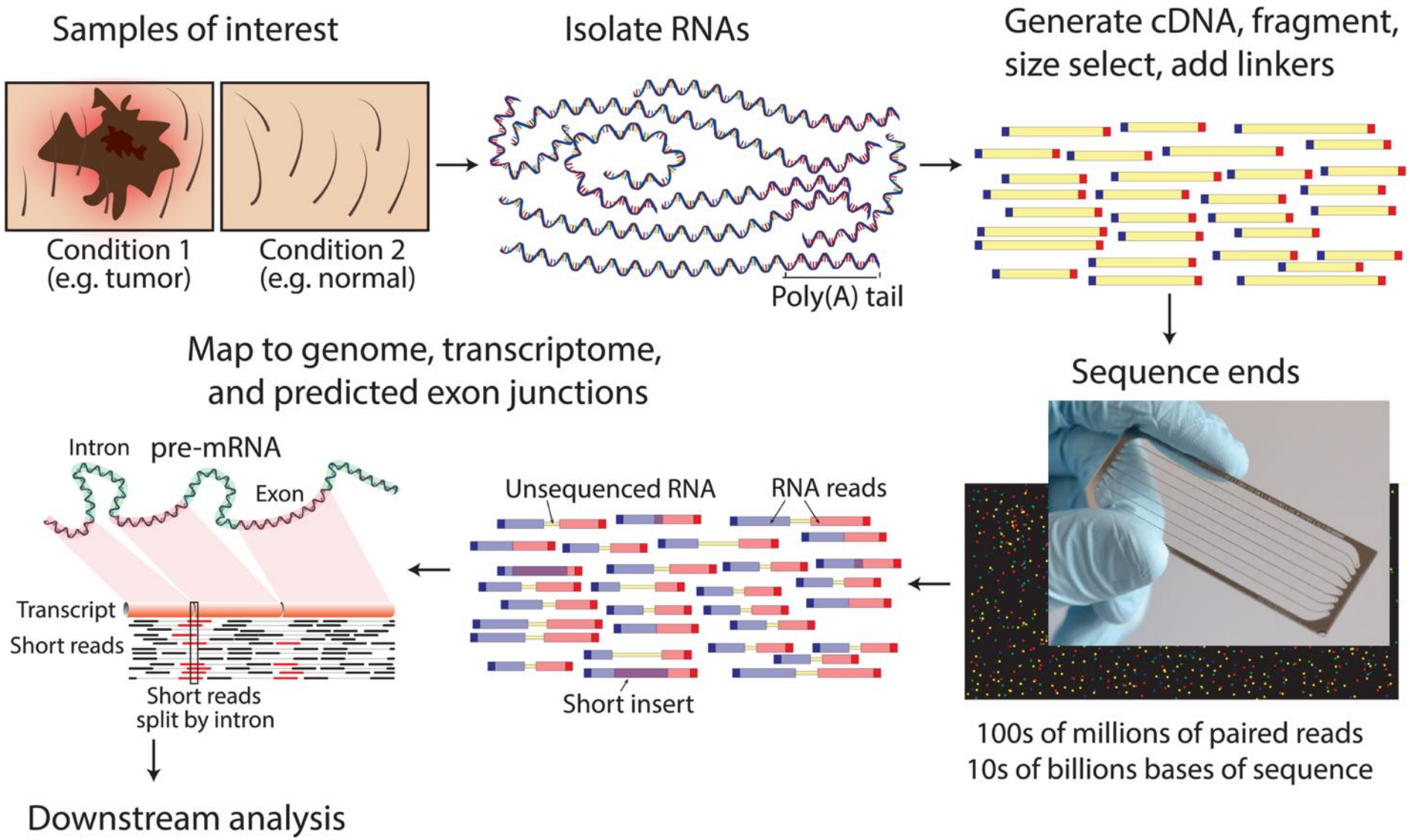
Transcription





https://www.thoughtco.com/dna-transcription-373398

Profile Gene Expression Levels by RNA-sequencing



Transcriptomic Data

- Gene expression Quantitative Traits
  - Profiled by RNA sequencing (RNA-seq)
  - TPM (Transcripts Per Kilobase Million) per gene
    - Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
    - Count up all the RPK values in a sample and divide this number by 1,000,000. This is your “per million” scaling factor.
    - Divide the RPK values by the “per million” scaling factor. This gives you TPM.
- 20K ~ 25K genes in human genome
- Tissue specific
  - Subject to relevant tissue availability
- Could be time specific
  - Vary over development or disease course
- Could be cell-type specific
  - Bulk RNA-seq vs. Single Cell RNA-seq

eQTL

- Consider the profiled gene expression levels as the quantitative trait E_g in the following single variant linear regression model:

$$E_g = \beta_0 + \alpha C + \beta_1 X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

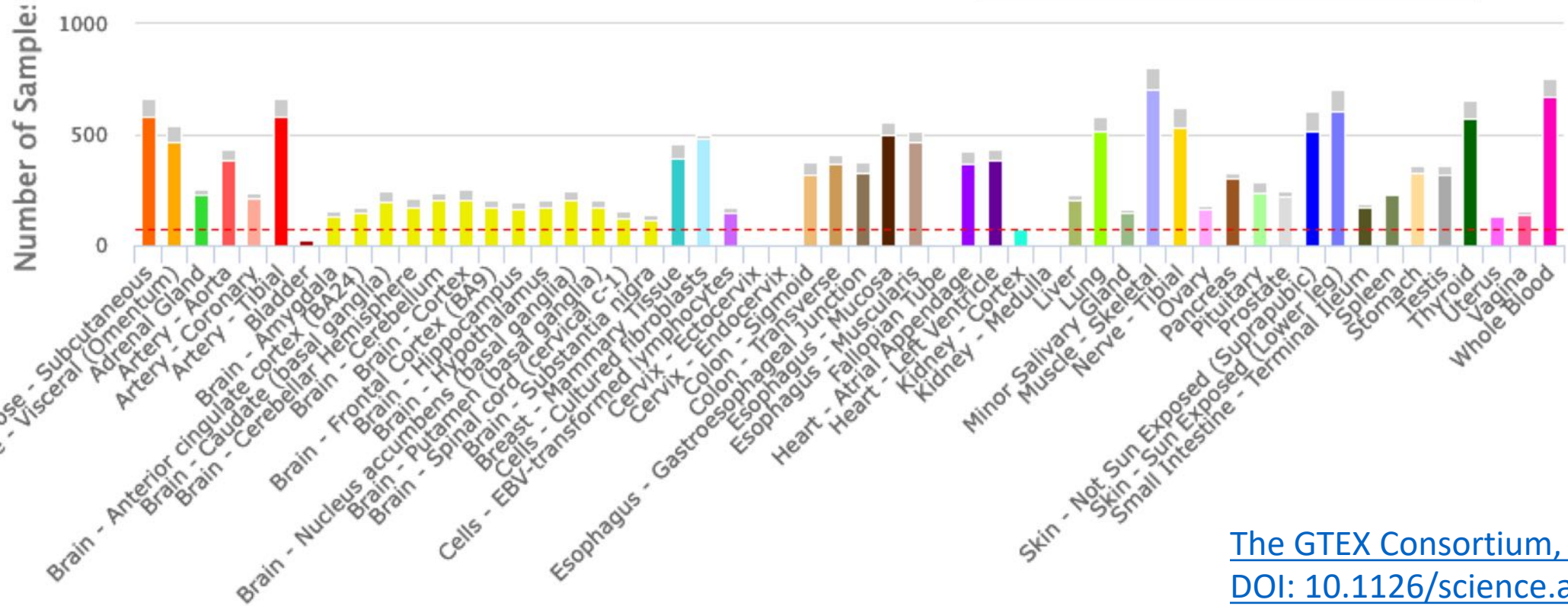
  - X represents the genotype data (0, 1, 2) or dosage [0, 2] of the test SNP
  - C represents the confounding covariates or other environmental variables
  - ϵ represents the error term, other unknown factors
- eQTL :SNPs significantly associated with a gene expression quantitative trait ($H_0: \beta_1 = 0$ is significantly rejected) are referred as **e**xpression **Q**uantitative **T**rait **L**oci (eQTL)
- Cis-eQTL : eQTL nearby the test gene (e.g., located within the +-1MB region of the transcription starting site; thousands cis-SNPs per gene).
- Trans-eQTL : eQTL distant from the test gene (e.g., located out of the +-1MB region of the transcription starting site, or on different chromosome; ~10M trans-SNPs per gene).

