Some Biology Knowledge in LD Score Regression (LD) and Mediated Expression Score Regression (MESC)

Date: 2022.7.1 　　　　　　　　Name：Chenxiao Tian

1. Some Basic Notations in Bioinformatics

eQTL：即表达数量性状基因座（Expression Quantitative Trait Loci）,比如体重、身高都是一个数量性状，则其对应的控制基因的位点就是一个数量性状基因, 而 eQTL 就是能控制数量性状基因（如身高基因）表达水平高低的那些位点。

在研究遗传突变与疾病的关系时，有时候这两者之间差的太远，不太直观，而 eQTL 则改为直接研究遗传突变与基因表达的相关性， 即改为用某个基因的差异表达作在中间斡旋（Mediated）:

突变 A（Genetic variants affect）——》B 基因表达变化（Regulation of gene expression levels）——》表现型 C（traits）。

cis-eQTL(顺式): 　　指的是与所调控的基因相距较近的 eQTL，一般多位于所调控基因的上下游 1Mb 位置。

trans-eQTL(反式): 　指的是与所调控的基因相距较远的 eQTL，远者可距离所调控基因 5Mb 的位置。

eQTL 分析的两个指标：SNP 和基因表达水平的关联度以及 SNP 与基因的距离。

SNP：进化过程中随机产生的单点突变，并能稳定的在群体中遗传。

GWAS：GWAS 相比 eQTL 更倾向于直接大范围分析基因位点与性状（疾病）的关系，即全基因组关联分析（Genome Wide Association Study，GWAS）是指在全基因组层面上，开展多中心、大样本、反复验证的基因与疾病的关联研究，是通过对大规模的群体 DNA 样本进行全基因组高密度遗传标记（如 SNP 或 CNV 等）分型，进而将基因型与可观测的性状，即表型，进行群体水平的统计学分析，根据统计量或显著性 p 值筛选出最有可能影响该性状的遗传变异（标记），全面揭示疾病发生、发展与治疗相关的遗传基因。

GWAS 的条件和局限：

（1）GWAS 通常需要大量的样本，以得到足够的 power，不过对每个样本实际操作时往往只做常见的10^6个位于编码区的 SNP 位点，区别于 WES（含所有编码区的 SNP 位点）和全基因检测 WGS（含所有编码区和非编码区）。

（2）GWAS 分析结果中，大部分显著的 SNP 位点都位于非编码区，很难直接挖掘这些位点的调控机制。通常假设与疾病关联的 SNP 位点先通过调控基因表达（Regulation of gene expression levels）再来发挥作用。

（3）据 GWAS 的结果来筛选基因时，只能筛选出显著关联的 SNP 位点所在的基因，这种做法结果类似前面提及的 cis-eQTL，会受到距离限制，无法全面挖掘后续基因。

GWAS 和 eQTL 的结合的优势：

eQTL 可以直接识别 SNP 与基因表达间的调控关系，若能将 eQTL 和 GWAS 结果相结合，可以进一步筛选候选基因。

遗传力（Heritability）：又称遗传率。遗传力是指遗传方差在表型总方差（表型总方差，即遗传方差与环境方差的和）中所占的比例。遗传力越大，表型越由遗传因素决定，环境因素相对较小。但数量性状如身高一般受到环境因素影响较大。

GTEx：一个收集正常人基因表达的数据集

2. Mediated Expression Score Regression (MESC)

1.方法的动机或者问题背景：Several different causal scenarios can result in similar patterns of overlap between GWAS loci and eQTLs, summarized in Figure 1a: (1) mediation （调解）, (2) pleiotropy (多效性) (3) linkage （连锁）
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).
所以如何区分或衡量（1）和（2）与（3）对 overlap 的贡献是一个重要的问题。

Remark：特别的，虽然只有(1) is informative of the SNP's mechanism of action on disease，但是（1）不一定是 overlap 的主要因素。

Example：在自身免疫疾病中，（3）linkage 可能比（1）mediation 更加显著。

2.本文具体解决的问题：quantify the proportion of disease heritability mediated in cis by assayed expression levels (对应（1）)

3.MESC：基于外部测定的 GWAS summary statistics, linkage disequilibrium (LD) scores，eQTL effect sizes 去估计(1)因素对应的 expression-mediated heritability.

