Summer Presentation(2):

MESC Regression

Name: Chenxiao Tian

Date: 2022/7/18

OUTLINE

- 1.Motivation
- 2. Method and Derivations:
- (1)Definition of expression-mediated heritability (both the casual case and the assayed case)
- (2) Unstructured Case:
- idealized casual case, idealized assayed case, summary statistics casual case(a special case of unstructured LD regression), summary statistics assayed case
- (3) Population Stratification Case:
- the casual case(a special case of structured LD regression), the assayed case
- (4) Estimation of the expression scores
- 3.Simulations and Real Data
- 4.Conclusion and Discussion

We model trait $\mathbf{y}$ for N individuals as follows:

$$\mathrm{y} = \mathrm{X}\gamma + \mathrm{XB}\alpha + \epsilon \quad (1)$$

where $\mathbf{y}$ is an N-vector of phenotypes (standardized to mean 0 and variance 1), $\mathbf{X}$ is an $N \times M$ genotype matrix for M SNPs (standardized to mean 0 and variance 1), $\boldsymbol{\gamma}$ is an M vector of non-mediated SNP effect sizes on the trait (including pleiotropic, linkage, and trans-eQTL-mediated effects), $\mathbf{B}$ is an $M \times G$ matrix of cis-eQTL effect sizes *in the causal cell types/contexts* for G genes, $\boldsymbol{\alpha}$ is a G-vector of causal gene expression effect sizes on the trait, and $\boldsymbol{\epsilon}$ is an N-vector of environmental effects. We treat all variables as random. We define $h^2_{med;\,causal}$ as follows:

$$\phi = X\beta + \epsilon$$

Under the generative model (1), the total effect of SNP k on the complex trait is

$$\omega_k = \sum_i^G \beta_{ik}\alpha_i + \gamma_k$$

Given conditional independence of α and γ given β, upon squaring ω_k we have

$$E\left[\omega_k^2 \mid \beta_{1k}\ldots\beta_{ik}\right] = \sum_i^G E\left[\alpha_i^2 \mid \beta_{1k}\ldots\beta_{ik}\right]\beta_{ik}^2 + E\left[\gamma_k^2 \mid \beta_{1k}\ldots\beta_{ik}\right]$$

1.Motivation

- Background：
- In the past decade, genome-wide association studies (GWAS) have shown that most disease associated variants lie in noncoding regions of the genome , leading to the hypothesis that ：
- Regulation of gene expression levels is the primary biological mechanism through which genetic variants affect complex traits, and motivating large scale expression quantitative trait loci (eQTL) studies.

1.Motivation

- Many statistical methods have been developed to integrate eQTL data with GWAS data to gain functional insight into the genetic architecture of disease：
- Eg1： Colocalization Tests：

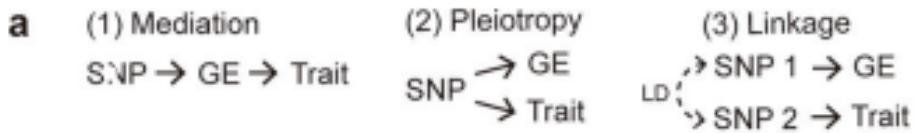
  Have shown that many genes have eQTLs that colocalize with GWAS loci
- Eg2： Transcriptome-wide association studies：
- Have shown that many genes exhibit significant cis-genetic correlations between their expression and disease.
- Eg3： Partitioning of disease heritability：
- Have shown that eQTLs as a whole are significantly enriched for disease heritability

1.Motivation

- Challenge and Problems：
- 1.It remains unclear the extent to which eQTLs from available studies capture mechanistic effects of gene expression on disease.
- 2.EQTLs from the largest available gene expression reference panels are measured in bulk tissues in steady-state cellular conditions, which may not reflect the specific cell types or cellular contexts in which gene expression is causal for disease.

1.Motivation

- 3.In addition, several different causal scenarios can result in similar patterns of enrichment/overlap between GWAS loci and eQTLs：
- (1) mediation, (2) pleiotropy, and (3) linkage.
- Of these three scenarios, only scenario (1) is informative of the SNP's mechanism of action on disease, but existing methods are unable to consistently distinguish scenarios (2) and (3) from scenario (1).



- Note：GE：Gene Expression Level

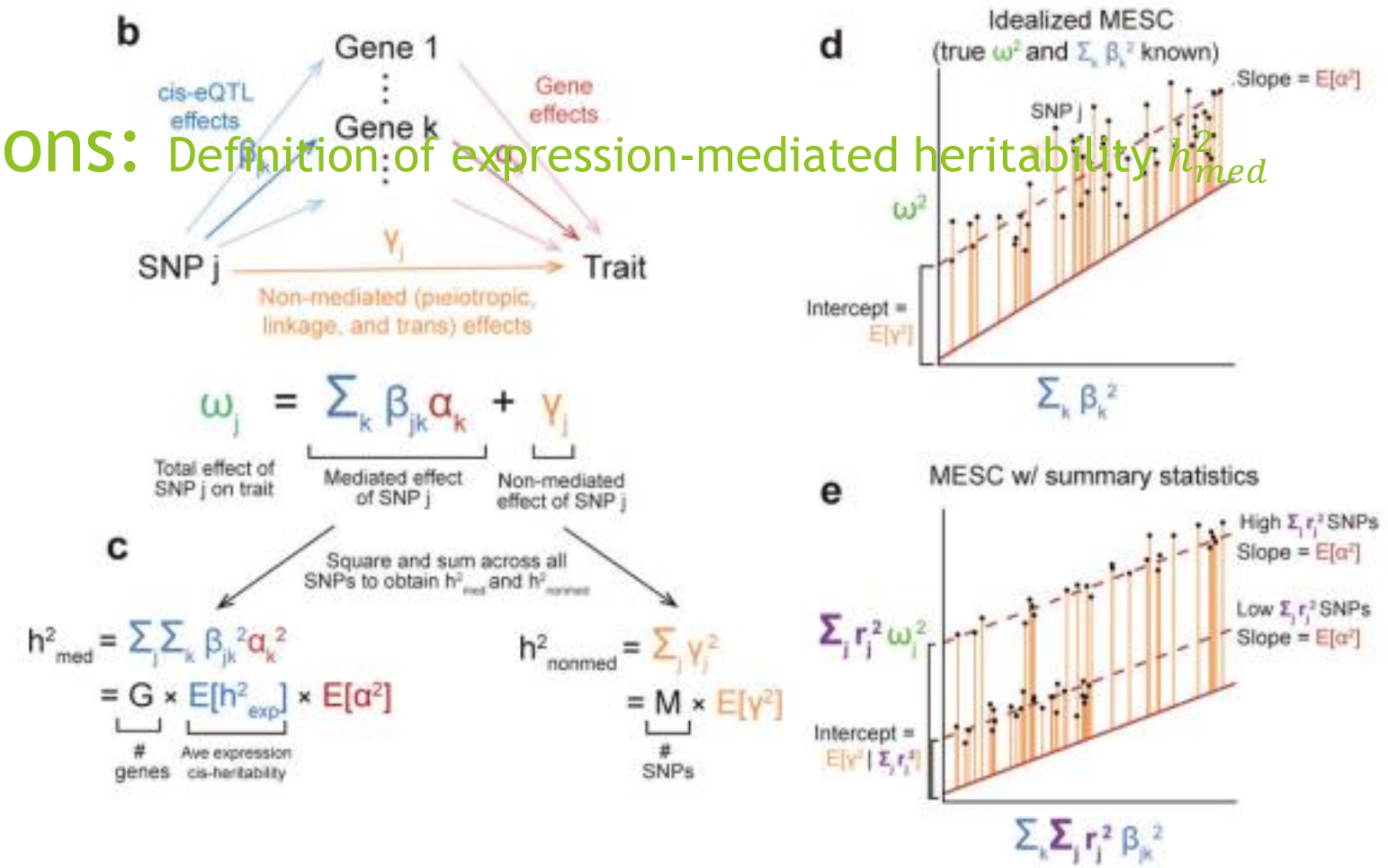
1.Motivation

- Main Goal：
- We aim to quantify the proportion of disease heritability mediated in cis by assayed expression levels (scenario (1) from above).

2.Method and Derivations: Definition of expression-mediated heritability h^2_{med}

- 1. Definition of expression-mediated heritability h^2_{med}

- (b) Total SNP effect sizes are modeled as the sum of a mediated component (defined as causal cis-eQTL effect sizes β multiplied by gene-trait effect sizes α) and a non-mediated component γ.



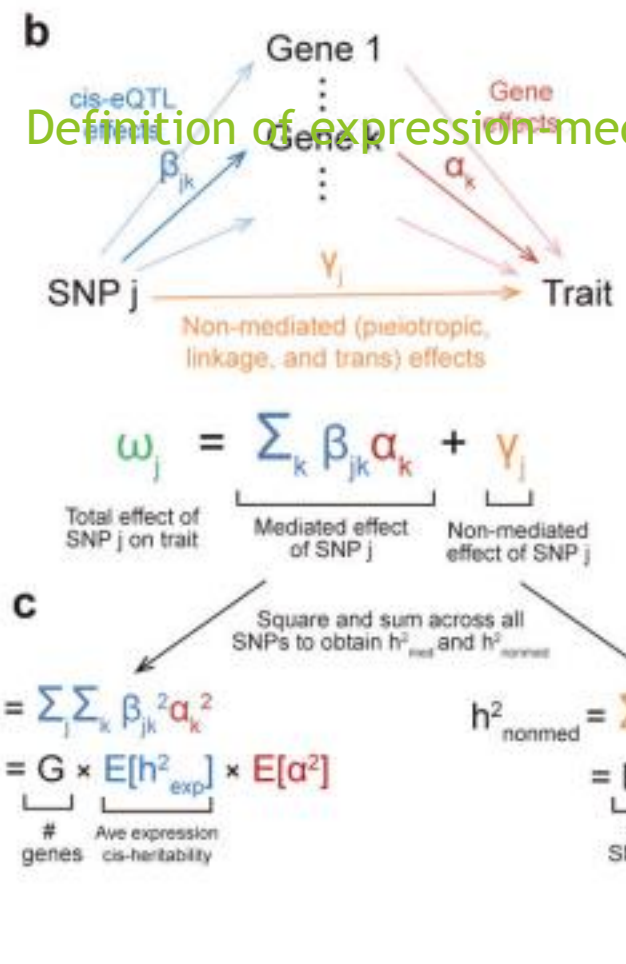
2.Method and Derivations: : Definition of expression-mediated heritability h^2_{med}

