A Novel Statistical Framework To Jointly Model Genetic and Epigenetic Regulation of Tissue-Specific Gene Expression

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen

February 7, 2016

Contents

1 Background & Motivation

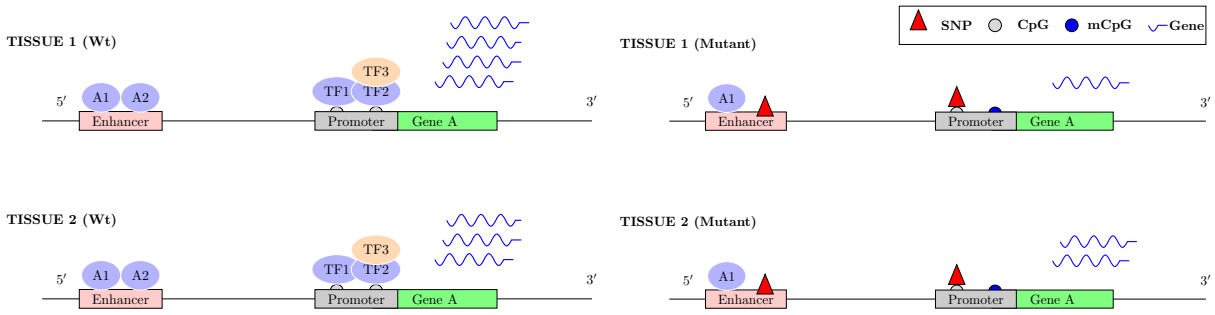


Figure 1: An illustration of tissue-specific gene expression of Gene A (quantified by blue squiggly lines) and the effect of the same genetic variant (denoted by red triangle labeled SNP) and the methylation status of the proximal CpG island (denoted by shaded semi-circle) in its expression in tissues 1 and 2

It has been long established that regulatory regions in higher eukaryotes activate gene transcription in a tissue-specific manner [13, 6]. These regulatory regions are susceptible to both genetic variation and epigenetic modifications that play a coordinated role in regulating tissue-specific gene expression [2, 7, 16, 20, 12]. One form of epigenetic variation is DNA methylation that targets non-methylated CpG sites, which constitute approximately 70% of all annotated promoters [3]. DNA methylation is linked to transcriptional silencing, and many CpG island promoters are active in a tissue-specific manner. Previous studies have shown that genetic variants at CpG sites are likely to disrupt DNA methylation and thus drastically change the methylation status at a single CpG site [19, 10, 9]. Since an increased DNA methylation at any of the CpG sites located in either the promoter or the intronic regions leads to a decreased gene expression, any variation in the CpG sites could lead to an opposite effect. For example, Catechol-O -methyltransferase (COMT) gene, which is implicated in schizophrenia has a functional single nucleotide polymorphism (SNP), Val^{158} Met (rs4680) that is associated with differential COMT expression across regions of the brain during the course of the illness [18]. More specifically, the substitution of a methionine (Met) for a valine (Val) at position 158 results in reduced activity of the COMT enzyme due to reduced protein stability. Besides the environmental stressors, methylation of CpG islands associated with the aforementioned variant seem to affect the expression of COMT across different brain regions[18]. Identifying and studying the mechanisms through which genetic variation, DNA methylation and gene expression interact may provide us with yet another clue to understanding complex phenotypes.