MESC 区分（1），（2），（3）的思想：（1）mediation 相比于（2）和（2）因素，与 eQTL effect sizes 和 diseases effect sizes 的数量呈现一个线性关系（liner relationship）

4. MESC 在本文的具体应用：本文对来自 GTEx 数据库中的 42 种疾病和复杂性状，以及对应的 48 种组织的 cis-eQTL 数据，将 MESC 应用到 GWAS summary

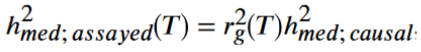
statistics 上。从而获得了对全基因集合和部分功能基因集的，（1）号 mediation 因素所占总的 disease heritability 的比例。

5.MESC Method 总览

h^2_{med}：Heritability mediated by the cis-genetic component of gene expression levels.

$h^2_{med;assayed}$(T)和$h^2_{med;casual}$：

前者是对实际测定的 T 基因集的 mediated 遗传力，后者则是对全部潜在的的具有因果因素基因集。本文没特别说明，文中使用h^2_{med}代表实际测定基因集 T 的 mediated 遗传力，即：



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

$h^2_{med}(D)$：代表其他感兴趣的基因集 D 实际测定的 mediated 遗传力。

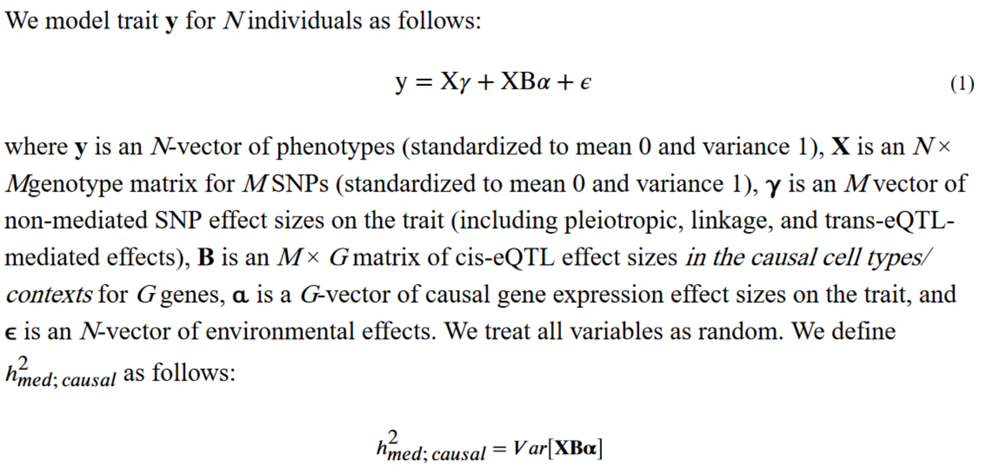
h^2_{med}的数学模型定义：y 是总的表现型向量

Xγ 一项是 non-mediated SNPs effect 的权重

XBα 是对应 mediated 的 cis-eQTL effect 所占的权重，注意这里是理论上的 casual 的情形不是 assayed，最终是可以导出具体的 assayed（D）的情形。

ε是环境因素的权重

则$h^2_{med;casual}$定义为 XBα 的方差：



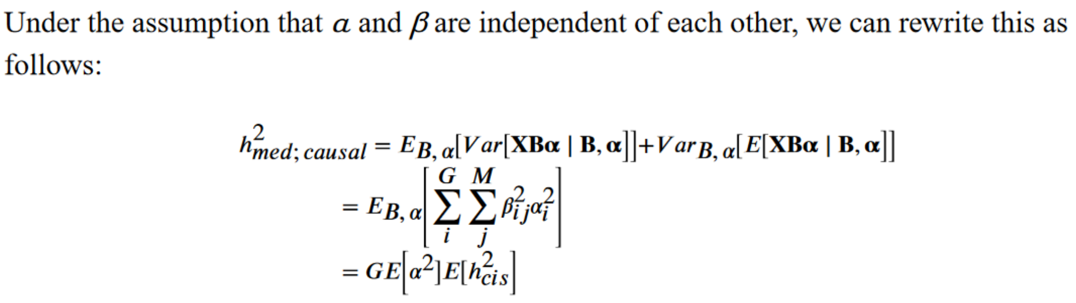
We model trait **y** for N individuals as follows:

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

where **y** is an N-vector of phenotypes (standardized to mean 0 and variance 1), **X** is an $N\times M$genotype matrix for MSNPs (standardized to mean 0 and variance 1), **γ** is an Mvector 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 Ggenes, **α** is a G-vector of causal gene expression effect sizes on the trait, and **ϵ** is an N-vector of environmental effects. We treat all variables as random. We define $h^2_{med;\,causal}$ as follows:

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

进一步$h^2_{med;casual}$的条件方差公式展开为 average squared per-gene effect of

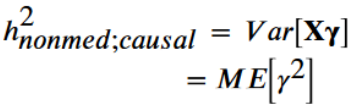


Under the assumption that α and β are independent of each other, we can rewrite this as follows:

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

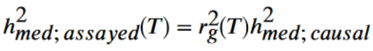
expression（mediated by gene expression）和 the average cis-heritability of expression across all genes 之乘积：

而 $h^2_{nonmed;casual}$:类似可定义并展开为：

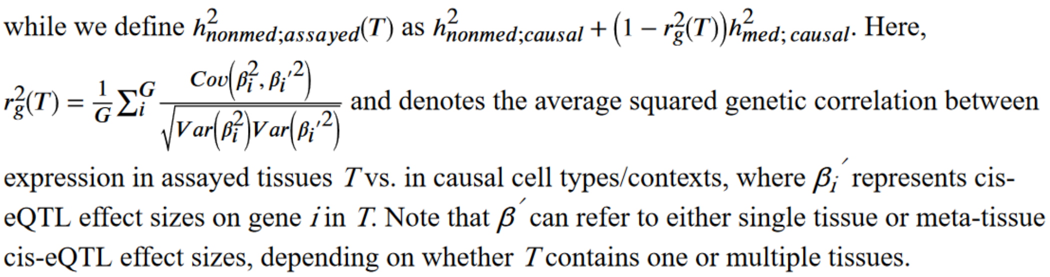


$$\begin{aligned} h^2_{nonmed;causal} &= Var[\mathbf{X\gamma}] \\ &= ME\left[\gamma^2\right] \end{aligned}$$

对 assayed 的情况和之前的相关系数r_g^2(T)，可以定义为：



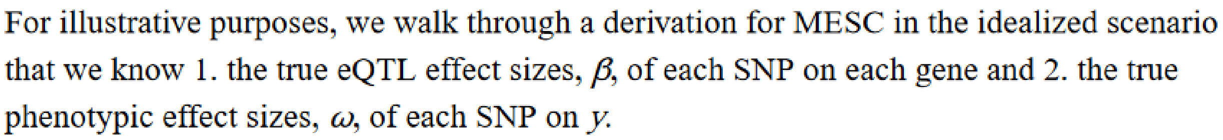
$$h^2_{med;\,assayed}(T) = r_g^2(T) h^2_{med;\,causal}$$



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

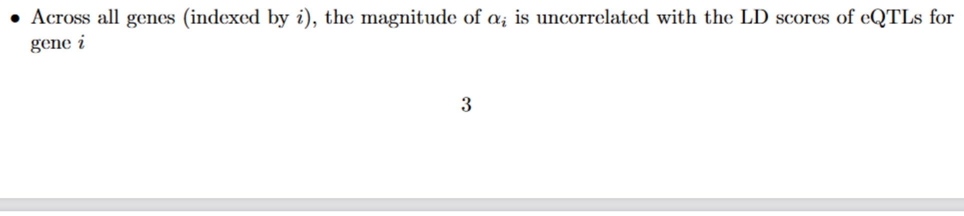
Unstratified MESC（只有一个基因组对h^2_{med}有影响的情况）

（原文中先推导的是知道如下两种具体信息的数值时的简化情形，后面在单独用两种方法区估计它们）

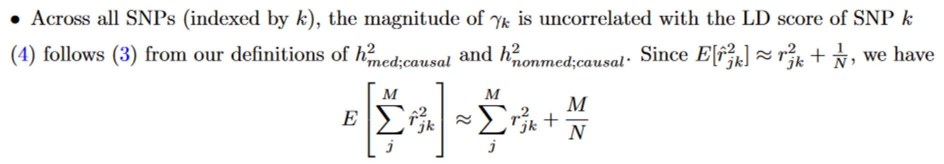


For illustrative purposes, we walk through a derivation for MESC in the idealized scenario that we know 1. the true eQTL effect sizes, β, of each SNP on each gene and 2. the true phenotypic effect sizes, ω, of each SNP on y.