Genotype-Tissue Expression (GTEx) project

https://www.gtexportal.org/home/

V8 Release	# Tissues	# Donors	# Samples
Total	54	948	17382
With Genotype	54	838	15253
Has eQTL Analysis*	49	838	15201

** Number of samples with genotype >= 70*



The GTEX Consortium, Science 2020.
DOI: 10.1126/science.aaz1776

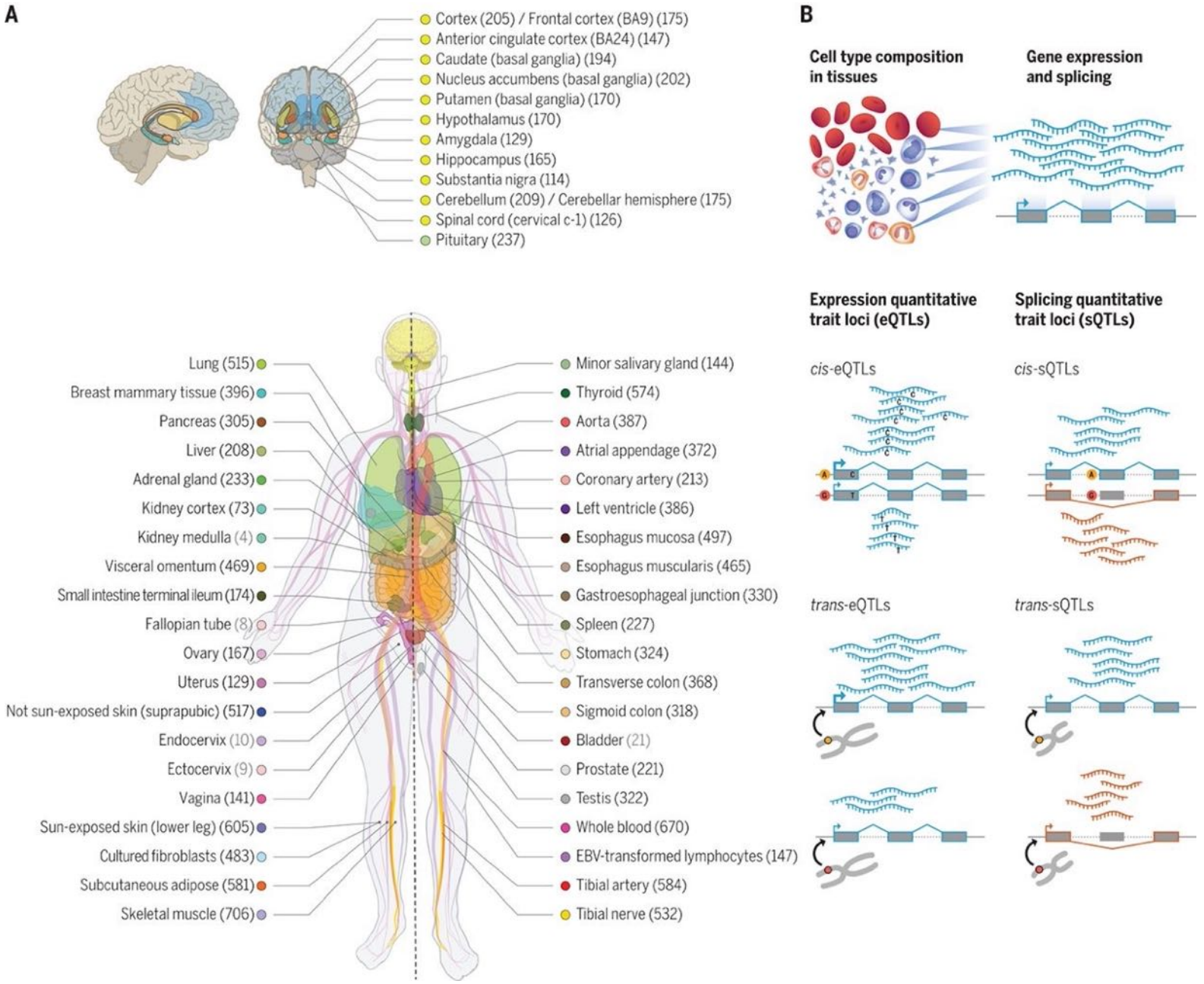


Fig. 1 Sample and data types in the GTEx v8 study.

(**A**) Illustration of the 54 tissue types examined (including 11 distinct brain regions and two cell lines), with sample numbers from genotyped donors in parentheses and color coding indicated in the adjacent circles. Tissues with 70 or more samples were included in QTL analyses.

(**B**) Illustration of the core data types used throughout the study. Gene expression and splicing were quantified from bulk RNA-seq of heterogeneous tissue samples, and local and distal genetic effects (cis-QTLs and trans-QTLs, respectively) were quantified across individuals for each tissue.

The GTEX Consortium, Science 2020.
DOI: 10.1126/science.aaz1776

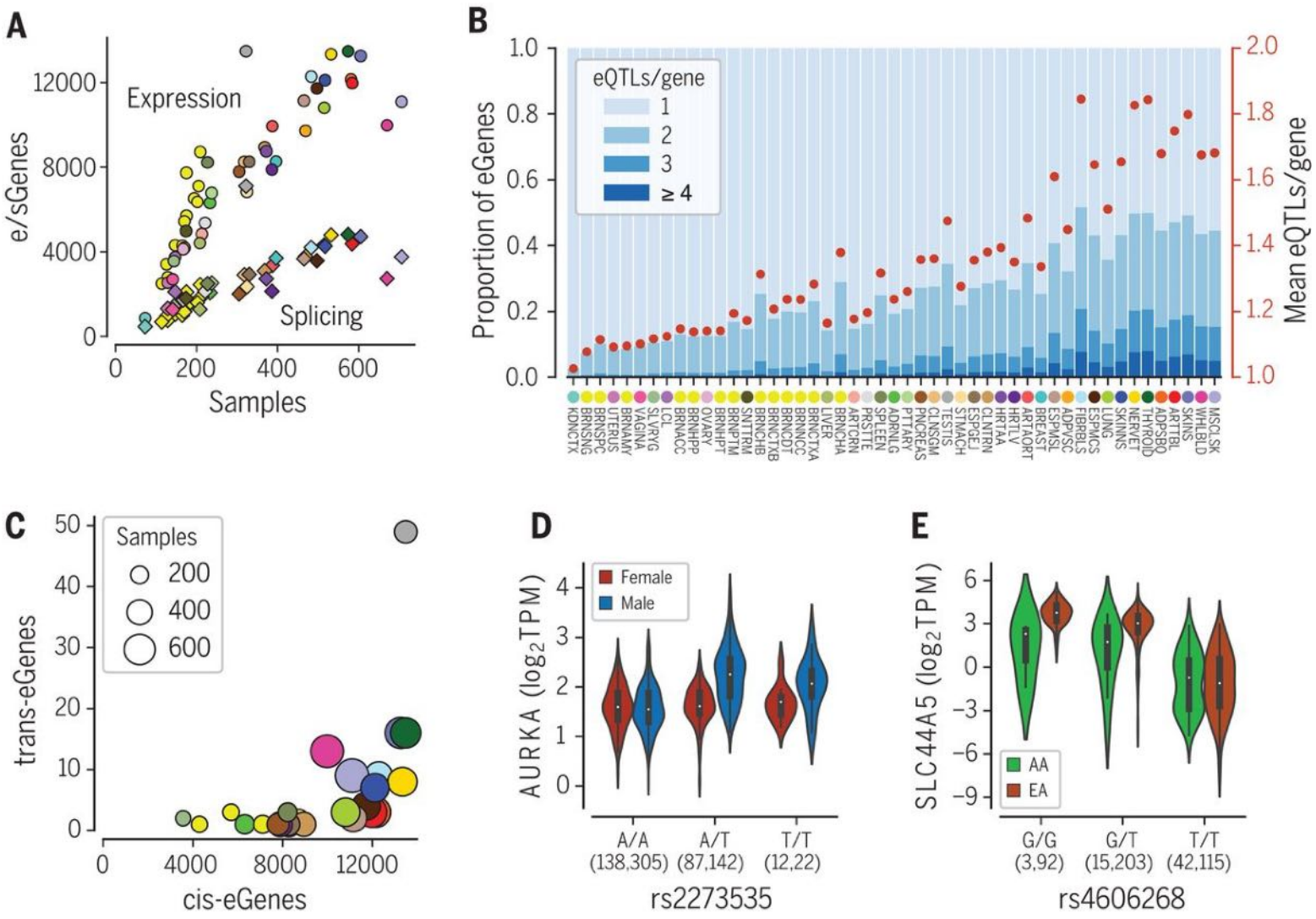


Fig 2. QTL discovery.