Genetic control of gene expression can be defined in terms of SNPs and their associations with gene expression. Expression quantitative trait loci (eQTL) are correlations between SNPs and gene expression. Similarly, epigenetic control of gene expression can be defined in terms of CpG island methylation and their interactions with SNPs. Expression quantitative trait methylations (eQTM) are correlations between gene expression and methylation while methylation quantitative trait loci (mQTL) are correlations between SNPs and methylation. Current approaches involve performing independent eQTL and mQTL analyses within each tissue followed by identifying pairs of statistically significant CpG-SNP and mRNA-SNP [7, 9]. These pairs are then expanded to

triplets of SNP and CpG-mRNA pair where the SNP was significantly correlated with either mRNA or CpG site of the CpG-mRNA pair. However, such an approach fails to fully exploit expression patterns across the tissues either by pooling information when a variant has a similar effect across multiple tissues or by explicitly identifying effects that differ across tissues. Even if we identify a variant that effects gene expression of a gene in a given tissue, it is not clear whether the effect is either due to an interaction effect between the genotype and tissue or an interaction effect between genotype, methylation status of a CpG site and tissues.

We propose a score test-based approach that explicitly models the aforementioned interactions among genotypes, methylation status and tissues as random effects. Our approach does not require parameter estimation under the alternative hypothesis, thus making it computationally tractable. Further, our score-based approach only requires estimation of first two moments of the random effects and, as a result, is robust to misspecification of the random effect distribution. We show using simulations that our joint score test approach is better than a tissue-by-tissue (TBT) approach.

We demonstrate the effectiveness of our method by applying it to a publicly available expression and methylation dataset from adult normal brains and show that by jointly analyzing multiple brain regions (tissues), we identify more mRNA-CpG-SNP triplets relative to a TBT analysis.

2 Our model

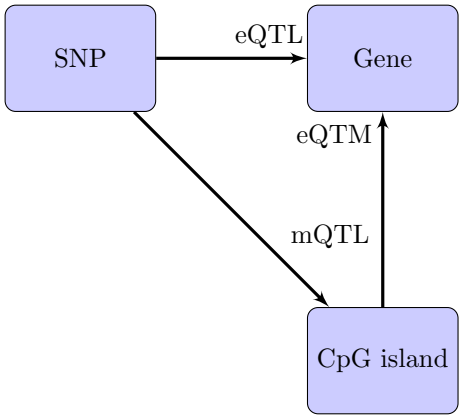


Figure 2: Tissue-specific gene expression is controlled by genetic, epigenetic and transcriptional regulatory mechanisms. Genetic control of gene expression can be defined in terms of SNPs and their associations with gene expression. Expression quantitative trait loci (eQTL) are correlations between SNPs and gene expression. Similarly, epigenetic control of gene expression can be defined in terms of CpG island methylation and their interactions with SNPs. Expression quantitative trait methylations (eQTM) are correlations between gene expression and methylation while methylation quantitative trait loci (mQTL) are correlations between SNPs and methylation. Gene expression can be modeled as a function of SNP and CpG islands (if all the requisite data types are available).

For a given gene-SNP pair, gene expression is modeled as a function of genotype and methylation -

$$Y = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi \tag{1}$$

where Y is a nt-dimensional vector of expression levels in t tissues and n individuals, α is a vector of tissue-specific intercepts, G is a nt-dimensional vector of genotypes, β is a fixed effect of genotype across tissue, λ is an overall methylation-specific fixed effect, ϕ is genotype $\times$ methylation interaction effect (fixed effect), $u \sim N\left(0, \tau AA^T\right)$ is a vector of subject-specific random effect, $v \sim N\left(0, \gamma BB^T\right)$ is a vector of tissue-specific random effects, $w \sim N\left(0, \delta CC^T\right)$ is a vector of tissue-specific random effects that describe the interaction effect between genotype and methylation is a vector of random effects describing the interaction between genotype, methylation and tissue, $x \sim N\left(0, \theta DD^T\right)$ is a vector of tissue-specific random effects describing methylation

effects and $\xi \sim N(0, \epsilon I_{nt})$. The matrices J, A, B, C, and D are design matrices with B being a function of genotype, C is a function of both genotype and methylation data and finally, D is a function of just the methylation data. J is $nt \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. A is $nt \times nt$ design matrix for the subject-specific intercepts. B is a $nt \times t$ design matrix of stacked genotypes. C is a $nt \times t$ design matrix of stacked (product of) tissue-specific methylation and genotype data. D is $nt \times t$ design matrices of stacked tissue-specific methylation data.