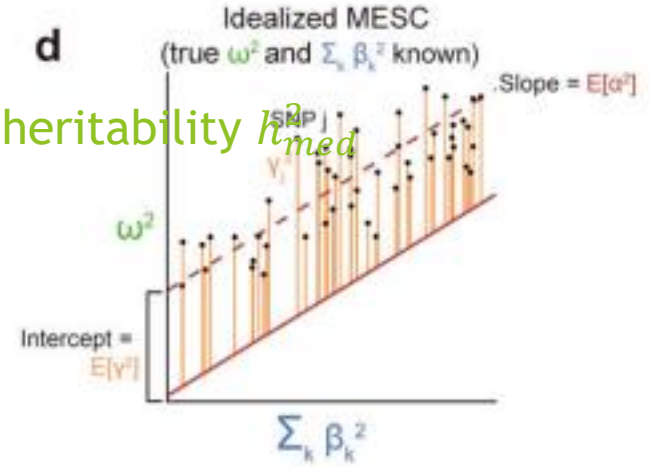
- 1. Definition of expression-mediated heritability h^2_{med}

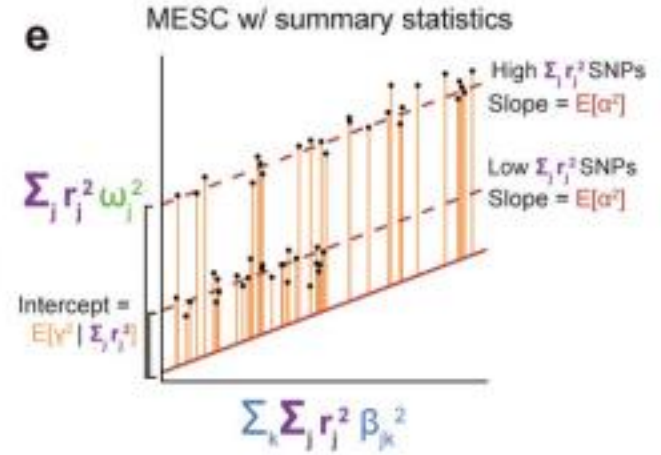
(c)

A. Heritability mediated by the cis-genetic component of gene expression levels (h^2_{med}) is defined as the squared mediated component of SNP effect sizes summed across all SNPs (assuming that genotypes and phenotypes are standardized)

B. h^2_{med} can also be rewritten as the product of the number of genes G, the average expression cis-heritability E[h^2_{cis}], and the average gene-trait effect size E[α^2]





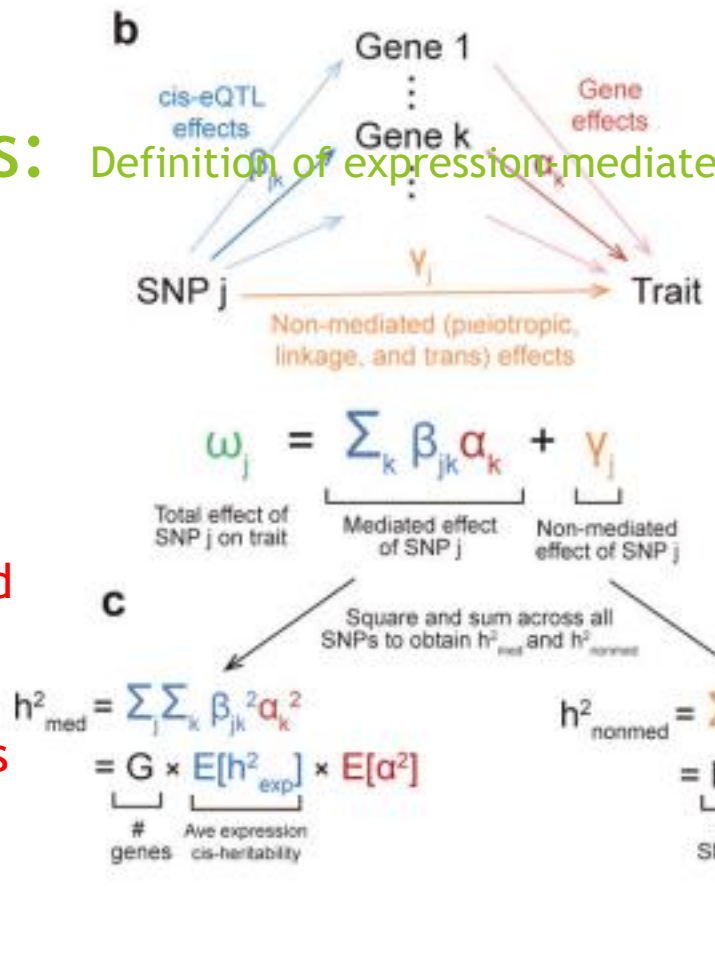


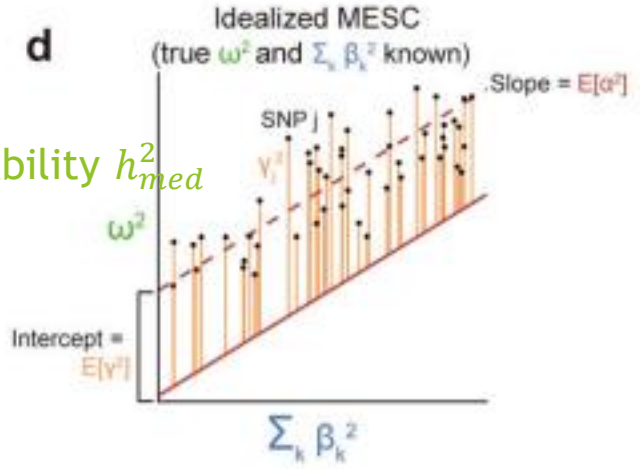
2.Method and Derivations:

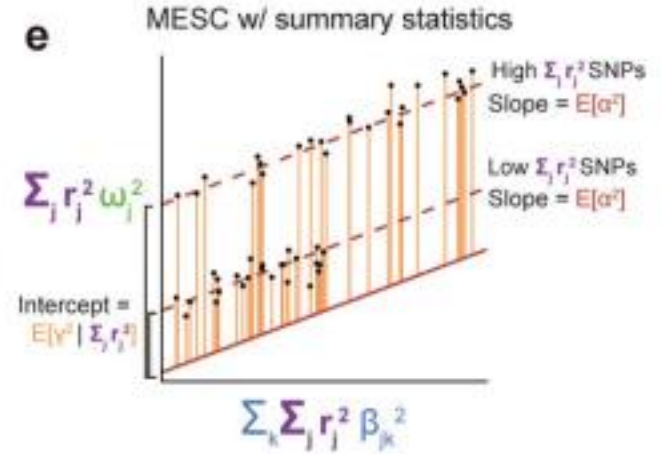
Definition of expression-mediated heritability h^2_{med}

- 1. Definition of expression-mediated heritability h^2_{med}

(d) The basic premise behind MESC is to regress squared GWAS effect sizes on squared eQTL effect sizes. Non-directional non-mediated effects are captured by the intercept, while directional mediated effects are captured by the slope, which equals E[α2] given appropriate effect size independence assumptions (see Methods).



