Quantifying genetic effects on disease mediated by assayed gene expression levels

Douglas W. Yao, Luke J. O'Connor, Alkes L. Price, and Alexander Gusev
Nature Genetics Volume 52, Number 6, pp. 626-633 (June 2020)

Presentation by Mykhaylo M. Malakhov

GWAS: a historical overview

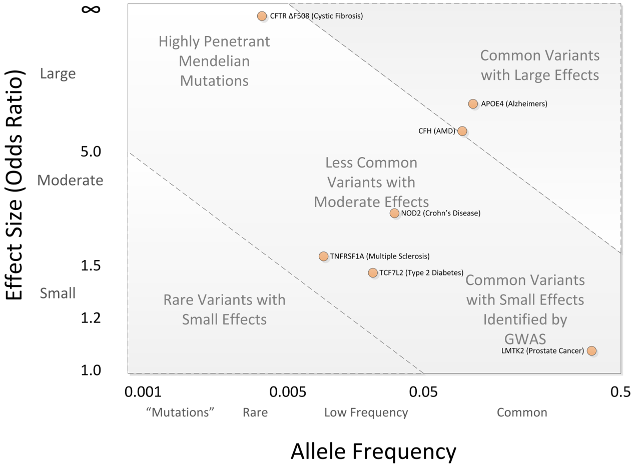


Image source: Bush WS, Moore JH (2012) PLOS Computational Biology 8(12): e1002822.

From genetics to molecular mechanisms

DNA (e.g. SNPs) → Gene expression (e.g. mRNA) → Protein → Biological activity

- Most GWAS hits fall in non-coding regions
- Understanding the functional pathways by which SNPs affect phenotypes can provide targets for clinical interventions

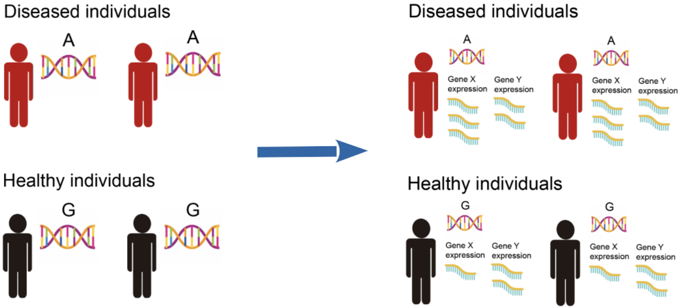
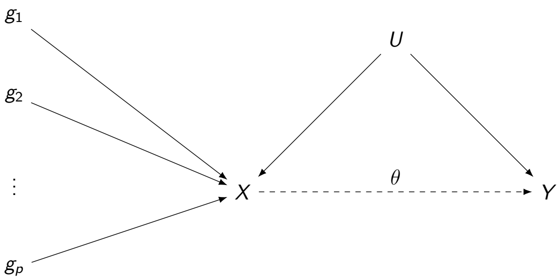


Image source: adapted from Douglas Yao's blog

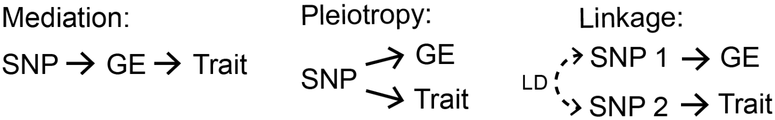
Transcriptome-wide association studies (TWAS)



1. Regress $X \sim g_1 + \cdots + g_k$ in an expression reference panel to obtain eQTL weights
2. Combine eQTL weights with GWAS summary statistics to predict $\hat{X}$
3. Regress $Y \sim \hat{X}$ to obtain (putatively) causal effect size $\hat{\theta}$

Problems with TWAS-based approaches

- The expression that is relevant to disease likely occurs in specific cell types under specific stimuli, but currently available expression data is from post-mortem samples of healthy individuals
- Widespread pleiotropy and linkage are hypothesized to exist:



Mediation:

SNP → GE → Trait

Pleiotropy:

SNP → GE

SNP → Trait

Linkage:

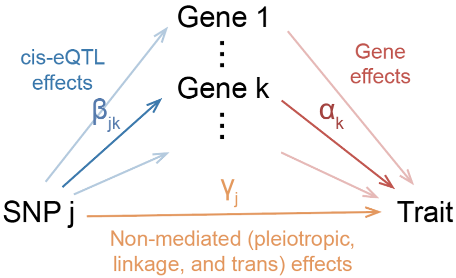
LD

SNP 1 → GE

SNP 2 → Trait

What proportion of trait heritability is mediated by gene expression levels, across all genes?

Quantifying mediation



$$\underset{\text{Total effect of SNP j on trait}}{\omega_j} = \underbrace{\sum\nolimits_k \beta_{jk}\alpha_k}_{\text{Mediated effect of SNP j}} + \underbrace{\gamma_j}_{\text{Non-mediated effect of SNP j}}$$

- β_{jk} is the change in expression of gene k in individuals carrying an allele of SNP j
- α_k is the change in the trait per unit change in expression of gene k
- γ_j is the additional change in the trait in individuals carrying an allele of SNP j that is not mediated by expression
- Gene expression is normalized across all genes

Quantifying mediation (continued)

Definition (Overall mediation)

$$h^2_{med} = \sum_j \sum_k \beta^2_{jk} \alpha^2_k$$

where β_j is scaled by the variance of SNP j's minor allele count

Definition (Heritability mediated by expression)

$$\frac{h^2_{med}}{h^2_g} = \frac{\sum_j \sum_k \beta^2_{jk} \alpha^2_k}{\sum_j \omega^2_j}$$

where h^2_g is the total SNP heritability of the trait

A technicality

- These definitions assume that gene expression is only measured in the causal cell types and/or cellular contexts for the trait of interest, so actually $h^2_{med} = h^2_{med;causal}$
- However, it is not known which cell types/contexts are causal and only tissue-level expression data is available

Definition (Mediation by assayed expression)

$$h^2_{med;assayed}(T) = r^2_g(T) h^2_{med;causal}$$

where T is the set of assayed tissues and $r^2_g(T)$ is the squared genetic correlation between expression in T and expression in the causal contexts averaged across all genes

cis vs *trans* regulation of expression

More precisely, this paper focuses on estimating the proportion of heritability that is mediated by the *cis* genetic component of assayed gene expression levels.

- *trans*-eQTLs have much weaker effects, which cannot be estimated even from the largest available expression reference panels
- Hence, the authors only consider *cis*-eQTLs and *cis-by-trans* eQTLs
- The proportion of heritability mediated by the entire genetic component of assayed gene expression may be higher – we simply can't know

Mediated expression score regression (MESC)

- One idea: use Mendelian randomization to estimate $\beta_{jk}\alpha_k$ for each gene k
- Better idea:

$$\sum_j \sum_k \beta_{jk}^2 \alpha_k^2 = E[\sum_j \beta_j^2] E[\alpha^2] G$$

where G is the total number of genes and expectations are taken over genes

Estimating $E[\sum_j \beta_j^2]$

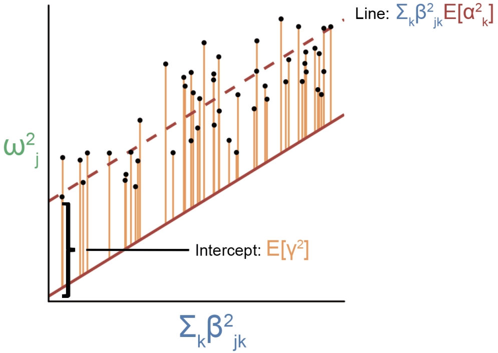
$E[\sum_j \beta_j^2]$ is the average *cis* heritability of gene expression across all genes, so it can be estimated using standard methods:

- The authors used REML as implemented in the genome-wide complex trait analysis (GCTA) software
- Linkage disequilibrium (LD) score regression could be used instead

Estimating $E[\alpha^2]$

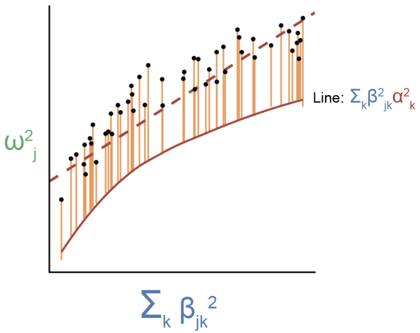
$$E[\omega_j^2|\beta_{1j},\cdots,\beta_{kj}] = E[\alpha^2]\sum_{k=1}^{G}\beta_{jk}^2 + E[\gamma^2]$$

Note: in practice we only have marginal estimates of ω_j and β_{jk}. To correct for LD, also include the LD score of the SNP as a covariate.



Assumption 1

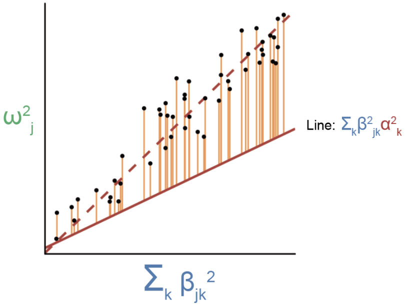
β_{jk}^2 must be uncorrelated with α_k^2



- Likely violated in practice (large-effect genes tend to have weak eQTLs)
- Can be mitigated by splitting genes into bins with approximate independence, and then estimating $E[\alpha^2]$ in each bin

Assumption 2

β_{jk}^2 must be uncorrelated with γ_j^2



- Likely violated in practice (biologically active genome regions have both larger expression-mediated and non-expression-mediated effects vs inactive regions)
- Can be mitigated by splitting SNPs into bins according to the baselineLD annotations, and then estimating $E[\alpha^2]$ in each bin

Assumption 3

LD scores must be uncorrelated with both α_k^2 and γ_j^2

- Likely violated in practice (Gazal et al. 2017 showed that LD is correlated with causal effect size of SNP)
- Can be mitigated by splitting SNPs into bins according to the baselineLD model, which takes MAF and other LD-associated metrics into account

Assumption 4

There is no sampling noise in eQTL effect size estimates

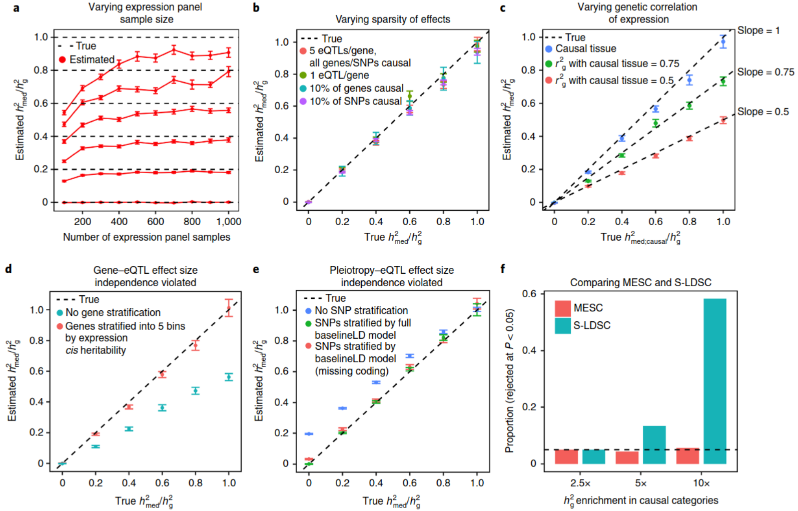
- Certainly violated in practice (we only have finite gene expression samples)
- Gusev et al. 2016 showed that for samples of over 500 individuals, there is negligible noise
- In smaller samples, violation will downwardly bias $r_g^2(T)$
- Not an issue if the goal is to estimate expression-mediated heritability for your specific gene expression data set rather than in general