其中推导出后面 Unstratified 情形的（2）号回归方程的独立性假设前提为：



- Across all genes (indexed by i), the magnitude of α_i is uncorrelated with the LD scores of eQTLs for gene i

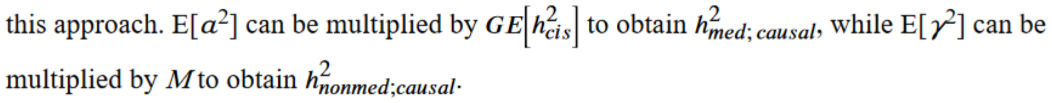


- Across all SNPs (indexed by k), the magnitude of γ_k is uncorrelated with the LD score of SNP k

(4) follows (3) from our definitions of $h^2_{med;causal}$ and $h^2_{nonmed;causal}$. Since $E[\hat{r}^2_{jk}] \approx r^2_{jk} + \frac{1}{N}$, we have

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

用回归方程（2）可估计$h^2_{nonmed;casual}$和$h^2_{med;casual}$
若对于 non-casual tissues T 的情形，可类似估计$h^2_{med;assayed}$和$h^2_{nonmed;assayed}$



this approach. E$[\alpha^2]$ can be multiplied by $GE\left[h^2_{cis}\right]$ to obtain $h^2_{med;\,causal}$, while E$[\gamma^2]$ can be multiplied by M to obtain $h^2_{nonmed;causal}$.

如下（2）为 Unstratified 情形的回归方程：

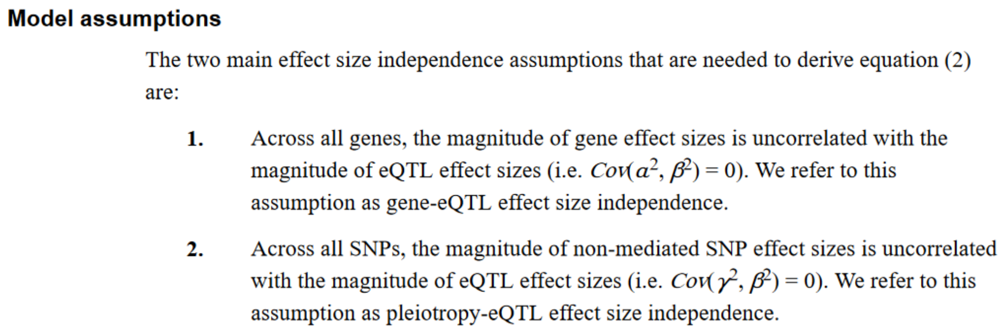
$$\mathrm{y} = \mathrm{X}\gamma + \mathrm{XB}\alpha + \epsilon \tag{1}$$