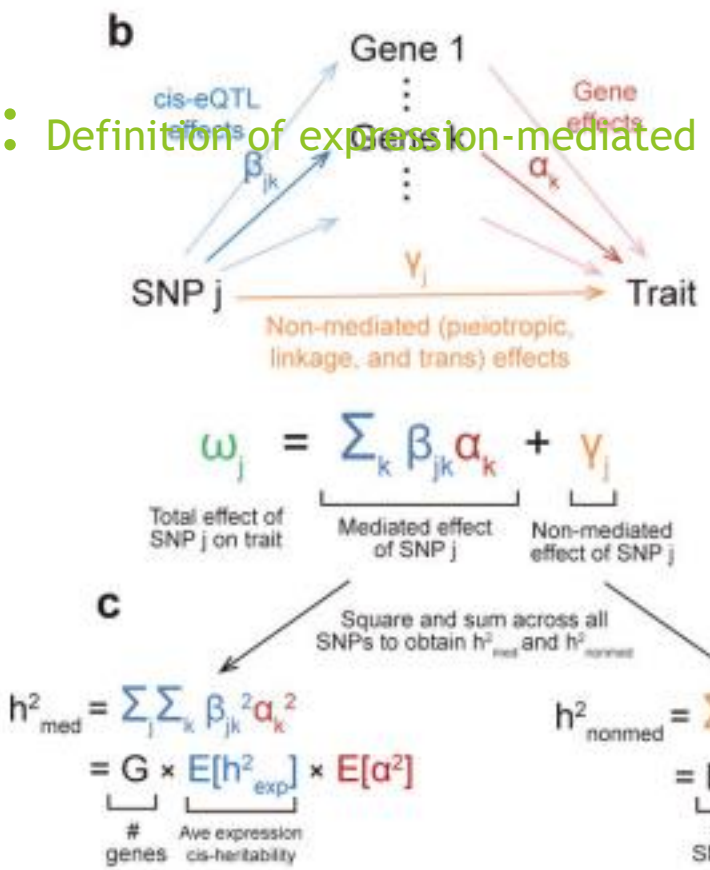


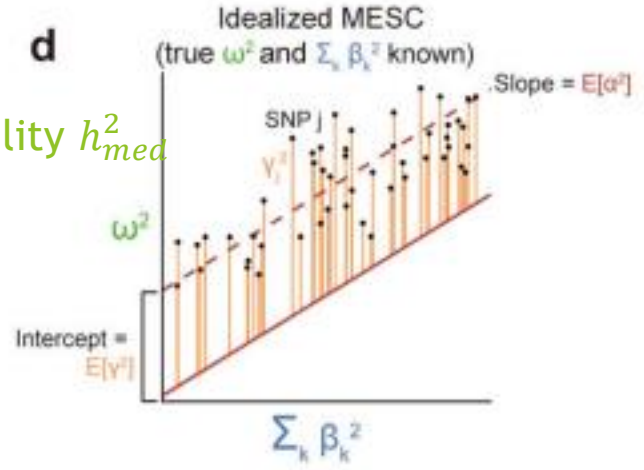


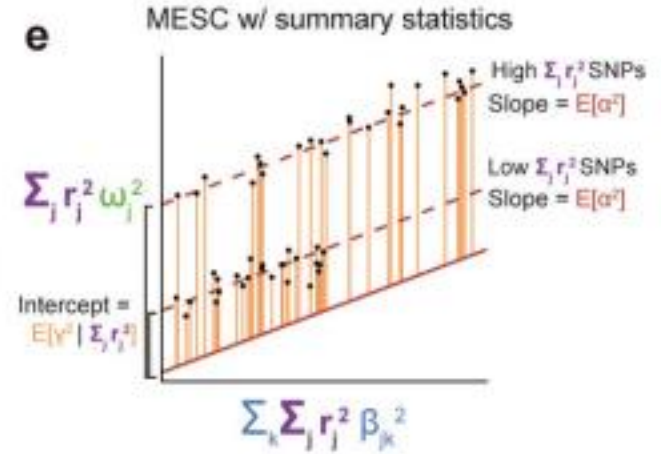
2.Method and Derivations: Definition of expression-mediated heritability h^2_{med}

- 1. Definition of expression-mediated heritability h^2_{med}

(e) In practice, MESC involves regressing squared GWAS summary statistics on squared eQTL summary statistics.







2.Method and Derivations: : Definition of expression-mediated heritability h^2_{med}

- 1. Definition of expression-mediated heritability h^2_{med}
- $h^2_{med,casual}$ and $h^2_{med,assayed}$
- $h^2_{med,casual}$ in which cis-eQTL effect sizes are hypothetically obtained in the causal cell types and contexts for the disease(More in theory speaking)
- $h^2_{med,assayed}$: in which cis-eQTL effect sizes are obtained in a given set of assayed tissues T (e.g. from GTEx, More in practical speaking)

2.Method and Derivations : Definition of expression-mediated heritability h^2_{med}

- 1. Definition of expression-mediated heritability h^2_{med}
- $h^2_{med,casual}$ and $h^2_{med,assayed}$
  - Relationship:

$$h^2_{med;\,assayed}(T) = r_g^2(T) h^2_{med;\,causal},$$

  - r_g^2(T) is the average squared genetic correlation between expression in T and expression in the unobserved causal cell types/contexts for the disease.

  - In practice, we only aim to estimate $h^2_{med,assayed}$ (T)
  - $h^2_{med,casual}$ has a more direct mechanistic interpretation
  - For brevity, we refer to $h^2_{med,assayed}$ (T) as simply h^2_{med} for the remainder, where the set of tissues T is implicit
  - h^2_{med}(D):

We also define a quantity $h^2_{med}(D)$ corresponding to the heritability mediated by the expression levels of gene category D, where D can be arbitrarily defined over any set of genes (e.g. genes in a specific molecular pathway). See Methods for a more detailed definition of h^2_{med} and $h^2_{med}(D)$.

3.Method and Derivations: Unstratified MESC

(1) Mathematical Model and Mathematical Definition of h^2_{med}

We model trait $\mathbf{y}$ for N individuals as follows:

$$\mathrm{y} = \mathrm{X}\gamma + \mathrm{XB}\alpha + \epsilon \qquad (1)$$

where $\mathbf{y}$ is an N-vector of phenotypes (standardized to mean 0 and variance 1), $\mathbf{X}$ is an $N \times M$ genotype matrix for M SNPs (standardized to mean 0 and variance 1), $\boldsymbol{\gamma}$ is an M vector of non-mediated SNP effect sizes on the trait (including pleiotropic, linkage, and trans-eQTL-mediated effects), $\mathbf{B}$ is an $M \times G$ matrix of cis-eQTL effect sizes *in the causal cell types/ contexts* for G genes, $\boldsymbol{\alpha}$ is a G-vector of causal gene expression effect sizes on the trait, and $\boldsymbol{\epsilon}$ is an N-vector of environmental effects. We treat all variables as random. We define

3.Method and Derivations: Unstratified MESC

(1) Mathematical Model and Mathematical Definition of h^2_{med}

$\boldsymbol{\epsilon}$ is an N-vector of environmental effects. We treat all variables as random. We define $h^2_{med;\,causal}$ as follows:

$$h^2_{med;\,causal} = Var[\mathbf{XB\alpha}]$$

Assume 1: We treat all variables as random

3.Method and Derivations: Unstratified MESC

- (1) Mathematical Model and Mathematical Definition of h^2_{med}
- Under the Assume 2:α and β (also X、ε) are independent of each other
- Also since X has been standardized to mean 0 and variance 1, so the second line above follows the first because E[XBα | B, α] = 0
- Where E[α^2] is the average squared per-gene effect of expression on trait and E[h^2_{cis}] is the average cis-heritability of expression across all genes.

Under the assumption that α and β are independent of each other, we can rewrite this as follows:

$$h^2_{med;\,causal} = E_{B,\alpha}[Var[\mathbf{XB\alpha} \mid \mathbf{B}, \boldsymbol{\alpha}]] + Var_{B,\alpha}[E[\mathbf{XB\alpha} \mid \mathbf{B}, \boldsymbol{\alpha}]]$$
$$= E_{B,\alpha}\left[\sum_i^G \sum_j^M \beta^2_{ij}\alpha^2_i\right]$$
$$= GE\left[\alpha^2\right]E[h^2_{cis}]$$

3.Method and Derivations: Unstratified MESC

- (2) Mathematical Model and Mathematical Definition of $h^2_{nonmed;casual}$
- We define $h^2_{nonmed;casual}$ in a similar fashion:

$$h^2_{nonmed;causal} = Var[\mathbf{X}\boldsymbol{\gamma}] = ME[\gamma^2]$$

where $E[\gamma^2]$ is the average squared per-SNP effect on trait that is not mediated by gene expression. We consider additional expression causality scenarios, such as reverse

3.Method and Derivations: Unstratified MESC

- (2) Mathematical Model and Mathematical Definition of $h^2_{nonmed;assayed}$(T)
- In practice, expression levels in causal cell types/contexts for the complex trait are likely not assayed. Given a set of assayed tissues T (which may or may not be causal for the complex trait), we define $h^2_{nonmed;assayed}$(T) as follows:

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

while we define $h^2_{nonmed;assayed}(T)$ as $h^2_{nonmed;causal} + \left(1 - r^2_g(T)\right) h^2_{med;\,causal}$. Here,

3.Method and Derivations: Unstratified MESC

$$h^2_{med;\,assayed}(T) = r_g^2(T)h^2_{med;\,causal}$$

while we define $h^2_{nonmed;assayed}(T)$ as $h^2_{nonmed;causal} + \left(1 - r_g^2(T)\right)h^2_{med;\,causal}$. Here,

- (2) Mathematical Model and Mathematical Definition of $h^2_{nonmed;assayed}$(T)
- Where r_g^2(T) denotes the average squared genetic correlation between expression in assayed tissues T vs. in causal cell types/contexts

  Where β_i' represents cis-eQTL effect sizes on gene i in assayed tissues T.

$r_g^2(T) = \frac{1}{G}\sum_i^G \frac{Cov\left(\beta_i^2, \beta_i'^2\right)}{\sqrt{Var\left(\beta_i^2\right)Var\left(\beta_i'^2\right)}}$ and denotes the average squared genetic correlation between expression in assayed tissues T vs. in causal cell types/contexts, where β_i' represents cis-eQTL effect sizes on gene i in T. Note that β' can refer to either single tissue or meta-tissue cis-eQTL effect sizes, depending on whether T contains one or multiple tissues.

3.Method and Derivations: Unstratified MESC

We model trait $\mathbf{y}$ for N individuals as follows:

$$\mathbf{y} = \mathbf{X}\gamma + \mathbf{XB}\alpha + \epsilon \quad (1)$$

where $\mathbf{y}$ is an N-vector of phenotypes (standardized to mean 0 and variance 1), $\mathbf{X}$ is an $N \times M$ genotype matrix for M SNPs (standardized to mean 0 and variance 1), $\boldsymbol{\gamma}$ is an M vector of SNP-mediated effect sizes on the trait (including pleiotropic, linkage, and trans-eQTL-mediated effects), $\mathbf{B}$ is an $M \times G$ matrix of cis-eQTL effect sizes *in the causal cell types/contexts* for G genes, $\boldsymbol{\alpha}$ is a G-vector of causal gene expression effect sizes on the trait, and $\boldsymbol{\epsilon}$ is an N-vector of environmental effects. We treat all variables as random. We define

- (3)The idealized casual Regression Equation for Unstratified MESC
- Assume 3.(In the idealized scenario) We know that:
- 1. the true eQTL effect sizes, ß, of each SNP on each gene and
- 2. the true phenotypic effect sizes, ω, of each SNP on y. (We don't have to consider ε, β is known and we already "have" its given value in this idealized simple model)
- Firstly， if we have the extra condition that Given conditional independence of α and γ given ß, upon squaring ωk we have：
- Implicit Assume: $\mathrm{E}(\alpha) = 0$; $E(\gamma) = 0$;
- We will firstly conduct the estimate case

corresponding to the theory casual case, then we will prove that it also works for the specific assayed case.