Assumption 5

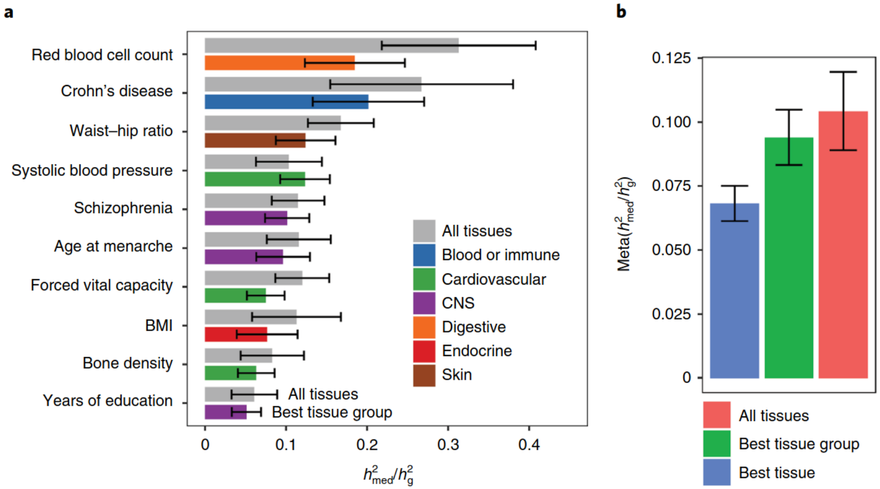
Expression-mediated effects of SNPs on the trait are linear

- Likely violated in practice
  - Lin, Xue, Malakhov, et al. 2022 provide evidence for non-linear eQTL effects using the TWAS framework
- Might not be an issue since most genes have a single lead eQTL
- Authors suggest that this violation can be mitigated by binning genes into categories with approximately linear effects, but they do not perform this binning

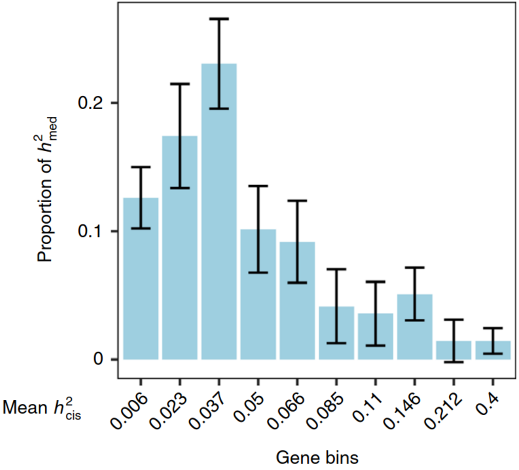
Simulation results



Estimates of $h^2_{med;assayed}$ in GTEx



Low h^2_g genes have high $h^2_{med}(D)/h^2_{med}$



Upshot: Using eQTL effect sizes estimated via meta-analysis from 48 human tissues, the average h^2_{med}/h^2_g across 42 independent traits is 0.11 ± 0.02. Of those 42 traits, only 10 had significantly nonzero h^2_{med}/h^2_g estimates ($P < 0.05/42$).

Possible interpretations

- SNPs might primarily affect phenotypes by changing protein-coding sequences, through post-transcriptional modifications, or through post-translational modifications instead of by regulating gene expression
- SNP effects on phenotypes might be mediated by weak effects on the expression of distant genes, which are not detectable from currently available gene expression panels (i.e. expression mediated by *trans*-eQTLs)
- SNP effects on phenotypes might be mediated by gene expression, but only in specific cell types and specific disease state or cellular contexts

Questions?