(**A**) The number of genes with a cis-eQTL (eGenes) or cis-sQTL (sGenes) per tissue, as a function of sample size. See Fig. 1A for the legend of tissue colors.

(**B**) Allelic heterogeneity of cis-eQTLs depicted as proportion of eGenes with one or more independent cis-eQTLs (blue stacked bars; left *y* axis) and as a mean number of cis-eQTLs per gene (red dots; right *y* axis). The tissues are ordered by sample size.

(**C**) The number of genes with a trans-eQTL as a function of the number of cis-eGenes.

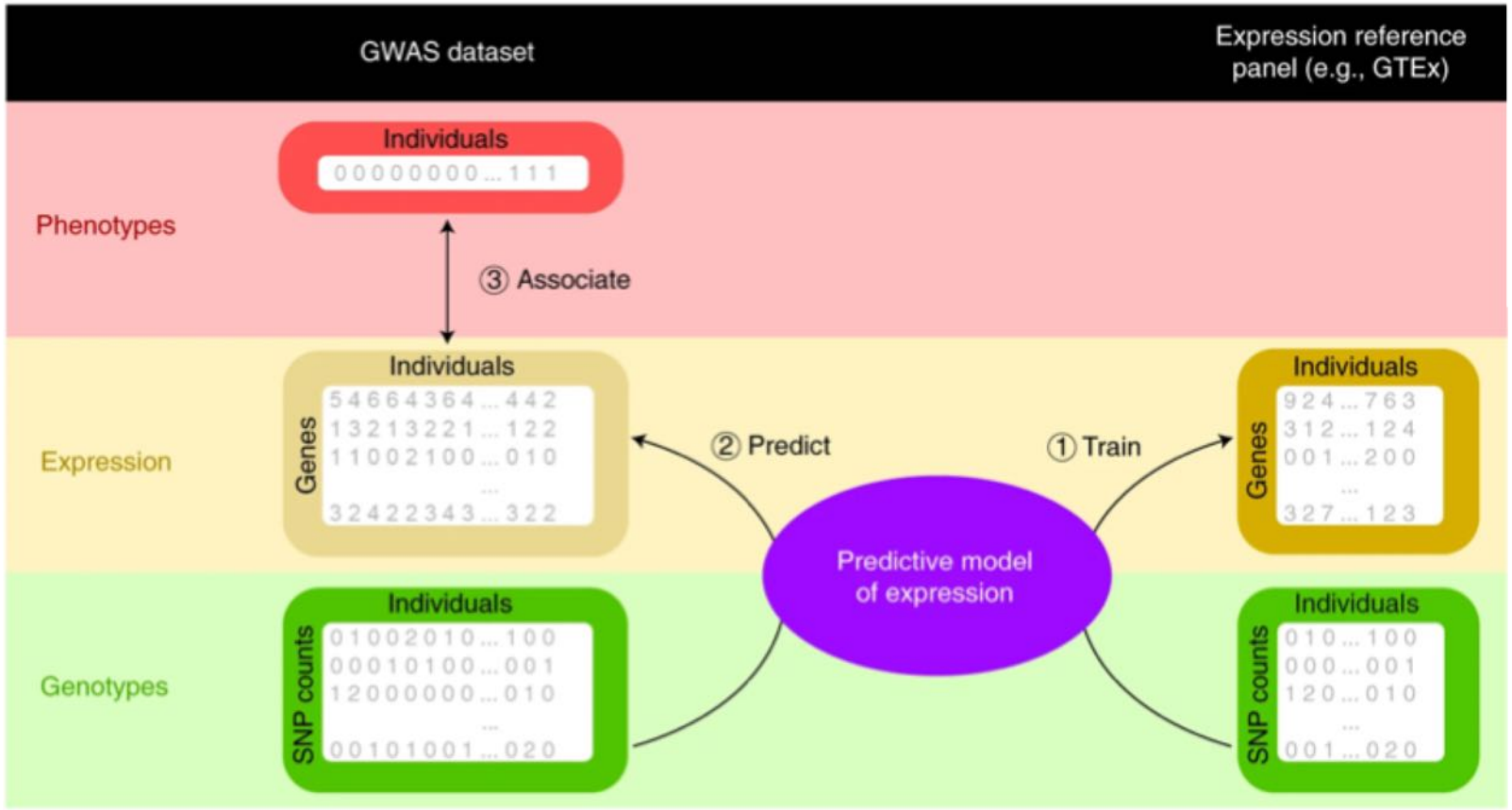
(**D**) Sex-biased cis-eQTL for *AURKA* in skeletal muscle, where rs2273535-T is associated with increased *AURKA* expression in males ($P = 9.02 \times 10^{-27}$) but not in females ($P = 0.75$).

(**E**) Population-biased cis-eQTL for *SLC44A5* in esophagus mucosa (aFC = −2.85 and −4.82 and in African Americans (AA) and European Americans (EA), respectively; permutation *P* value = 1.2×10^{-3}).

TPM, transcripts per million.

The GTEX Consortium, Science 2020.
DOI: 10.1126/science.aaz1776

Transcriptome-wide Association Studies (TWAS) Framework



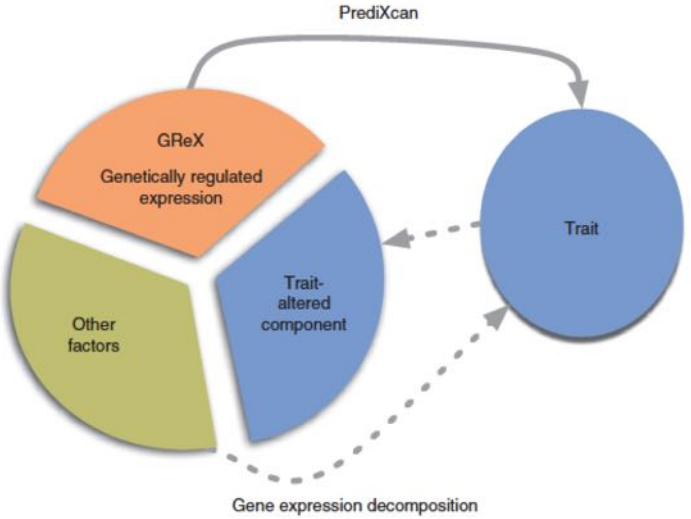
Wainberg M. et. al. Nat. Genetics. 2019.