Under the generative model (1), the total effect of SNP k on the complex trait is

$$\omega_k = \sum_i^G \beta_{ik}\alpha_i + \gamma_k$$

Given conditional independence of α and γ given β, upon squaring ω_k we have

$$E\left[\omega_k^2 \mid \beta_{1k} \cdots \beta_{ik}\right] = \sum_i^G E\left[\alpha_i^2 \mid \beta_{1k} \cdots \beta_{ik}\right]\beta_{ik}^2 + E\left[\gamma_k^2 \mid \beta_{1k} \cdots \beta_{ik}\right]$$

3.Method and Derivations: Unstratified MESC

We model trait $\mathbf{y}$ for N individuals as follows:

$$\mathbf{y} = \mathbf{X}\gamma + \mathbf{X}\mathbf{B}\alpha + \epsilon \quad (1)$$

where $\mathbf{y}$ is an N-vector of phenotypes (standardized to mean 0 and variance 1), $\mathbf{X}$ is an $N \times M$ genotype matrix for M SNPs (standardized to mean 0 and variance 1), $\boldsymbol{\gamma}$ is an M vector of non-mediated SNP effect sizes on the trait (including pleiotropic, linkage, and trans-eQTL-mediated effects), $\mathbf{B}$ is an $M \times G$ matrix of cis-eQTL effect sizes *in the causal cell types/contexts* for G genes, $\boldsymbol{\alpha}$ is a G-vector of causal gene expression effect sizes on the trait, and $\boldsymbol{\epsilon}$ is an N-vector of environmental effects. We treat all variables as random. We define

- (3)The idealized casual Regression Equation for Unstratified MESC
- Furthermore, now we assume unconditional independence of α and γ to make more derivations, which is the following independence assume 4:
- Assume 4:Assuming unconditional independence of α and γ (which requires that we make additional effect size independence assumptions involving β; see the following "Model assumptions"),

Model assumptions

The two main effect size independence assumptions that are needed to derive equation (2) are:

1. Across all genes, the magnitude of gene effect sizes is uncorrelated with the magnitude of eQTL effect sizes (i.e. $Cov(\alpha^2, \beta^2) = 0$). We refer to this assumption as gene-eQTL effect size independence.
2. Across all SNPs, the magnitude of non-mediated SNP effect sizes is uncorrelated with the magnitude of eQTL effect sizes (i.e. $Cov(\gamma^2, \beta^2) = 0$). We refer to this assumption as pleiotropy-eQTL effect size independence.

3.Method and Derivations: Unstratified MESC

- (3)The idealized casual Regression Equation for Unstratified MESC
- Under the extra independence Assume 4, we get the following simplified regression equation(2):

Under the generative model (1), the total effect of SNP k on the complex trait is

$$\omega_k = \sum_i^G \beta_{ik}\alpha_i + \gamma_k$$

Given conditional independence of α and γ given β, upon squaring ω_k we have

$$E\left[\omega_k^2 \mid \beta_{1k}\ldots\beta_{ik}\right] = \sum_i^G E\left[\alpha_i^2 \mid \beta_{1k}\ldots\beta_{ik}\right]\beta_{ik}^2 + E\left[\gamma_k^2 \mid \beta_{1k}\ldots\beta_{ik}\right]$$

Model assumptions

The two main effect size independence assumptions that are needed to derive equation (2) are:

1. Across all genes, the magnitude of gene effect sizes is uncorrelated with the magnitude of eQTL effect sizes (i.e. $Cov(\alpha^2, \beta^2) = 0$). We refer to this assumption as gene-eQTL effect size independence.
2. Across all SNPs, the magnitude of non-mediated SNP effect sizes is uncorrelated with the magnitude of eQTL effect sizes (i.e. $Cov(\gamma^2, \beta^2) = 0$). We refer to this assumption as pleiotropy-eQTL effect size independence.

Assuming *unconditional* independence of α and γ (which requires that we make additional effect size independence assumptions involving β; see "Model assumptions"), this simplifies to

$$E\left[\omega_k^2 \mid \beta_{1k}\ldots\beta_{ik}\right] = E\left[\alpha^2\right]\sum_i^G \beta_{ik}^2 + E\left[\gamma^2\right] \qquad (2)$$

3.Method and Derivations: Unstratified MESC

Assuming *unconditional* independence of α and γ (which requires that we make additional effect size independence assumptions involving β; see "Model assumptions"), this simplifies to

$$E[\omega_k^2 \mid \beta_{1k} \ldots \beta_{ik}] = E[\alpha^2] \sum_i^G \beta_{ik}^2 + E[\gamma^2] \quad (2)$$

- (4) The application of the idealized casual regression equation(2)

- Estimate E[α^2] :
- We use equation (2) to estimate E[α^2] by regressing ω2 for all SNPs on $\sum_i^G \beta_{ik}^2$ and taking the slope,
- Estimate E[γ^2] :
- we estimate E[γ^2] by taking the intercept,
- Estimate $h^2_{med;casual}$:
- E[α^2] can be multiplied by GE[h^2_{cis}] to obtain $h^2_{med;casual}$
- Estimate $h^2_{nonmed;casual}$:
- E[γ^2] can be multiplied by M to obtain $h^2_{nonmed;casual}$

3.Method and Derivations: Unstratified MESC

(5)Some More Theory and Practice Problems to be Solved for the Unstratified Case:

A. In the previous idealized casual scenario case that SNP effect sizes are given, the previous derivation is only for the theory casual case(rather than assayed case). But we will show that when we carry out the regression procedure described in "Unstratified MESC" (Methods) using eQTL effect sizes assayed in non-causal tissues T, we exactly obtain an estimate of the quantity $h^2_{med;casual}$ as defined in "Definition of expression-mediated heritability"

When we perform this regression using eQTL effect sizes obtained from non-causal tissues T with squared genetic correlation $r_g^2(T)$ with the causal tissue(s), we obtain an estimate of the quantity $h^2_{med;\,assayed}(T)$ rather than $h^2_{med;\,causal}$ (Supplementary Note). Moreover, in practice we perform this regression using GWAS and eQTL summary statistics, in which case we account for differences in LD between SNPs with an LD score covariate (see Supplementary Note for derivation and regression equation).

3.Method and Derivations: Unstratified MESC

(5)Some More Theory and Practice Problems to be Solved for the Unstratified Case:

B. The previous method is under the idealized scenario case that SNP effect sizes are given. In practice, we instead use the GWAS summary statistics, which are affected by sampling noise and by LD. Fortunately, it has previously been shown that LD and sampling noise can be accounted for by regressing GWAS $\chi 2$ statistics on LD scores, which measure the total of LD for each SNP .

So when in practice we perform this regression using GWAS and eQTL summary statistics, in which case we have to take the differences in LD between SNPs into consideration, so we put an extra LD score covariate term into the equation. In fact, It's just a general case of the traditional LD Score Regression.

When we perform this regression using eQTL effect sizes obtained from non-causal tissues T with squared genetic correlation $r_g^2(T)$ with the causal tissue(s), we obtain an estimate of the quantity $h^2_{med;\,assayed}(T)$ rather than $h^2_{med;\,causal}$ (Supplementary Note). Moreover, in practice we perform this regression using GWAS and eQTL summary statistics, in which case we account for differences in LD between SNPs with an LD score covariate (see Supplementary Note for derivation and regression equation).

3.Method and Derivations: Unstratified MESC

- (5) Problem A: In the idealized SNPs effect given case, show that we can also succesfully obtain an estimate of the quantity $h^2_{med;assayed}$ as defined in "Definition of expression-mediated heritability", not only just get an estimate of the theory or " true" casual case:
- β represent cis-eQTL effect sizes in causal cell types/contexts for the trait
- β' represent cis-eQTL effect sizes in assayed tissues T
- We start with regression equation from Methods: The assume for the idealized case:

$$E[\omega_k^2] = E[\alpha^2] \sum_i^G \beta_{ik}^2 + E[\gamma^2]$$

For illustrative purposes, we walk through a derivation for MESC in the idealized scenario that we know 1. the true eQTL effect sizes, β, of each SNP on each gene and 2. the true phenotypic effect sizes, ω, of each SNP on y.

- The ordinary least squares estimate of the coefficient from regressing ω^2 on $\sum_i^G \beta'^2_i$:

$$\alpha'^2 = \frac{Cov(\omega^2, \sum_i^G \beta'^2_i)}{Var(\sum_i^G \beta'^2_i)}$$

3.Method and Derivations: Unstratified MESC

- The ordinary least squares estimate of the coefficient from regressing ω^2 on $\sum_i^G \beta'^2_i$:

$$\alpha'^2 = \frac{Cov(\omega^2, \sum_i^G \beta'^2_i)}{Var(\sum_i^G \beta'^2_i)} \approx \frac{1}{G}\sum_i^G \frac{Cov(\omega^2, \beta'^2_i)}{Var(\beta'^2_i)}$$

- Since we have the following assume, we also note that E [α^2] is the average squared per-gene effect of expression on trait, so we have:

The third line follows given the $Cov(\gamma^2, \beta'^2) = 0$ and $Cov(\alpha^2, \beta'^2) = 0$.

$$\begin{aligned}
\alpha'^2 &= \frac{Cov(\omega^2, \sum_i^G \beta'^2_i)}{Var(\sum_i^G \beta'^2_i)} \\
&\approx \frac{1}{G}\sum_i^G \frac{Cov(\omega^2, \beta'^2_i)}{Var(\beta'^2_i)} \\
&\approx \frac{1}{G}\sum_i^G \frac{Cov(\alpha_i^2\beta_i^2 + \gamma^2, \beta'^2_i)}{Var(\beta'^2_i)} \\
&\approx E[\alpha^2]\frac{1}{G}\sum_i^G \frac{Cov(\beta_i^2, \beta'^2_i)}{Var(\beta'^2_i)}
\end{aligned}$$

3.Method and Derivations: Unstratified MESC

- Furthermore, notice that we have defined that: the average squared genetic correlation between expression in T vs. in causal cell types:

$$r_g^2(T) = \frac{1}{G}\sum_i^G \frac{Cov(\beta_i^2, \beta_i'^2)}{\sqrt{Var(\beta_i^2)Var(\beta_i'^2)}}$$