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

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

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

其他模型推导的独立性假设前提：



Model assumptions

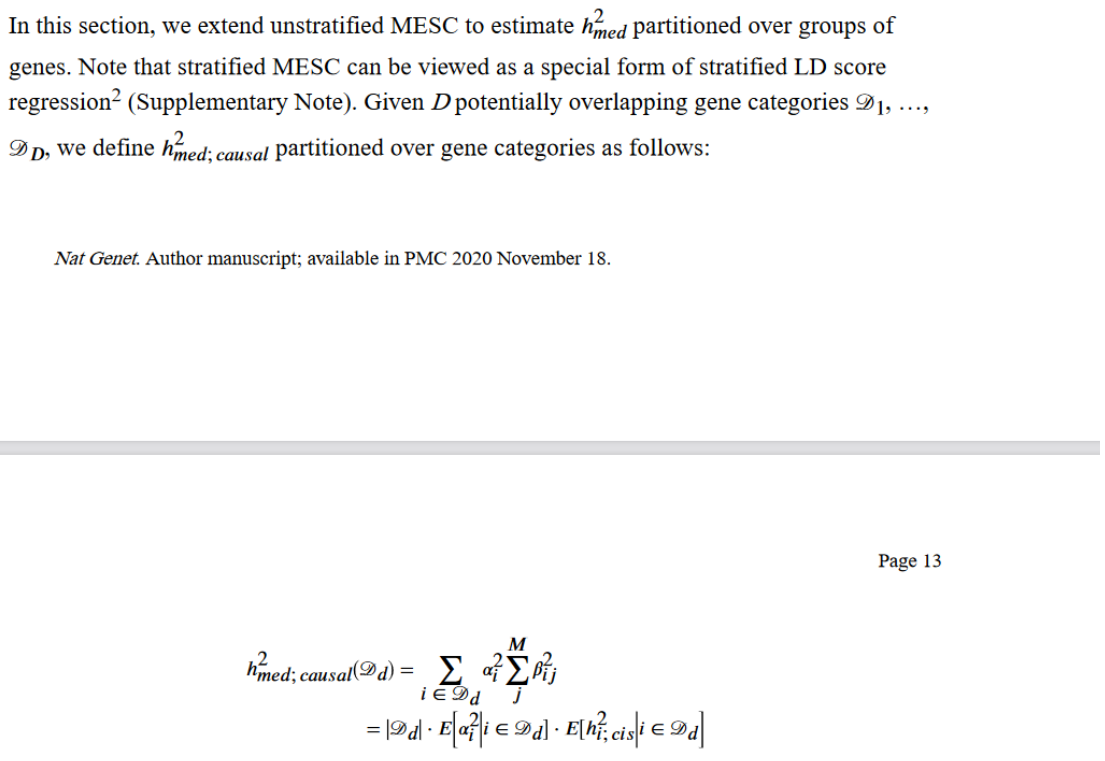
The two main effect size independence assumptions that are needed to derive equation (2) are:

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.

Stratified MESC（把 Unstratified MESC 推广到多个基因组对h^2_{med}有影响的情况）

多基因组 mediated 情形对$h^2_{med,casual}$的定义：



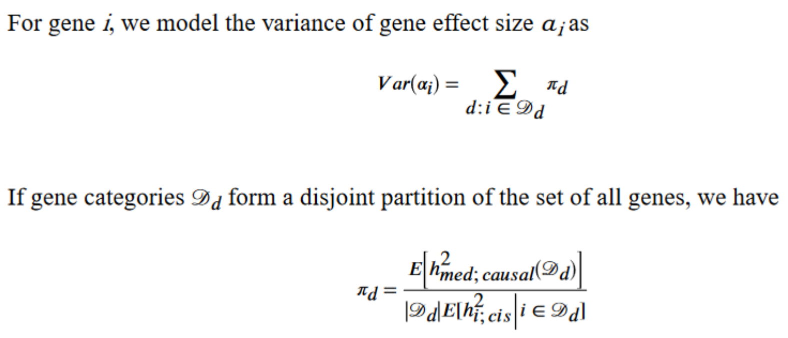
In this section, we extend unstratified MESC to estimate h^2_{med} partitioned over groups of genes. Note that stratified MESC can be viewed as a special form of stratified LD score regression[2] (Supplementary Note). Given D potentially overlapping gene categories $\mathcal{D}_1, \ldots, \mathcal{D}_D$, we define $h^2_{med;\,causal}$ partitioned over gene categories as follows:

Nat Genet. Author manuscript; available in PMC 2020 November 18.