- Quantitative gene expression trait T_g
- Genotype data of all cis-SNPs X
- SNP effect sizes on transcriptome (eQTL effect sizes) w_i

$$\mathbf{T}_g = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$$

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

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



Gamazon ER et. al., Nat Genetics, 2015

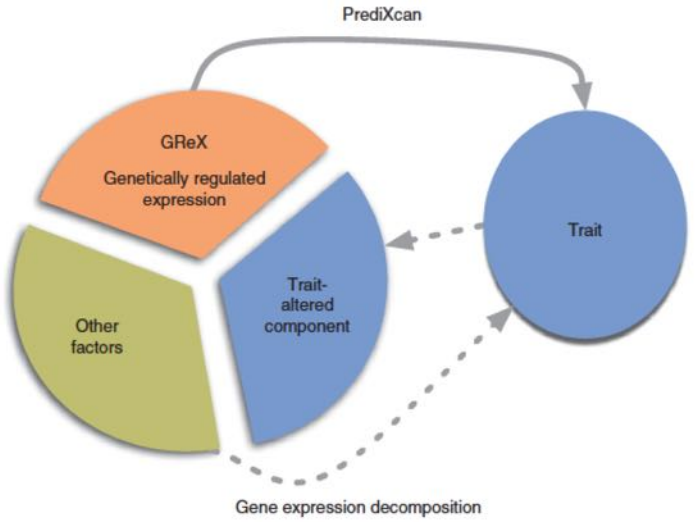
GReX: **G**enetically **R**egulated Gene **eX**pression

PrediXcan

- Stage I: Estimates eQTL weights $\widehat{\boldsymbol{w}}$ from a penalized regression model with Elastic-Net penalty, a combination of L_1 (LASSO) and L_2 (Ridge) penalties

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left(\left\| \boldsymbol{T_g} - \boldsymbol{X}\boldsymbol{w} \right\|_2^2 + \lambda \left(\alpha \|\boldsymbol{w}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{w}\|_2^2 \right) \right)$$

  - where $\|\cdot\|_2$ denotes L_2 norm, $\|\cdot\|_1$ denotes L_1norm, $\alpha \in [0, 1]$ denotes the proportion of L_1 penalty, and λ denotes the penalty parameter
  - Takes $\alpha = 0.5$ and tunes the penalty parameter λ by a 5-fold cross validation



Gamazon ER et. al., Nat Genetics, 2015

PrediXcan

- State II: Test gene-based association between GReX and phenotype of interest, using individual-level or summary-level (Z-scores Z_l from single variant tests) GWAS data

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

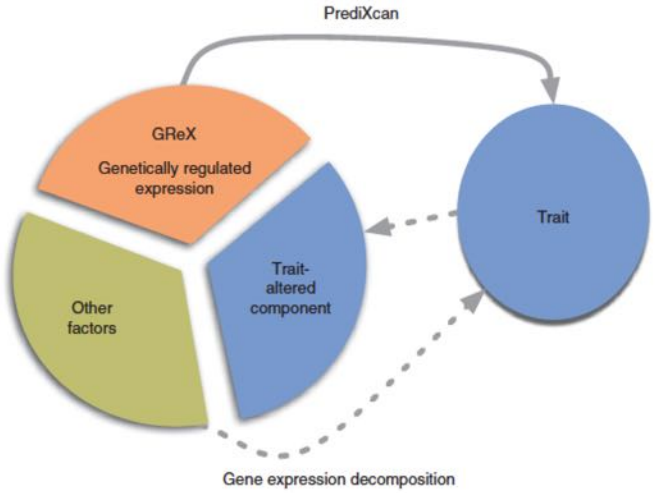
$$\tilde{Z}_{g,\text{FUSION}} = \frac{\sum_{l=1}^{m}(\widehat{w}_l Z_l)}{\sqrt{\widehat{\mathbf{w}}'\mathbf{V}\widehat{\mathbf{w}}}}, \quad \mathbf{V} = \text{Corr}(\mathbf{G}_0)$$

$$\tilde{Z}_{g,\text{S-PrediXcan}} = \frac{\sum_{l=1}^{m}(\widehat{w}_l\widehat{\sigma}_l Z_l)}{\sqrt{\widehat{\mathbf{w}}'\mathbf{V}\widehat{\mathbf{w}}}}, \quad \widehat{\sigma_l^2} = \text{Var}(\mathbf{G}_{0,l}), \quad \mathbf{V} = \text{Cov}(\mathbf{G}_0).$$

- Existing TWAS tools all assume SNP effect sizes on phenotype $\beta_i = \gamma\widehat{w}_i,\ i = 1, \cdots, m.$
- Burden test: $H_0 : \gamma = 0$

$\widehat{\sigma_l^2}$: genotype variance

$\mathbf{G}_0$: reference genotype data



Gamazon ER et. al., Nat Genetics, 2015

TIGAR: Transcriptome-Integrated Genetic Association Resource

Estimates eQTL weights $\hat{w}$ from Bayesian Dirichlet Process Regression Model

- Considering gene expression levels $\mathbf{E_g}$ of gene g
  genotype data matrix $\mathbf{X_{n\times p}}$ of all cis-SNPs
- $\mathbf{E_g}$ were normalized and adjusted for covariates such as age, sex, genotype PCs, PEER factors of transcriptomic data
- The nonparametric Bayesian Dirichlet process regression (DPR) model 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

Nagpal et. al, AJHG 2019

Bayesian Dirichlet Process Regression 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

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

Nagpal et. al, AJHG 2019

TIGAR

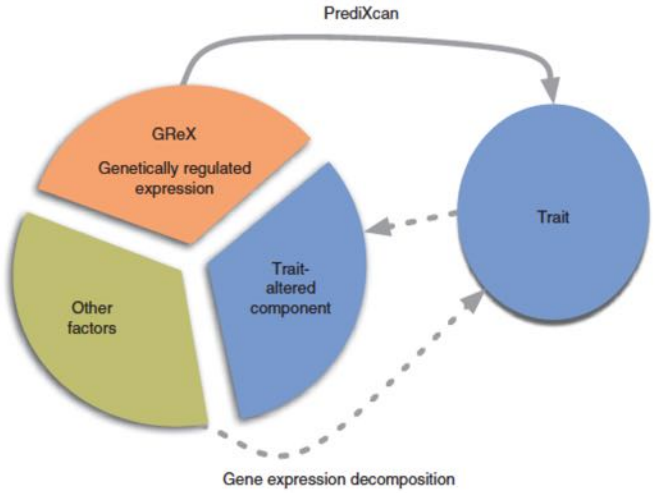
- State II (the same procedure as PrediXcan): Test gene-based association between GReX and phenotype of interest, using individual-level or summary-level (Z-scores Z_l from single variant tests) GWAS data

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

$$\tilde{Z}_{g,\text{FUSION}} = \frac{\sum_{l=1}^{m}(\widehat{w}_l Z_l)}{\sqrt{\widehat{\mathbf{w}}'\mathbf{V}\widehat{\mathbf{w}}}}, \quad \mathbf{V} = \text{Corr}(\mathbf{G}_0)$$

$$\tilde{Z}_{g,\text{S-PrediXcan}} = \frac{\sum_{l=1}^{m}(\widehat{w}_l\widehat{\sigma}_l Z_l)}{\sqrt{\widehat{\mathbf{w}}'\mathbf{V}\widehat{\mathbf{w}}}}, \quad \widehat{\sigma_l^2} = \text{Var}(\mathbf{G}_{0,l}), \quad \mathbf{V} = \text{Cov}(\mathbf{G}_0).$$

- Existing TWAS tools all assume SNP effect sizes on phenotype $\beta_i = \gamma\widehat{w}_i,\ i = 1,\cdots,m.$
- Burden test: $H_0 : \gamma = 0$



$\widehat{\sigma_l^2}$: genotype variance
$\mathbf{G}_0$: reference genotype data
These two Z-score TWAS statistics are equivalent when eQTL weights were derived from standardized expression traits and cis-SNP genotype data.

Gamazon ER et. al., Nat Genetics, 2015

Variance Component TWAS

- Tests if the phenotype variance component due to $GReX_g$ is non-zero:

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

$$\beta_i \sim N(0, \widehat{w}_i^2 \tau)$$

- Estimated eQTL effect sizes $\widehat{w}_i$ from the reference panel will be taken as SNP weights
- Variance Component test: $H_0 : \tau = 0$

- $\widehat{\mu}$ denotes the phenotype mean under the null model

$$Q = (\mathbf{Y} - \widehat{\mu})\mathbf{K}(\mathbf{Y} - \widehat{\mu}),\ \mathbf{K} = \mathbf{XWX}'$$

$$\mathbf{W} = diag(\widehat{w_1}^2, \cdots, \widehat{w_m}^2)$$

- Test statistic Q follows a mixture of chi-square distributions under the null hypothesis
- P-value can be easily calculated by Davies exact method as used by SKAT
- P-value calculation is also derived for using GWAS summary statistics

Tang S. et. al. PLOS Genetics, 2021.
https://doi.org/10.1371/journal.pgen.1009482