- Given this definition, we have:

$$\alpha'^2 \approx r_g^2(T)E[\alpha^2]\frac{1}{G}\sum_i^G \sqrt{\frac{Var(\beta_i^2)}{Var(\beta'^2_i)}}$$

- For simplicity, we make the assumption that V ar(β_i^2) ≈ V ar(β'^2_i) across genes:

$$\alpha'^2 \approx r_g^2(T)E[\alpha^2]$$

3.Method and Derivations: Unstratified MESC

- By the previous two relationships between the
- assayed mediated heritability and the
- casual mediated heritability, also we have the
- rewrite form of the casual mediated heritability,

Under the assumption that α and β are independent of each other, we can rewrite this as follows:

$$\begin{aligned} h^2_{med;\,causal} &= E_{B,\alpha}[Var[\mathbf{XB\alpha} \mid \mathbf{B}, \boldsymbol{\alpha}]] + Var_{B,\alpha}[E[\mathbf{XB\alpha} \mid \mathbf{B}, \boldsymbol{\alpha}]] \\ &= E_{B,\alpha}\left[\sum_i^G \sum_j^M \beta_{ij}^2 \alpha_i^2\right] \\ &= GE\left[\alpha^2\right] E[h^2_{cis}] \end{aligned}$$

So we finally have the following unbiased estimate method

$$\alpha'^2 \approx r_g^2(T) E[\alpha^2]$$

of the assayed mediated heritability and prove it only work for the casual mediated case bust also can work for the assayed mediated case:

$$h^2_{med;\,assayed}(T) = r_g^2(T) h^2_{med;\,causal}$$

while we define $h^2_{nonmed;assayed}(T)$ as $h^2_{nonmed;causal} + \left(1 - r_g^2(T)\right) h^2_{med;\,causal}$. Here,

We can then multiply α'^2 by $GE[h'^2_{cis}]$ to obtain an unbiased estimate of $h^2_{med;assayed}(T)$, where $E[h'^2_{cis}]$ is the average expression cis-heritability of genes in T.

3.Method and Derivations: Unstratified MESC

- Remark: the case for the different variance still holds:

$$Estimates\ of\ h^2_{med;assayed}(T)\ \ when\ Var(\beta_i^2) \neq Var(\beta_i'^2).$$

- e.g. if causal genes for the trait are primarily influenced by cell type-specific eQTLs that are weaker/absent in assayed tissues
- To illustrate the remark, we consider the following specific situation, that is:

$$\beta'^2_i = c\beta^2_i$$

To illustrate this, consider a scenario where β'^2 is both correlated with β^2 and scaled by a factor c relative to β^2. (1) thus becomes

3.Method and Derivations: Unstratified MESC

- Derivations of the remark:

$$\alpha'^2 = r_g^2(T)E[\alpha^2]\frac{1}{G}\sum_i^G \sqrt{\frac{Var(\beta_i^2)}{Var(c\beta_i^2)}}$$
$$= \frac{1}{c}r_g^2(T)E[\alpha^2]$$

Note that scaling β'^2 by c will not change the average squared correlation $r_g^2(T)$ between β'^2 and β. We then have

$$h^2_{med;assayed}(T) = GE[\beta'^2]\frac{1}{c}r_g^2(T)E[\alpha^2]$$
$$= GE[c\beta^2]\frac{1}{c}r_g^2(T)E[\alpha^2]$$
$$= r_g^2(T)h^2_{med;causal}$$

3.Method and Derivations: Unstratified MESC

$$\phi = X\beta + \epsilon$$

- (5) The summary GWAS statistics case of the unstratified MESC(A special case of the LD score)
- Under our generative model, take the total effect size ω_k as a whole term:

We model trait **y** for N individuals as follows:

$$\mathrm{y} = \mathrm{X}\gamma + \mathrm{XB}\alpha + \epsilon \quad (1)$$

where **y** is an N-vector of phenotypes (standardized to mean 0 and variance 1), **X** is an $N \times M$ genotype matrix for M SNPs (standardized to mean 0 and variance 1), $\boldsymbol{\gamma}$ is an M vector of non-mediated SNP effect sizes on the trait (including pleiotropic, linkage, and trans-eQTL-mediated effects), **B** is an $M \times G$ matrix of cis-eQTL effect sizes *in the causal cell types/ contexts* for G genes, $\boldsymbol{\alpha}$ is a G-vector of causal gene expression effect sizes on the trait, and $\boldsymbol{\epsilon}$ is an N-vector of environmental effects. We treat all variables as random. We define $h^2_{med;\,causal}$ as follows:

Under the generative model (1), the total effect of SNP k on the complex trait is

$$\omega_k = \sum_i^G \beta_{ik}\alpha_i + \gamma_k$$

$$E\left[\omega_k^2 \mid \beta_{1k} \ldots \beta_{ik}\right] = E\left[\alpha^2\right]\sum_i^G \beta_{ik}^2 + E\left[\gamma^2\right] \quad (2)$$

- We just firstly follow the derivations of the traditional LD score regression method to conduct the estimate as following:

3.Method and Derivations: Unstratified MESC

- (6) The summary GWAS statistics case of the unstratified MESC(A special case of the LD score)

Just like the LD Score Regression, the marginal OLS estimate of the total effect size of a SNP k on the trait is given by:

$$
\begin{aligned}
\hat{\omega}_k &= \frac{1}{N}\boldsymbol{X_k^T y} \\
&= \frac{1}{N}(\boldsymbol{X_k^T}(\boldsymbol{X\gamma} + \boldsymbol{XB\alpha} + \boldsymbol{\epsilon})) \\
&= \frac{1}{N}\boldsymbol{X_k^T X\gamma} + \frac{1}{N}\boldsymbol{X_k^T XB\alpha} + \frac{1}{N}\boldsymbol{X_k^T\epsilon} \\
&= \sum_j^M \gamma_j \hat{r}_{jk} + \sum_i^G \alpha_i \sum_j^M \hat{r}_{jk}\beta_{ij} + \boldsymbol{\epsilon'}
\end{aligned}
$$

Where we let:

Let $\hat{\boldsymbol{R}} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$ denote the in-sample LD matrix, and let $\boldsymbol{\epsilon'} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{\epsilon}$ denote the noise term in the summary statistics. The χ^2 statistic for SNP k (defined as $N\hat{\omega}_k^2$) is:

3.Method and Derivations: Unstratified MESC

- (6) The summary GWAS statistics case of the unstratified MESC(A special case of the LD score)
- Just like the idealized case, we still need some independence assumptions to derivate an equation like (2):

Assuming *unconditional* independence of a and γ (which requires that we make additional effect size independence assumptions involving β; see "Model assumptions"), this simplifies to

$$E\left[\omega_k^2 \mid \beta_{1k} \ldots \beta_{ik}\right] = E\left[\alpha^2\right] \sum_i^G \beta_{ik}^2 + E\left[\gamma^2\right] \quad (2)$$

Model assumptions

The two main effect size independence assumptions that are needed to derive equation (2) are:

1. Across all genes, the magnitude of gene effect sizes is uncorrelated with the magnitude of eQTL effect sizes (i.e. $Cov(a^2, \beta^2) = 0$). We refer to this assumption as gene-eQTL effect size independence.
2. Across all SNPs, the magnitude of non-mediated SNP effect sizes is uncorrelated with the magnitude of eQTL effect sizes (i.e. $Cov(\gamma^2, \beta^2) = 0$). We refer to this assumption as pleiotropy-eQTL effect size independence.

3.Method and Derivations: Unstratified MESC

- In order for the equation regarding the unconditional expectation of χ^2 to hold true, we must make two independence assumptions involving LD-dependent genetic architecture, in addition to the independence assumptions described in "Model assumptions" (Methods). LD-dependent architecture, if not accounted for, is known to produce bias in heritability estimates. These assumptions are:
- • Across all genes (indexed by i), the magnitude of α_i is uncorrelated with the LD scores of eQTLs for gene i
- • Across all SNPs (indexed by k), the magnitude of γ_k is uncorrelated with the LD score of SNP k

3.Method and Derivations: Unstratified MESC

- Given conditional independence of γ_j and α_i given *B* and $\hat{R}$, upon squaring ω k (to a constant N) we still have:

$$E[\chi_k^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}] = N \sum_j^M E[\gamma_j^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}] \hat{r}_{jk}^2 + N \sum_i^G E[\alpha_i^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}] \sum_j^M \hat{r}_{jk}^2 \beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \quad (2)$$

- By the previous independence assumptions, we can delete the conditional independence of γ_j and α_i :

$$E[\chi_k^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}] = N \sum_j^M E[\gamma_j^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}] \hat{r}_{jk}^2 + N \sum_i^G E[\alpha_i^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}] \sum_j^M \hat{r}_{jk}^2 \beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \quad (2)$$

$$= NE[\gamma^2] \sum_j^M \hat{r}_{jk}^2 + NE[\alpha^2] \sum_i^G \sum_j^M \hat{r}_{jk}^2 \beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \quad (3)$$

3.Method and Derivations: Unstratified MESC

- By the previous rewrite form of both the non-mediated casual heritability and the mediated casual heritability, as well as the variance assume of the ϵ', we have that:

$$= NE[\gamma^2] \sum_{j}^{M} \hat{r}_{jk}^2 + NE[\alpha^2] \sum_{i}^{G} \sum_{j}^{M} \hat{r}_{jk}^2 \beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \tag{3}$$

$$= \frac{Nh_{nonmed;causal}^2}{M} \sum_{j}^{M} \hat{r}_{jk}^2 + \frac{Nh_{med;causal}^2}{GE[h_{cis}^2]} \sum_{i}^{G} \sum_{j}^{M} \hat{r}_{jk}^2 \beta_{ij}^2 + 1 - h_{med;causal}^2 - h_{nonmed;causal}^2 \tag{4}$$

- Now by the approximation formular of the sample LD score and the real LD score:

Since $E[\hat{r}_{jk}^2] \approx r_{jk}^2 + \frac{1}{N}$, we have

3.Method and Derivations: Unstratified MESC

- So we totally have that:

$$E\left[\sum_{j}^{M}\hat{r}_{jk}^{2}\right]\approx\sum_{j}^{M}r_{jk}^{2}+\frac{M}{N}$$

and

$$E\left[\sum_{i}^{G}\sum_{j}^{M}\hat{r}_{jk}^{2}\beta_{ij}^{2}\right]\approx\sum_{i}^{G}\sum_{j}^{M}\left(r_{jk}^{2}\beta_{ij}^{2}+\frac{\beta_{ij}^{2}}{N}\right)$$
$$\approx\sum_{i}^{G}\sum_{j}^{M}r_{jk}^{2}\beta_{ij}^{2}+\frac{GE[h_{cis}^{2}]}{N}$$

Thus,

$$E[\chi_{k}^{2}]\approx\frac{Nh_{nonmed;causal}^{2}}{M}\left(\sum_{j}^{M}r_{jk}^{2}+\frac{M}{N}\right)+\frac{Nh_{med;causal}^{2}}{GE[h_{cis}^{2}]}\left(\sum_{i}^{G}\sum_{j}^{M}r_{jk}^{2}\beta_{ij}^{2}+\frac{GE[h_{cis}^{2}]}{N}\right)+1-h_{nonmed;causal}^{2}-h_{med;causal}^{2}$$
$$\approx\frac{Nh_{nonmed;causal}^{2}}{M}\sum_{j}^{M}r_{jk}^{2}+\frac{Nh_{med;causal}^{2}}{GE[h_{cis}^{2}]}\sum_{i}^{G}\sum_{j}^{M}r_{jk}^{2}\beta_{ij}^{2}+1$$

3.Method and Derivations: Unstratified MESC

- We finally define and rewrite the LD scores and the expression scores:

$$\text{Defining LD scores } \ell_k = \sum_j^M r_{jk}^2 \text{ and expression scores } \mathcal{L}_k = \sum_i^G \sum_j^M r_{jk}^2 \beta_{ij}^2,$$

- We arrive at our main equation for summary MESC regression:

$$E[\chi_k^2] \approx \frac{Nh_{nonmed;causal}^2}{M}\ell_k + \frac{Nh_{med;causal}^2}{GE[h_{cis}^2]}\mathcal{L}_k + 1 \qquad (5)$$

3.Method and Derivations: Unstratified MESC

- Similar to the derivation in “Unstratified MESC with non-causal eQTL effect sizes” (the idealized case), we show that if we perform this regression using expression scores in assayed tissues T rather than expression scores in causal cell types/contexts, we still can obtain an estimate of the $h^2_{med,assayed}$(T):

3.Method and Derivations: Unstratified MESC

Defining LD scores $\ell_k = \sum_j^M r_{jk}^2$ and expression scores $\mathcal{L}_k = \sum_i^G \sum_j^M r_{jk}^2 \beta_{ij}^2$, we arrive at our main equation for summary MESC regression:

$$E[\chi_k^2] \approx \frac{Nh_{nonmed;causal}^2}{M}\ell_k + \frac{Nh_{med;causal}^2}{GE[h_{cis}^2]}\mathcal{L}_k + 1 \tag{5}$$

- Upon regressing GWAS χ^2 statistics on expression scores in assayed tissues (see equation (5))

χ^2 statistics. Upon regressing GWAS χ^2 statistics on expression scores in assayed tissues (see equation (5)), we have

$$\begin{aligned}
\alpha'^2 &\approx \frac{1}{G}\sum_i^G \frac{Cov(\chi^2, \sum_j^M r_j^2 \beta'^2_{ij})}{Var(\sum_j^M r_j^2 \beta'^2_{ij})} \\
&\approx \frac{1}{G}\sum_i^G \frac{Cov(\sum_j^M r_j^2 \alpha_i^2 \beta_{ij}^2, \sum_j^M r_j^2 \beta'^2_{ij})}{Var(\sum_j^M r_j^2 \beta'^2_{ij})} \\
&\approx E[\alpha^2]\frac{1}{G}\sum_i^G \frac{\ell^2 Cov(\beta_i^2, \beta'^2_i)}{\ell^2 Var(\beta'^2_i)} \\
&\approx E[\alpha^2]\frac{1}{G}\sum_i^G \frac{Cov(\beta_i^2, \beta'^2_i)}{Var(\beta'^2_i)}
\end{aligned}$$

Here, $\ell^2 = Var(\sum_j^M r_j^2)$. The third and fourth line follow given that r^2 is independent of α, β, and β'. See "Unstratified MESC with non-causal eQTL effect sizes" (above) for the remainder of the derivation.

3.Method and Derivations: Stratified MESC

- In this section, we extend unstratified MESC to estimate $h^2_{med,assayed}$(T) partitioned over groups of genes. Note that stratified MESC can also be viewed as a special form of stratified LD score regression. We still start from the casual case:
- (1) definition of the $h^2_{med,casual}(\mathcal{D}_d)$
- We define the $h^2_{med,casual}(\mathcal{D}_d)$ partitioned over gene categories $\mathcal{D}_1$,…$\mathcal{D}_d$ as follow:

$$h^2_{med;\,causal}(\mathcal{D}_d) = \sum_{i \in \mathcal{D}_d} \alpha_i^2 \sum_{j}^{M} \beta_{ij}^2$$

3.Method and Derivations: Stratified MESC

- We can also similarly rewrite it as the following average form also similarly under the assume that there exist independence assumption between α and β:

$$h^2_{med;\,causal}(\mathscr{D}_d) = \sum_{i \in \mathscr{D}_d} \alpha_i^2 \sum_j^M \beta_{ij}^2$$
$$= |\mathscr{D}_d| \cdot E\left[\alpha_i^2 \middle| i \in \mathscr{D}_d\right] \cdot E\left[h^2_{i;\,cis} \middle| i \in \mathscr{D}_d\right]$$

where $h^2_{med;\,causal}(\mathscr{D}_d)$ is the heritability mediated in cis through the expression of genes in category $\mathscr{D}_d$, $|\mathscr{D}_d|$ is the number of genes in $\mathscr{D}_d$, $E[\alpha_i^2 | i \in \mathscr{D}_d]$ is the average squared causal effect of expression on trait for genes in $\mathscr{D}_d$, and $E[h^2_{i;\,cis} | i \in \mathscr{D}_d]$ is the average cis-heritability of expression of genes in $\mathscr{D}_d$. Similar to our definition of $h^2_{med;\,causal}$, the second line above relies on an independence assumption between α and β, namely that $\alpha_i \perp \beta_i | \not\in \mathscr{D}_d$.

3.Method and Derivations: Stratified MESC

- (2) Model of the variance of the gene effect size

For gene i, we model the variance of gene effect size α_i as

$$Var(\alpha_i) = \sum_{d: i \in \mathscr{D}_d} \pi_d$$

If gene categories $\mathscr{D}_d$ form a disjoint partition of the set of all genes, we have

$$\pi_d = \frac{E\left[h^2_{med;\,causal}(\mathscr{D}_d)\right]}{|\mathscr{D}_d| E[h^2_{i;\,cis} \,|\, i \in \mathscr{D}_d]}$$

- Remark: If it is not disjoint, then we have the following conceptually interpretation of the model:

On the other hand, if gene categories are overlapping, then π_d can be conceptualized as the contribution of annotation $\mathscr{D}_d$ to $h^2_{med;\,causal}$ conditional on contributions from all other gene categories included in the model.

3.Method and Derivations: Stratified MESC

- (3) the definition case for the $h^2_{nonmed,casual}$
- We still firstly define the $h^2_{nonmed,casual}$ partitioned over SNP categories as follows:

$$h^2_{nonmed;causal}(\mathscr{C}_c) = \sum_{j \in \mathscr{C}_c} \gamma_j^2$$
$$= |\mathscr{C}_c| \cdot E[\gamma_j^2 \Big| j \in \mathscr{C}_c]$$

where $h^2_{nonmed;causal}(\mathscr{C}_c)$ is the non-mediated heritability of SNPs in category $\mathscr{C}_c$, $|\mathscr{C}_c|$ is the number of SNPs in $\mathscr{C}_c$, and $E[\gamma_j^2 | j \in \mathscr{C}_c]$ is the average squared non-mediated effect size of SNPs in $\mathscr{C}_c$.

3.Method and Derivations: Stratified MESC

- For SNP j, we model the variance of non-mediated effect size γj as follows:

$$Var(\gamma_j) = \sum_{c:\, j \in \mathscr{C}_c} \tau_c$$

If SNP categories $\mathscr{C}_c$ form a disjoint partition of the set of all SNPs, we have

$$\tau_c = \frac{E\left[h^2_{nonmed;causal}(\mathscr{C}_c)\right]}{|\mathscr{C}_c|}$$

- Remark: If it is not disjoint, then we have the following conceptually interpretation of the model:

On the other hand, if SNP categories are overlapping, then τ_c can be conceptualized of as the contribution of annotation $\mathscr{C}_c$ to $h^2_{nonmed;causal}$ conditional on contributions from all other SNP categories included in the model.

3.Method and Derivations: Stratified MESC

- (4) the regression equation for the casual case:

The equation for stratified MESC is

$$E[\chi_k^2] = N\sum_c \tau_c \ell_{k;c} + N\sum_d \pi_d \mathscr{L}_{k;d} + 1 \qquad (3)$$

where χ_k^2 is the GWAS χ^2-statistic of SNP k, N is the number of samples, $\ell_{k;c}$ is the LD score of SNP k with respect to SNP category $\mathscr{C}_c$ (defined as $\ell_{k;c} = \sum_{j \in \mathscr{C}_c} r_{jk}^2$), and $\mathscr{L}_{k;d}$ is the expression score of SNP k with respect to gene category $\mathscr{D}_d$ (defined as $\mathscr{L}_{k;d} = \sum_{i \in \mathscr{D}_d} \sum_j^M r_{jk}^2 \beta_{ij}^2$). Here, r_{jk} refers to the LD between SNPs j and k. See Supplementary Note for a derivation of this equation. Analogous to unstratified MESC,

- Analogous to unstratified MESC, when we perform this regression using expression scores in assayed tissues T rather than expression scores in causal cell types/contexts, we will estimate:

$h_{med;\,assayed}^2(T, \mathscr{D}_d) = r^2(T, \mathscr{D}_d) h_{med;\,causal}^2(\mathscr{D}_d)$, where $r^2(T, \mathscr{D}_d)$ is the average squared genetic correlation of expression between T and causal cell types/contexts for genes in $\mathscr{D}_d$.