Page 13

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

（mediated）gene effect size α_i的方差模型：



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

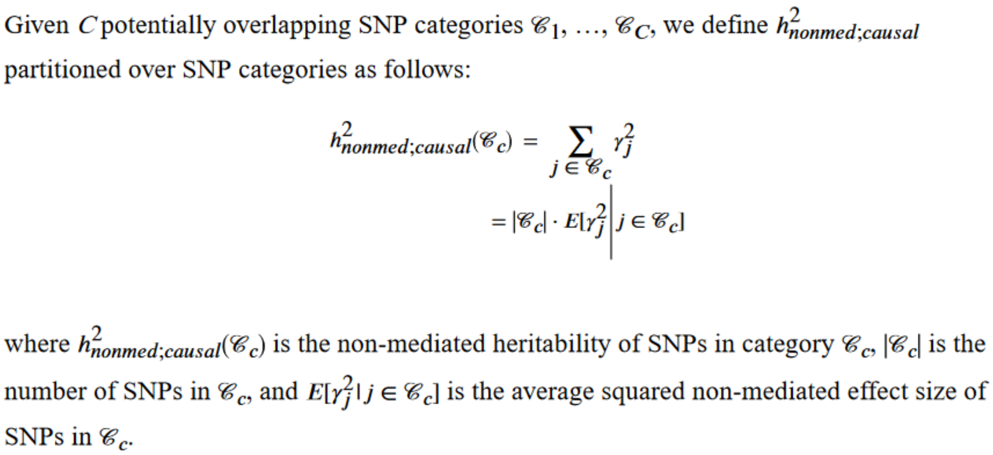
$$Var(\alpha_i) = \sum_{d:i \in \mathscr{D}_d} \pi_d$$

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

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

多基因组 non-mediated 情形对 h 的定义：

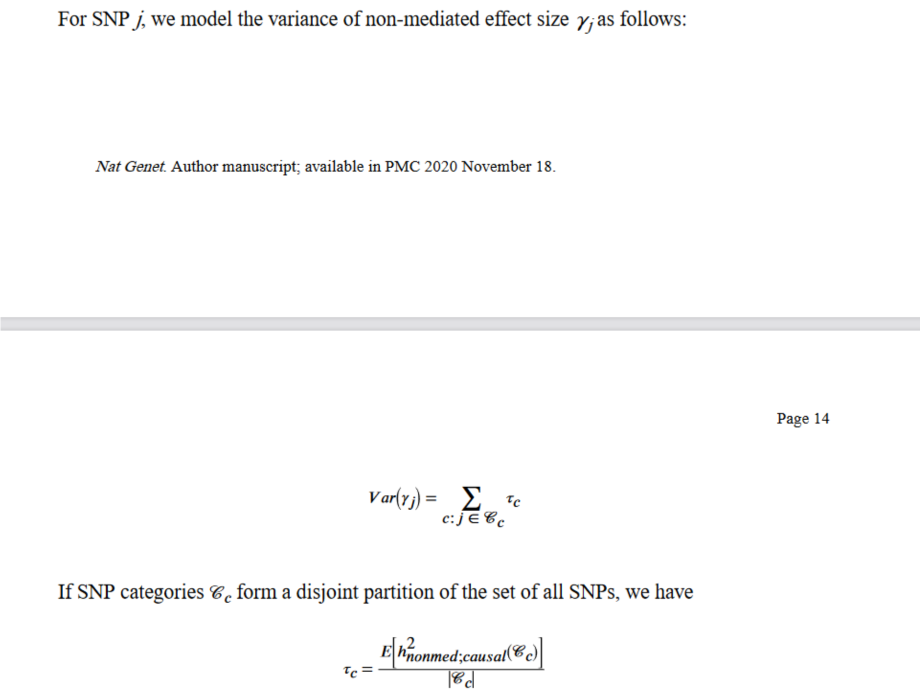
以及类似的 Non-mediated gene effect size 的方差模型：



Given C potentially overlapping SNP categories $\mathscr{C}_1, \ldots, \mathscr{C}_C$, we define $h^2_{nonmed;causal}$ partitioned over SNP categories as follows:

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

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 $\mathscr{C}_c$.



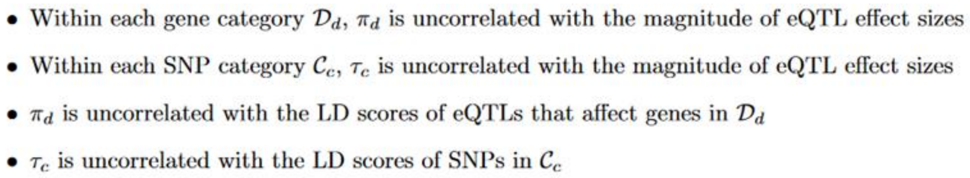
For SNP *j*, we model the variance of non-mediated effect size γ_j as follows:

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

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

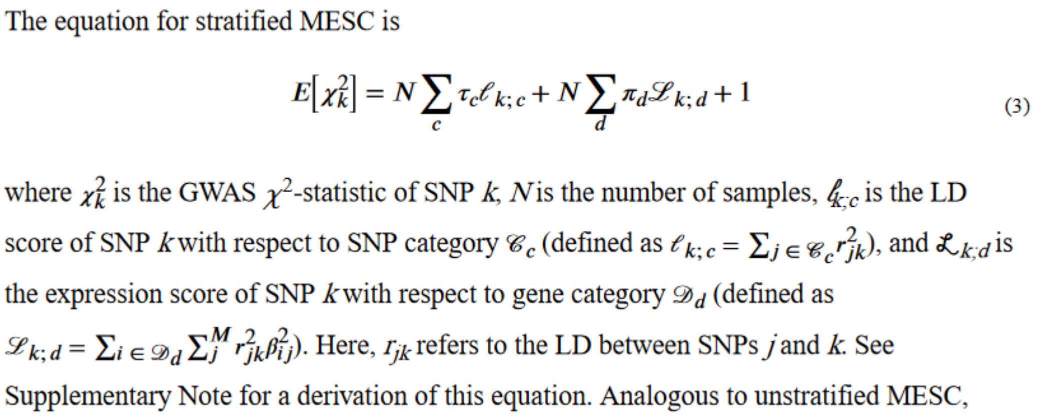
$$\tau_c = \frac{E\left[h^2_{nonmed;causal}(\mathscr{C}_c)\right]}{|\mathscr{C}_c|}$$

最终仍然由之前 unstratified 情形的（2），在如下四个假设下可以导出的 stratified 情形的回归方程：



- 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 $\mathcal{C}_c$

注：类似 unstratified 的情形，也可以进一步自然推广到 assayed h(T)的情形

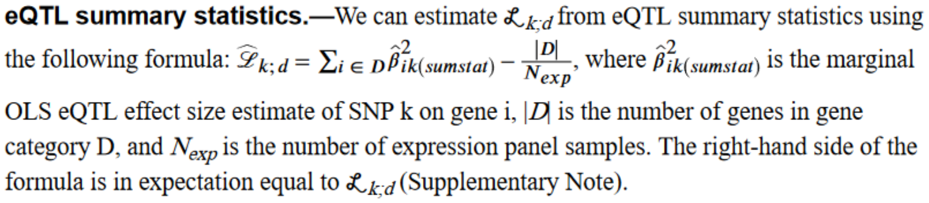


The equation for stratified MESC is

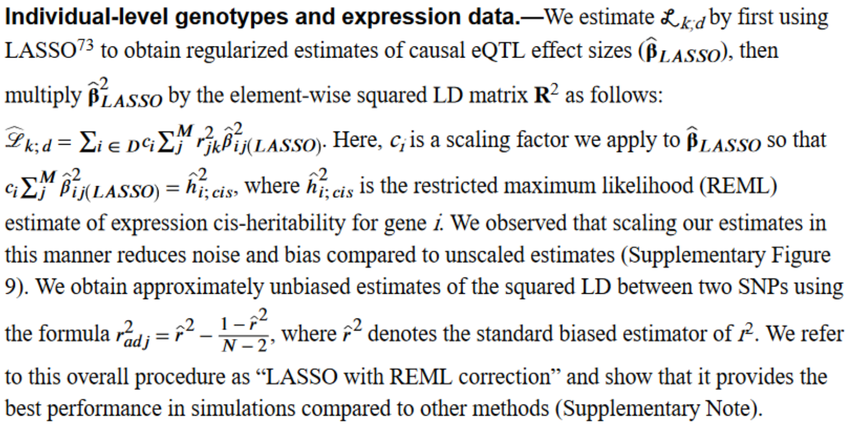
$$E\left[\chi_k^2\right] = N\sum_c \tau_c \ell_{k;c} + N\sum_d \pi_d \mathscr{L}_{k;d} + 1 \tag{3}$$

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 $\mathcal{L}_{k;d}$ 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,

额外最后 expression scores 参数$L_{k,d}$要提前单独用 eQTL summary statistics 或者 Individual-level genotypes and expression data 的方法区估计，后者会产生更少的噪声。



eQTL summary statistics.—We can estimate $\mathcal{L}_{k;d}$ from eQTL summary statistics using the following formula: $\widehat{\mathscr{L}}_{k;d} = \sum_{i\in D}\hat{\beta}_{ik(sumstat)}^2 - \frac{|D|}{N_{exp}}$, where $\hat{\beta}_{ik(sumstat)}^2$ is the marginal OLS eQTL effect size estimate of SNP k on gene i, $|D|$ is the number of genes in gene category D, and N_{exp} is the number of expression panel samples. The right-hand side of the formula is in expectation equal to $\mathcal{L}_{k;d}$ (Supplementary Note).