Variance Component TWAS with GWAS Summary Data

Assume the phenotype mean $\hat{\boldsymbol{\mu}}$ under $\boldsymbol{H_0}$ is 0, the test statistic is given by

$$Q = \sum_{j=1}^{m} w_j^2 s_j^2$$

- $s_j = \boldsymbol{X'_{\cdot j}Y}/\widehat{\sigma_Y}^2$ is the single variant score statistic of the j^{th} variant
- Numerator of s_j which can be estimated by

$$\boldsymbol{X'_{\cdot j}Y} = (n-1)\widehat{\beta}_j \Sigma_{j,j}$$

- Denominator of s_j is the estimated phenotype variance $\widehat{\sigma_Y}^2$, which can be estimated by,

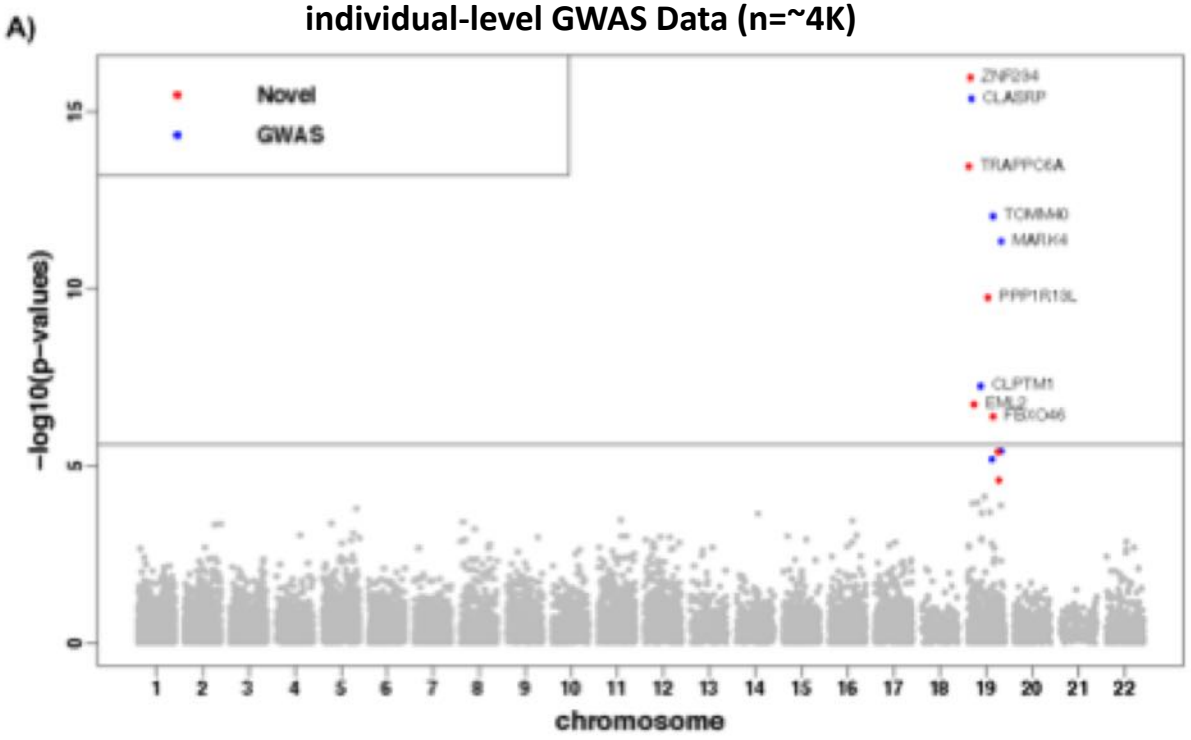
$$\widehat{\sigma_Y}^2 = median\left(\Sigma_{j,j}\widehat{\sigma_j^2}(n-1) + \Sigma_{j,j}\hat{\beta}_j{}^2; j = 1, \dots, m\right)$$

Tang S. et. al. PLOS Genetics, 2021.
https://doi.org/10.1371/journal.pgen.1009482

TWAS of Alzheimer's Disease by TIGAR

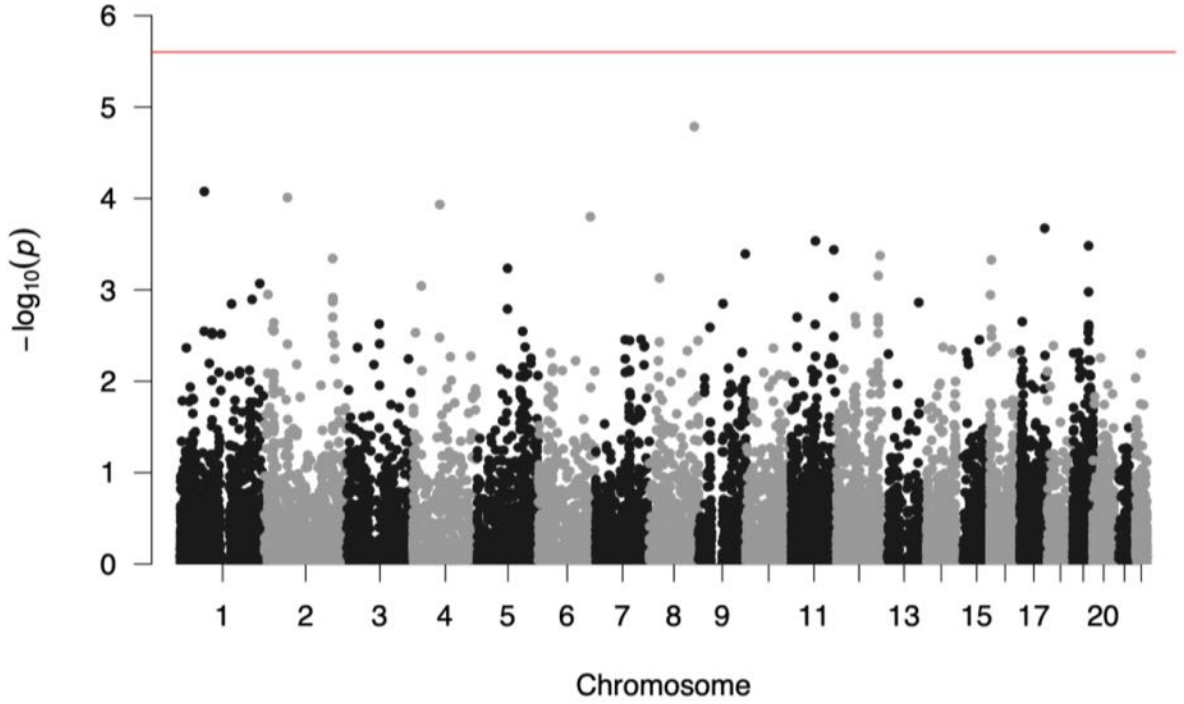
- eQTL weights trained by Bayesian DPR model with reference transcriptomic data of brain tissue were used
- Using ~4K individual-level GWAS data

Variance Component TWAS with individual-level GWAS Data (n=~4K)



Tang S. et. al. PLOS Genetics, 2021.
https://doi.org/10.1371/journal.pgen.1009482

Burden TWAS with individual-level GWAS Data (n=~4K)