3.Method and Derivations: Stratified MESC

- We still start from the following equation (2) from the unstratified MESC case:

Let $\hat{\boldsymbol{R}} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$ denote the in-sample LD matrix, and let $\boldsymbol{\epsilon}' = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{\epsilon}$ denote the noise term in the summary statistics. The χ^2 statistic for SNP k (defined as $N\hat{\omega}_k^2$) is:

$$E[\chi_k^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}] = N\sum_{\cdot}^{M} E[\gamma_j^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}]\hat{r}_{jk}^2 + N\sum_{\cdot}^{G} E[\alpha_i^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}]\sum_{\cdot}^{M}\hat{r}_{jk}^2\beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \quad (2)$$

- By the previous model of the variance, we have the (7):

$$E[\chi_k^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}] = N\sum_{j}^{M} E[\gamma_j^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}]\hat{r}_{jk}^2 + N\sum_{i}^{G} E[\alpha_i^2 \mid \hat{\boldsymbol{R}}, \boldsymbol{B}]\sum_{j}^{M}\hat{r}_{jk}^2\beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \quad (6)$$

$$= N\sum_{j}^{M}\left(\sum_{c:j\in\mathcal{C}_c}\tau_c \mid \hat{\boldsymbol{R}}, \boldsymbol{B}\right)\hat{r}_{jk}^2 + N\sum_{i}^{G}\left(\sum_{d:i\in\mathcal{D}_d}\pi_d \mid \hat{\boldsymbol{R}}, \boldsymbol{B}\right)\sum_{j}^{M}\hat{r}_{jk}^2\beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \quad (7)$$

3.Method and Derivations: Stratified MESC

By the following 4 independence assumptions:, we have (8):

In order for (8) to be true, we must make the following assumptions:

- Within each gene category $\mathcal{D}_d$, π_d is uncorrelated with the magnitude of eQTL effect sizes
- Within each SNP category $\mathcal{C}_c$, τ_c is uncorrelated with the magnitude of eQTL effect sizes
- π_d is uncorrelated with the LD scores of eQTLs that affect genes in $\mathcal{D}_d$
- τ_c is uncorrelated with the LD scores of SNPs in $\mathcal{C}_c$

$$= N \sum_{j}^{M} \left(\sum_{c:j \in \mathcal{C}_c} \tau_c \mid \hat{\boldsymbol{R}}, \boldsymbol{B} \right) \hat{r}_{jk}^2 + N \sum_{i}^{G} \left(\sum_{d:i \in \mathcal{D}_d} \pi_d \mid \hat{\boldsymbol{R}}, \boldsymbol{B} \right) \sum_{j}^{M} \hat{r}_{jk}^2 \beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \quad (7)$$

$$E[\chi_k^2] = N \sum_{c} \tau_c \sum_{j \in \mathcal{C}_c} \hat{r}_{jk}^2 + N \sum_{d} \pi_d \sum_{i \in \mathcal{D}_d} \sum_{j}^{M} \hat{r}_{jk}^2 \beta_{ij}^2 + NE[(\boldsymbol{\epsilon}')^2] \quad (8)$$

3.Method and Derivations: Stratified MESC

- By the approximation formular of the sample LD Score and also by the previous direct definition of the mediated and non-mediated heritability:

Since $E[\hat{r}_{jk}^2] \approx r_{jk}^2 + \frac{1}{N}$, we have

$$E[\chi_k^2] = N\sum_c \tau_c \sum_{j\in C_c}\left(r_{jk}^2 + \frac{1}{N}\right) + N\sum_d \pi_d \sum_{i\in\mathcal{D}_d}\sum_j^M \left(r_{jk}^2\beta_{ij}^2 + \frac{\beta_{ij}^2}{N}\right) + NE[(\boldsymbol{\epsilon}')^2]$$

$$= N\sum_c \tau_c \sum_{j\in C_c} r_{jk}^2 + \sum_c \sum_{j\in C_c} \tau_c + N\sum_d \pi_d \sum_{i\in\mathcal{D}_d}\sum_j^M r_{jk}^2\beta_{ij}^2 +$$

$$\sum_d \sum_{i\in\mathcal{D}_d}\sum_j^M \left(\pi_d|\mathcal{D}_d|E[h_{cis}^2(\mathcal{D}_d)]\right) + NE[(\boldsymbol{\epsilon}')^2]$$

$$= N\sum_c \tau_c \sum_{j\in C_c} r_{jk}^2 + h_{nonmed;causal}^2 + N\sum_d \pi_d \sum_{i\in\mathcal{D}_d}\sum_j^M r_{jk}^2\beta_{ij}^2 + h_{med;causal}^2 + 1 - h_{nonmed;causal}^2 - h_{med;causal}^2$$

$$= N\sum_c \tau_c \sum_{j\in C_c} r_{jk}^2 + N\sum_d \pi_d \sum_{i\in\mathcal{D}_d}\sum_j^M r_{jk}^2\beta_{ij}^2 + 1$$

3.Method and Derivations: Stratified MESC

- Finally, we rewrite the equation(8) and arrive at our main equation for stratified MESC:

Letting $\ell_{k;c} = \sum_{j \in C_c} r_{jk}^2$ and $\mathcal{L}_{k;d} = \sum_{i \in \mathcal{D}_d} \sum_j^M r_{jk}^2 \beta_{ij}^2$, we arrive at our main equation for stratified MESC:

$$E[\chi_k^2] = N \sum_c \tau_c \ell_{k;c} + N \sum_d \pi_d \mathcal{L}_{k;d} + 1 \tag{9}$$

3.Method and Derivations: Stratified MESC

- (5) Estimation of expression scores
- In order to carry out the regression described in equation (9), we must first estimate expression scores $\mathcal{L}_{k;d}$ from an external expression panel:

Letting $\ell_{k;c} = \sum_{j \in C_c} r_{jk}^2$ and $\mathcal{L}_{k;d} = \sum_{i \in \mathcal{D}_d} \sum_j^M r_{jk}^2 \beta_{ij}^2$, we arrive at our main equation for stratified MESC:

$$E[\chi_k^2] = N \sum_c \tau_c \ell_{k;c} + N \sum_d \pi_d \mathcal{L}_{k;d} + 1 \qquad (9)$$

- The paper estimates $\mathcal{L}_{k;d}$ from either eQTL summary statistics or individual-level genotypes, here we study the method conducted by the eQTL summary statistics:

3.Method and Derivations: Stratified MESC

- We can use summary statistics from eQTL studies to estimate expression scores $\mathcal{L}$k;d:
- which are equivalent to the sum of marginal OLS estimates of eQTL effect sizes for SNP k on genes in
- $\mathcal{D}_{\mathrm{d}}$, modulo an error term that depends on |D d | and the sample size of the eQTL study. $(\sum_{i\in\mathcal{D}_d}\hat{\beta}^2_{ik(sumstat)})$
- This error term will be captured by the intercept during regression.
- Now we write down the model:

$$\boldsymbol{y}_{i(exp)} = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_{i(exp)}$$

3.Method and Derivations: Stratified MESC

- In the model:

$$\boldsymbol{y}_{i(exp)} = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_{i(exp)}$$

5

where $\boldsymbol{y}_{i(exp)}$ is an N_{exp}-vector of gene expression measurements (standardized to mean 0 and variance 1), $\boldsymbol{X}$ is an $N_{exp} \times M$ genotype for M SNPs (standardized to mean 0 and variance 1), $\boldsymbol{\beta}_i$ is an M-vector of eQTL effect sizes, and $\boldsymbol{\epsilon}_{i(exp)}$ is an N_{exp}-vector of environmental effects. Under this model, we have

3.Method and Derivations: Stratified MESC

- Under this model, we have the following derivations:

$$
\begin{aligned}
E\left[\sum_{i\in\mathcal{D}_d}\hat{\beta}^2_{ik(sumstat)}\right] &= \sum_{i\in\mathcal{D}_d}\left(\sum_{j}^{M}\hat{r}^2_{jk}\beta^2_{ij}+\frac{E[\epsilon^2_{i(exp)}]}{N_{exp}}\right)\\
&= \sum_{i\in\mathcal{D}_d}\sum_{j}^{M}\hat{r}^2_{jk}\beta^2_{ij}+\sum_{i\in\mathcal{D}_d}\frac{1-E[h^2_{cis}]}{N_{exp}}\\
&= \sum_{i\in\mathcal{D}_d}\sum_{j}^{M}r^2_{jk}\beta^2_{ij}+\frac{|\mathcal{D}_d|E[h^2_{cis}]}{N_{exp}}+\frac{|\mathcal{D}_d|(1-E[h^2_{cis}])}{N_{exp}}\\
&= \mathcal{L}_{k;d}+\frac{|\mathcal{D}_d|}{N_{exp}}
\end{aligned}
$$

3.Method and Derivations: Stratified MESC

- Thus, we can use the following alternate form of equation (9) to perform regression:

$$E[\chi_k^2] = N\sum_c \tau_c \ell_{k;c} + N\sum_d \pi_d \sum_{i\in\mathcal{D}_d} \hat{\beta}^2_{ik(sumstat)} + 1 + \frac{Nh^2_{med;causal}}{N_{exp}E[h^2_{cis}]}$$

4.Simulations and Real Data

- (1) Basic Practice Applications of the MESC Methods:

- Totally speaking, throughout this study, we present estimates of three quantities that are a function of h^2_{med} or h^2_{med}(D) as the following (1), (2), (3):

and/or $h^2_{med}(D)$: (1) the proportion of heritability mediated by expression (defined as h^2_{med}/h^2_g), (2) the proportion of expression-mediated heritability for gene category D (defined as $h^2_{med}(D)/h^2_{med}$), and (3) the enrichment of expression-mediated heritability for D (defined as the proportion of expression-mediated heritability in D divided by the proportion of genes in D). We estimate standard errors and p-values for all quantities by jackknifing over blocks

4.Simulations and Real Data

- (2) The general goal and method of the simulations
- We performed simulations to assess the calibration and bias of MESC in estimating h^2_{med}/ h^2_g and its standard error from simulated complex trait and expression data under a variety of genetic architectures
- Method: We performed all simulations using real genotypes from UK Biobank 38 (N_{GWAS} = 10,000 GWAS samples; N_{eQTL} = 100–1000 expression samples, M = 98,499 SNPs from chromosome 1).