Individual-level genotypes and expression data.—We estimate $\mathcal{L}_{k;d}$ by first using LASSO[73] to obtain regularized estimates of causal eQTL effect sizes ($\widehat{\boldsymbol{\beta}}_{LASSO}$), then multiply $\widehat{\boldsymbol{\beta}}_{LASSO}^2$ by the element-wise squared LD matrix $\mathbf{R}^2$ as follows: $\widehat{\mathscr{L}}_{k;d} = \sum_{i\in D} c_i \sum_j^M r_{jk}^2 \hat{\beta}_{ij(LASSO)}^2$. Here, c_i is a scaling factor we apply to $\widehat{\boldsymbol{\beta}}_{LASSO}$ so that $c_i\sum_j^M \hat{\beta}_{ij(LASSO)}^2 = \hat{h}_{i;cis}^2$, where $\hat{h}_{i;cis}^2$ is the restricted maximum likelihood (REML) estimate of expression cis-heritability for gene i. We observed that scaling our estimates in this manner reduces noise and bias compared to unscaled estimates (Supplementary Figure 9). We obtain approximately unbiased estimates of the squared LD between two SNPs using the formula $r_{adj}^2 = \hat{r}^2 - \frac{1-\hat{r}^2}{N-2}$, where $\hat{r}^2$ denotes the standard biased estimator of r^2. We refer to this overall procedure as “LASSO with REML correction” and show that it provides the best performance in simulations compared to other methods (Supplementary Note).

3．LD Score Regression Summary

方法的动机：我们知道，通过 GWAS 分析可以识别到与表型相关的 SNP 位点，但是相关不等于就一定直接对应，因此这个结果并不一定真实客观的描述遗传因素对表型的效应，因为其实，这个结果是由以下两个因素共同构成的：

1. polygenic effects，基因对表型的效应
2. confounding factors，混淆因素，比如群落分层，样本间隐藏的亲缘关系等等

为了保证分析结果的准确性，LD Score 回归就是为了探究 1，2。比如混淆因素的占比，我们希望混淆因素占比比较低，这个时候分析才是有效的。

回归模型：
LDSC 本质是一个线性回归，其输入数据为 GWAS 的分析结果，具体来说：
回归的自变量：为 SNP 位点的 LD score 值，对于一个 SNP 位点，其 LD score 定义该位点与其邻近位点的连锁不平衡 R2 的总和，即如下表达式：

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2.$$

临近位点：对于一个 SNP 位点 j，取其邻近位点，比如 1CM 遗传距离，计算该遗传距离内内的其他位点与该位点的连锁不平衡情况下，用 R2 相加即得到了该位点的 LD score。

回归的因变量：是专门额外定义的一个符合卡方分布的统计量，其近似满足如下恒等式，

$$\mathbb{E}[\chi_j^2] \approx \frac{Nh_g^2}{M}\ell_j + 1.$$

其中 N 为样本总数，M 为窗口内的其他 SNP 位点数，h^2是遗传力，这几个值为常数.

从恒等式可以看出，卡方统计量和 LD score 之间是一个线性关系，而且对应到图像上，其截距为 1。

应用 1 判断混淆因素：上述公式是只假设只有遗传效应而无混淆因素的前提下得到，如果存在混淆因素，那么最后的截距就不是 1 了。通过 LDSC 回归分析的截距，可以判断 GWAS 结果中是否存在混淆因素。如果截距在 1 附近，说明没有混淆因素，如果解决超过这个范围，说明有混淆因素的存在。

应用 2 通过 LDSC 也可以评估遗传力的大小：因为公式回归系数含有遗传力。