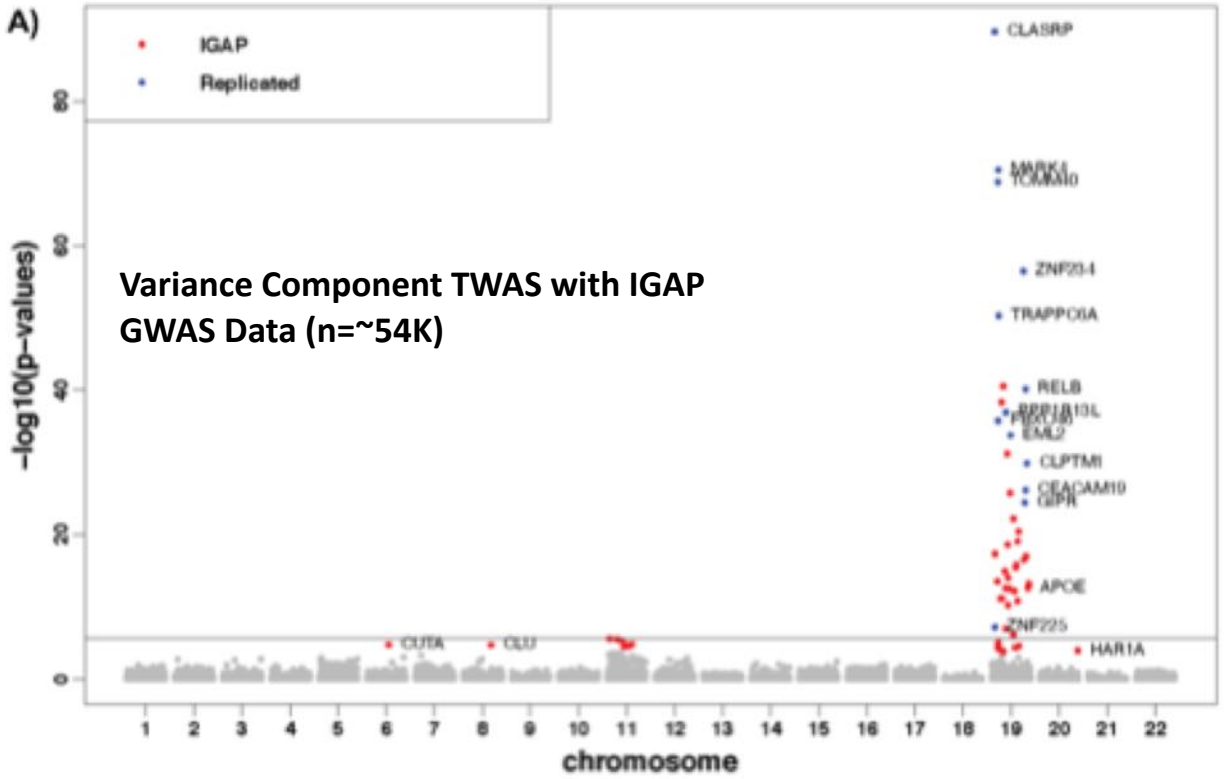


Luningham J.M. et. al. AJHG, 2020.
https://doi.org/10.1016/j.ajhg.2020.08.022

TWAS of Alzheimer's Disease by TIGAR

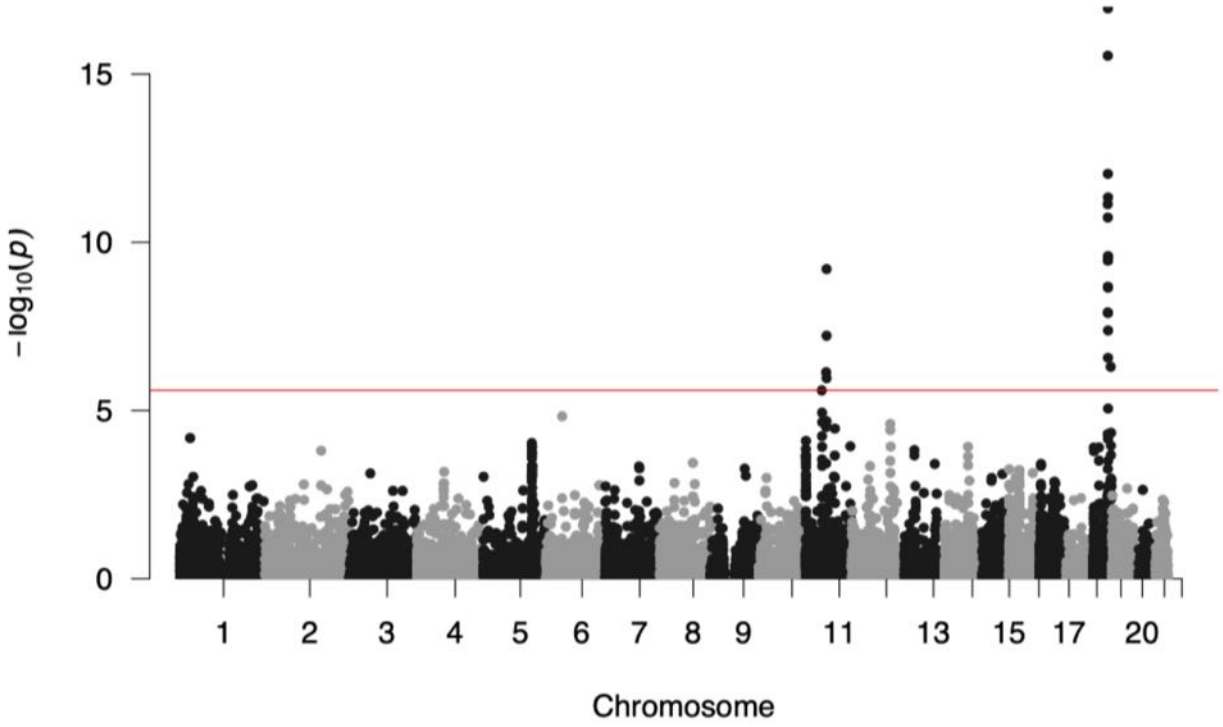
- eQTL weights trained by Bayesian DPR model with reference transcriptomic data of brain tissue were used
- Using IGAP summary-level GWAS data (n=~54K)

Variance Component TWAS



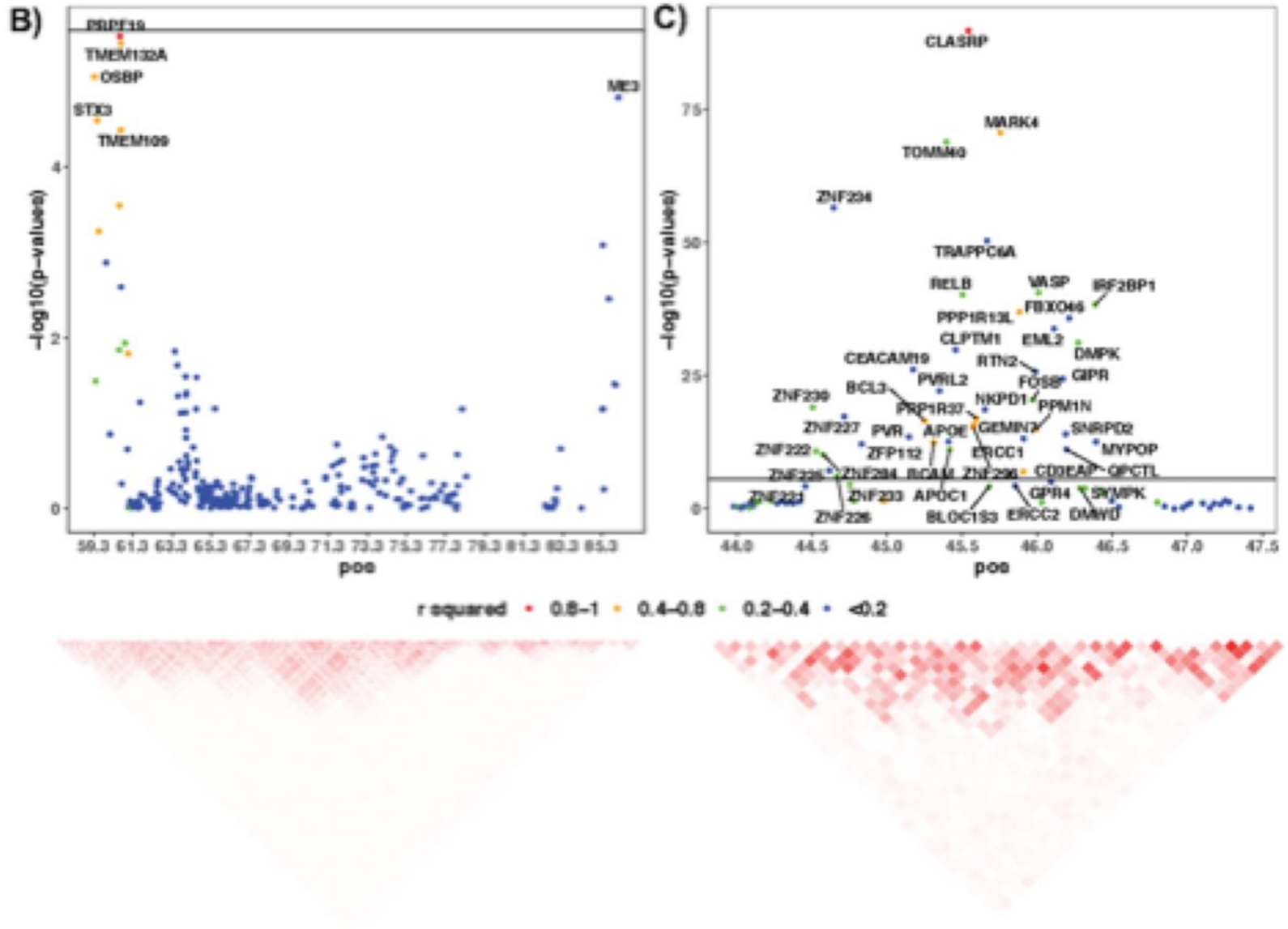
Tang S. et. al. PLOS Genetics, 2021.
https://doi.org/10.1371/journal.pgen.1009482