4.Simulations and Real Data

- (3) Evaluate the bias of MESC(A)
- We evaluated the bias of MESC in estimating various values of h^2_{med} / h^2_g in the following scenarios:

scenarios: (1) when varying expression panel sample size (Figure 2a), (2) when varying the proportion of SNPs and genes with nonzero effects (Figure 2b), (3) when simulating eQTL effect sizes in the gene expression panel that differ from those used to generate the complex trait phenotype, emulating the scenario in which assayed tissues differ from the causal tissue(s) for the disease (Figure 2c), (4) when using different methods to estimate expression scores (5 in total) (Supplementary Figure 1), (5) when varying total disease heritability (Supplementary Figure 2), and (6) when including rare variants and inducing an inverse relationship between eQTL/GWAS effect size magnitude and minor allele frequency (Supplementary Figure 3), consistent with negative selection acting on both gene expression[39,40] and complex trait[41,42]. We observed that MESC produced unbiased or nearly

4.Simulations and Real Data

- Result 1: We observed that MESC produced unbiased or nearly unbiased estimates of h^2_{med} / h^2_g across all simulated genetic architectures with expression panel sample size greater than 500.
- Condition: When using the best-performing method to estimate expression scores, LASSO with REML correction (Methods).

- Result 2:We also note that available expression panel sample sizes for individual tissues are typically smaller than 500, which necessitates meta-analysis across tissues to attain larger expression panel sample sizes

- Result 3: For scenario (3), we expect in theory that MESC will estimate the quantity: $r_g^2(T)h^2_{med;\,causal}$ when using expression scores from a non-causal tissue with average squared genetic correlation of expression $r_g^2(T)$ with the causal tissue. Our simulation results support this theoretical expectation.

4.Simulations and Real Data

- (4) Assessed the bias of MESC (B)
- We assessed the bias of MESC in two biologically plausible scenarios corresponding to violations of the two main effect size independence assumptions (Methods),
- Where the independence assumptions are:

assessed how well partitioning genes and SNPs ameliorated this bias. The assumptions can be summarized as: (1) gene-eQTL independence, where eQTL and gene effect size magnitude are independent within each gene category, and (2) pleiotropy-eQTL effect size independence, where eQTL and SNP non-mediated effect size magnitude are independent within each SNP category. We simulated violations of (1) by inducing a negative correlation

4.Simulations and Real Data

- Method: We simulated violations of (1) by inducing a negative correlation
- between eQTL and gene effect size magnitude across the genome.
- We simulated violations of (2) by inducing enrichment of eQTLs and non-mediated effects within the same SNP categories (e.g. coding regions, transcription start sites, or conserved regions).

4.Simulations and Real Data

- Results:
- 1.We observed that partitioning genes into 5 bins by the magnitude of their expression heritability enabled us to obtain approximately unbiased estimates of $h^2_{med}/\ h^2_g$ (Figure 2d).
- 2.We observed that partitioning SNPs by the baseline LD model (a set of comprehensive functional SNP annotations) enabled us to obtain approximately unbiased estimates of $h^2_{med}/\ h^2_g$ (Figure 2e)
- (even in extreme scenarios e.g. when 100% of mediated and non-mediated heritability were entirely concentrated in coding regions.)

4.Simulations and Real Data

- (5) Comparing MESC to other methods.
- Method: The closest analogues are approaches that measure the genome-wide heritability enrichment of eQTLs using GCTA or stratified LD score regression (S-LDSC)

- Result: In simulations, we found that S-LDSC detected significant heritability enrichment of a SNP category corresponding to the set of all eQTLs in the absence of any mediation(h^2_{med}/ h^2_g=0). While MESC had a well-calibrated false positive rate for detecting significantly non-zero in this scenario (Figure 2f).

4.Simulations and Real Data

- (6)Simulations Conclusions
- In summary, we show that MESC produces approximately unbiased estimates of h^2_{med} / h^2_g and well-calibrated standard errors under a wide variety of simulated genetic and gene architectures for expression panel sample sizes > 500, whereas other methods cannot distinguish mediated from non-mediated effects.

4.Simulations and Real Data

- (7) Real Data
- Method:
- A.We applied MESC to estimate the proportion of heritability mediated by the cis-genetic component of assayed expression levels (h^2_{med}/ h^2_g) for 42 independent diseases and complex traits from the UK Biobank 38 and other publicly available datasets (average N = 323K; see Supplementary Table 1 for list of traits).

4.Simulations and Real Data

- B.In total, we produced three different types of expression scores:
- (1) Expression scores for each individual GTEx tissue,
- (2) Expression scores meta-analyzed within groups of GTEx tissues with common biological origin (Supplementary Table 2),
- (3) Expression scores meta-analyzed across all 48 GTEx tissues.

Each type of expression score was used to estimate h^2_{med} / h^2_g for each complex trait.

- C. To avoid biases, we partitioned genes by 5 expression cis-heritability bins and SNPs by the baseline LD model.

4.Simulations and Real Data

- (D)As independent validation, we used cis-eQTL summary statistics from eQTLGen (NeQTL= 31,684 in blood only) to estimate h^2_{med}/ h^2_g for the same 42 traits we analyzed above.

4.Simulations and Real Data

- Results：
- (1) We performed several analyses evaluating the robustness of these SNP and gene categories, finding that our estimates of $h^2_{med}/\ h^2_g$ were similar with other reasonable choices of SNP and gene categories but very biased when not partitioning genes or SNPs at all (Supplementary Note)
- (2) Across all 42 traits, we observed an average $h^2_{med}/\ h^2_g$ of 0.11 (S.E. 0.02) from the all-tissue meta-analyzed expression scores.
- (3)We did not observe a relationship between $h^2_{med}/\ h^2_g$ and h^2_g across traits (R 2 = 0.004) (Extended Data 1).
- (4)Of the 42 traits, 26 had estimates greater than 0 at nominal significance (p-value < 0.05), with 10 reaching Bonferroni significance (p-value < 0.05 / 42).

4.Simulations and Real Data

- (5) In Figure 3a, we report h^2_{med}/ h^2_g estimates from all-tissue and tissue-group meta-analyzed expression scores for a representative set of 10 genetically uncorrelated traits (full results in Extended Data 2 and Supplementary Table 3,4).
- (6) We observed consistently lower estimates h^2_{med}/ h^2_g from individual-tissue expression scores than from meta-tissue expression scores, as well as a positive correlation between tissue sample size and magnitude of individual-tissue h^2_{med}/ h^2_g (R 2 = 0.71) (Extended Data 3). suggesting downward biases in h^2_{med}/ h^2_g estimates due to low sample size.

4.Simulations and Real Data

- (7)
- In the independent validation, we obtained very similar $h^2_{med}/\ h^2_g$ estimates as GTEx all-tissue expression for blood/immune traits and lower $h^2_{med}/\ h^2_g$ for non-blood/immune traits (Extended Data 4, Supplementary Table 5),
- which is consistent with the fact that that eQTLGen only captures expression levels in blood while GTEx all-tissue meta-analysis captures expression levels across diverse tissues.

Discussion and Conclusion

- (1)We have developed a new method, mediated expression score regression (MESC), to estimate complex trait heritability mediated by the cis-genetic component of assayed expression levels (h^2_{med}) from GWAS summary statistics and eQTL effect sizes estimated from an external expression panel.
- (2) Our method is distinct from existing methods that identify and quantify overlap between eQTLs and GWAS hits (including colocalization
- tests , transcriptome-wide association studies , and heritability partitioning by eQTL status) in that it specifically aims to distinguish directional mediated effects from non-directional pleiotropic and linkage effects.
- (3) Moreover, our polygenic approach does not require individual eQTLs or GWAS loci to be significant and is not impacted by the sparsity of eQTL effect sizes, so unlike other approaches.

Discussion and Conclusion

- (4) we do not exclude genes or SNPs from our analyses based on any significance thresholds.
- (5)We applied our method to summary statistics for 42 traits and eQTL effect sizes estimated from 48 GTEx tissues.
- (6) We show that across traits, a significant but modest proportion of complex trait heritability (0.11 ± 0.02) is mediated by the cis-genetic component of assayed expression levels.

Discussion and Conclusion

- (7) Problems on the low h^2_{med}/ h^2_g:
- On the other hand, the fact that our h^2_{med}/ h^2_g estimates are low for most traits suggests that:
- eQTLs estimated from steady-state expression in bulk post-mortem tissues from GTEx do not capture most of the mediated effect of complex trait heritability.
- This motivats additional assays to better identify molecular mechanisms impacted by regulatory GWAS variants.

Discussion and Conclusion

- (8) Explanations for our low h^2_{med}/ h^2_g:
- Explanation A: The proportion of complex trait heritability mediated by the cis-genetic component of gene expression levels is in fact high in causal cell types/contexts for the trait,
- but eQTL data from bulk assayed tissues from GTEx is a poor proxy for eQTL data in causal cell types/contexts.

Discussion and Conclusion

- Explanation B:
- The proportion of complex trait heritability mediated by the cis-genetic component of gene expression levels is low even in causal cell types/contexts for the trait.
- In particular, complex trait heritability may be mediated in ways other than through gene expression levels in cis, including through protein-coding changes, splicing, or expression levels in trans.

Thank you!