Burden TWAS



Luningham J.M. et. al. AJHG, 2020.
https://doi.org/10.1016/j.ajhg.2020.08.022

Locus Zoom Plot of TWAS Loci of Alzheimer's Disease by TIGAR



Tang S. et. al. PLOS Genetics, 2021. https://doi.org/10.1371/journal.pgen.1009482

TWAS vs. Mendelian Randomization (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 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 eQTL per gene by a multivariate regression model

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.

Test Methods

- Inverse-Variance Weighted (IVW) method : Estimate the ratio between SNP-outcome association and SNP-exposure association using a meta-analysis approach
- MR-Egger Regression (Bowden et al., 2015) : Weighted linear regression of the SNP-outcome coefficients on the SNP-exposure coefficients. Without an intercept term in the regression model, MR-Egger slope estimate will equal to the IVW estimate.
- SMR & GSMR (Zhu et. al., 2016, Zhu et. al., 2018):

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 $\hat{\beta}_{X_jY} \sim \hat{\beta}_{X_jM}, j = 1, \ldots, J$

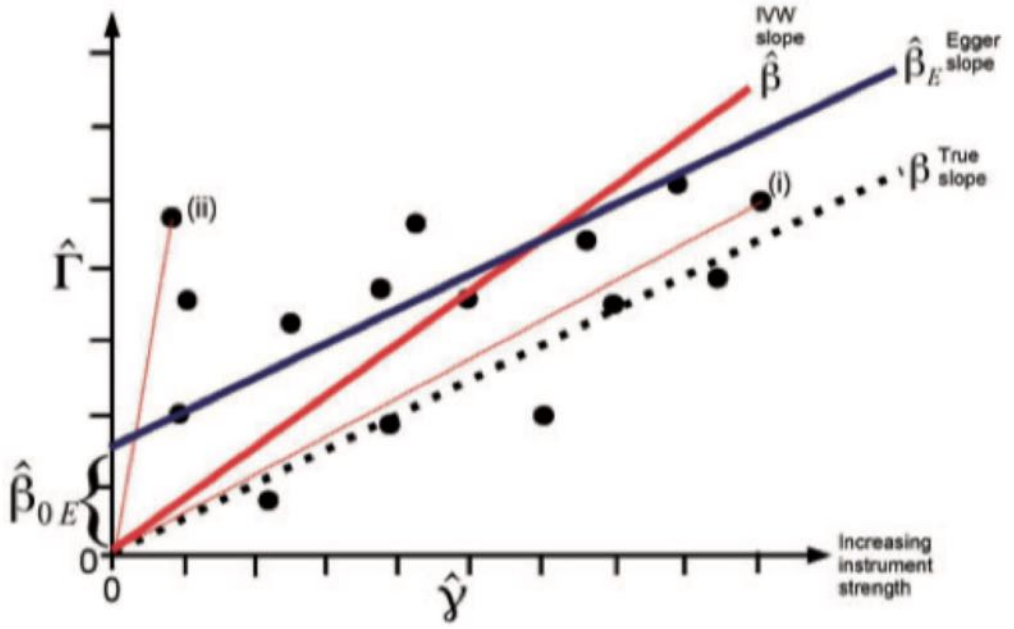


Figure 2. Plot of the gene–outcome ($\hat{\Gamma}$) vs gene–exposure ($\hat{\gamma}$) regression coefficients for a fictional Mendelian randomization analysis with 15 genetic variants. The true slope is shown by a dotted line, the inverse-variance weighted (IVW) estimate by a red line, and the MR-Egger 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)
- $\mathrm{H}_0\colon \gamma = 0$

Yuan Z. et. al. Nature Communications, 2020.

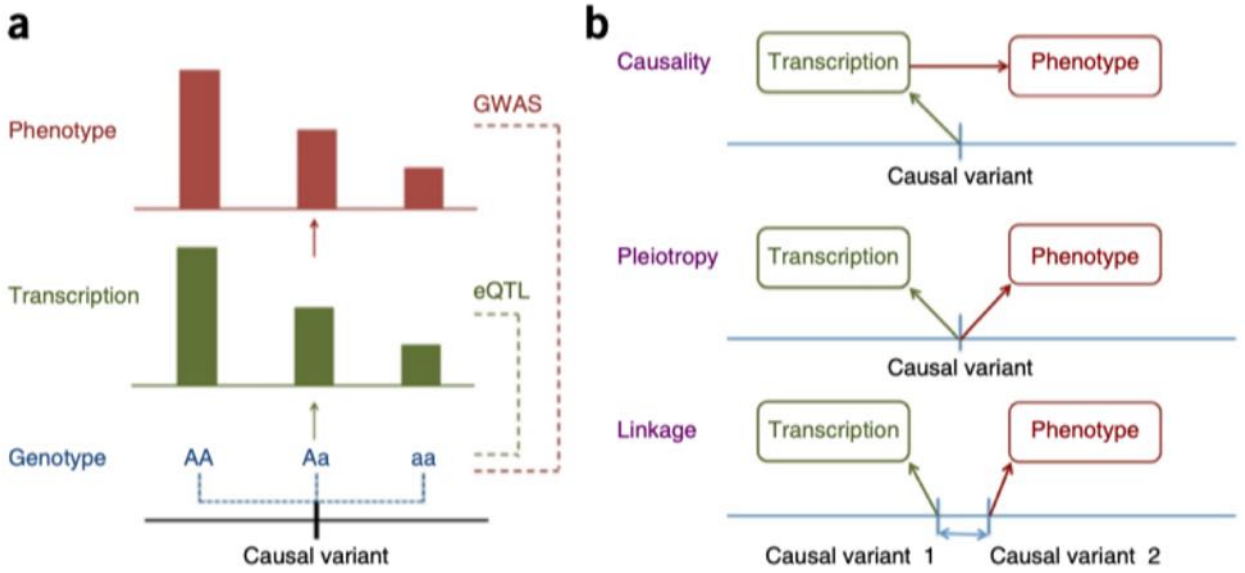


Figure 1 Association between gene expression and phenotype through genotypes. (**a**) A model of causality where a difference in phenotype is caused by a difference in genotype mediated by gene expression (transcription). (**b**) Three possible explanations for an observed association between a trait 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 or α
- Probabilistic as estimating and testing in a maximum likelihood framework
- See details in the Supplementary Notes of Zhu et. al., Nature Genetics, 2016.

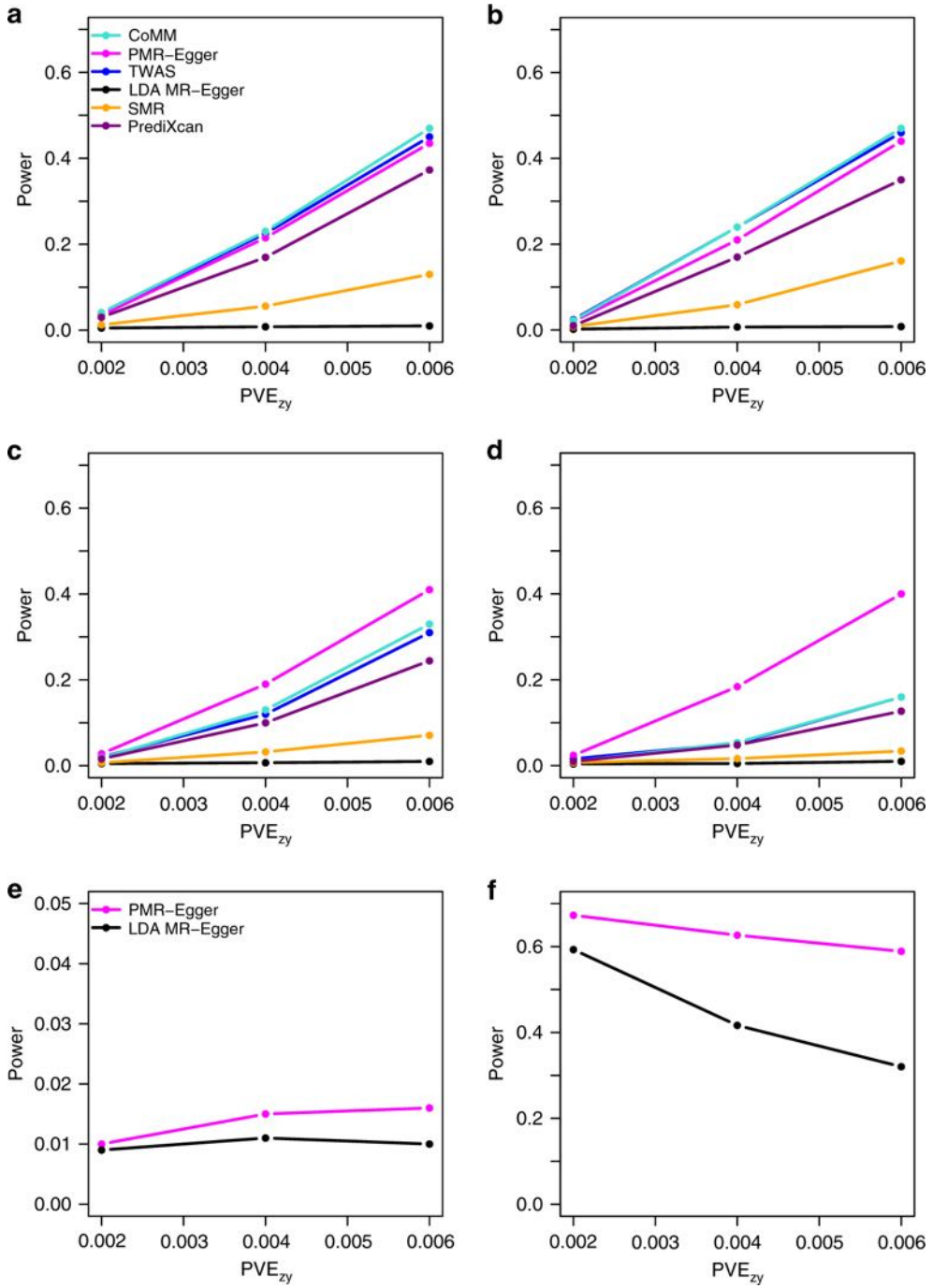


Fig. 2: Power of different methods under various simulation scenarios.
Power (*y* axis) at a false discovery rate of 0.1 to detect the causal effect (**a**–**d**) or the horizontal pleiotropic effect (**e**, **f**), plotted against different causal effect size characterized by PVE_{zy} (*x* axis).

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** $\gamma=0.001$.

Zhu et. al., Nature Genetics, 2016.

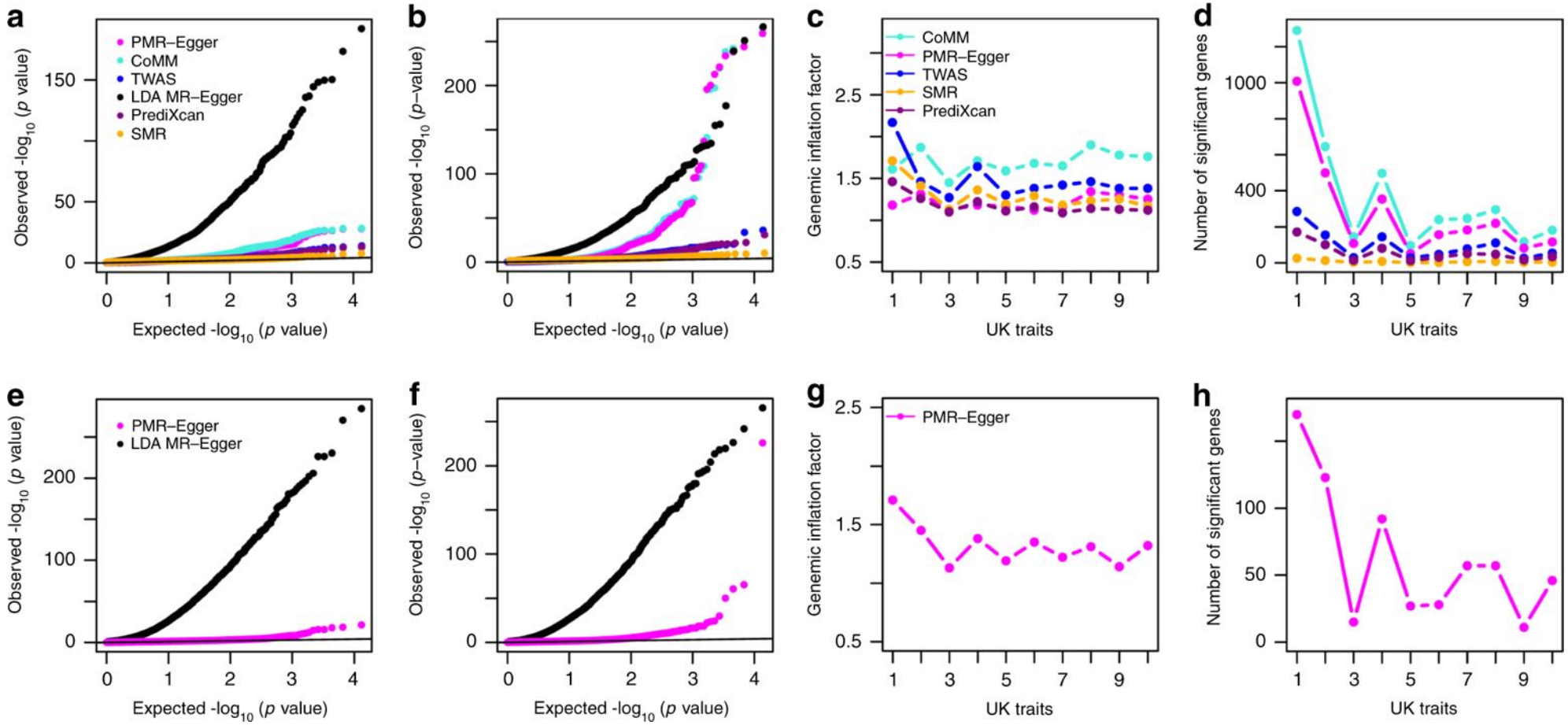


Fig. 6: TWAS analysis results by different methods for UK Biobank traits. a QQ plot for testing the causal effect for BMI. **b** QQ plot for testing the causal effect for platelet count. **c** Genomic inflation factor for testing the causal effect for each of the 10 traits by different methods. **d** 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 : Limitations

- Assume equal horizontal pleiotropic effect for all test SNPs
- Developed for continuous traits
- Computationally more expensive than two-stage TWAS methods

TWAS and MR

- Uses both eQTL (transcriptomic) and GWAS data that might be profiled for two independent cohorts
- Standard Two-stage TWAS: Test **association (pleiotropy or causality)** between multiple-SNPs-per-Gene and phenotype of interest, taking eQTL effect sizes as variant weights
- 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)

Ongoing Research Topics

- Transfer learning cross samples of different ethnicities
- Multi-tissue reference transcriptomic data (MultiXcan, https://doi.org/10.1371/journal.pgen.1007889)

TWAS Tools

- MetaXcan (PrediXcan, S-PrediXcan, MultiXcan, S-MultiXcan)
  - https://github.com/hakyimlab/MetaXcan
- TIGAR
  - https://github.com/yanglab-emory/TIGAR
- PMR-Egger/moPMR-Egger
  - https://github.com/yuanzhongshang/PMR