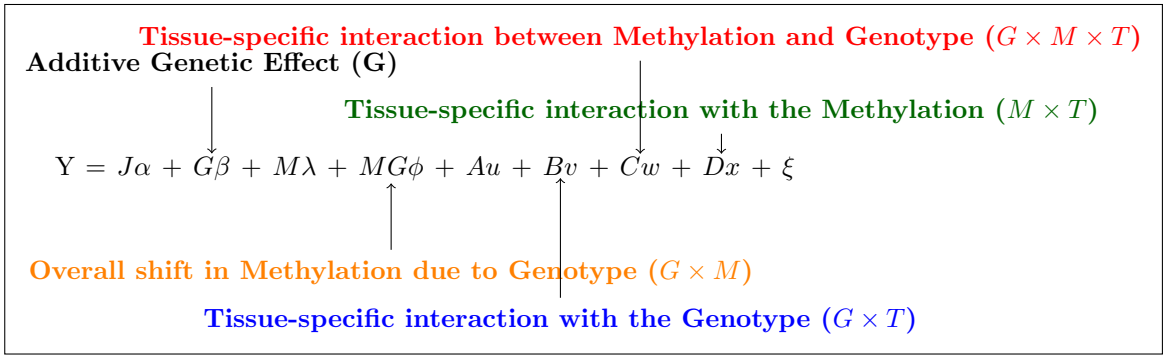


Figure 3: Model Description

Parameters of interest are γ, δ, β and ϕ; α, λ, τ, θ and ϵ are nuisance parameters. We test the null hypothesis that $H_0 : \beta = \phi = \gamma = \delta = 0$, i.e. the variant does not affect gene expression across any of the tissues. To do so, we compute the efficient scores for γ, δ, β and ϕ by projecting off components correlated with the nuisance parameters.

From equation 1, the log-likelihood function of Y conditioned on the genotype is –

$$\ell(\Theta; Y) = -c - \frac{1}{2}\log \mid \Sigma \mid -\frac{1}{2}(Y - J\alpha - G\beta - M\lambda - MG\phi)^T \Sigma^{-1} (Y - J\alpha - G\beta - M\lambda - MG\phi) \quad (2)$$

where Θ represents the vector of all the variance components involved in Σ and c is a constant. Alternatively, under normality, we have

$$Y \sim N(J\alpha + G\beta + M\lambda + MG\phi, \Sigma)$$

The efficient scores evaluated under the null are given by –

$$U_\beta = \hat{Y}^T \hat{\Sigma}_n^{-1} \left(G - \bar{G}\right)$$

$$U_\phi = \hat{Y}^T \hat{\Sigma}_n^{-1} \left(MG - \overline{MG}\right)$$

$$U_\gamma = \frac{1}{2} \hat{Y}^T \hat{\Sigma}_n^{-1} B B^T \hat{\Sigma}_n^{-1} \hat{Y}$$

$$U_\delta = \frac{1}{2} \hat{Y}^T \hat{\Sigma}_n^{-1} C C^T \hat{\Sigma}_n^{-1} \hat{Y}$$

where $\hat{Y}$ are the residuals and $\hat{\Sigma} = \hat{\epsilon} I + \hat{\tau} Z Z^T + \hat{\theta} D D^T$.

We propose a weighted sum of the above components to arrive at our joint score test statistic, U_ψ. Since U_β and U_ϕ are linear in Y while U_γ and U_δ are quadratic, we propose the following rule to combine them –

$$
\begin{aligned}
U_\psi &\equiv \left(\boldsymbol{a_\beta} U_\beta^2 + \boldsymbol{a_\phi} U_\phi^2 + \boldsymbol{a_\gamma} U_\gamma + \boldsymbol{a_\delta} U_\delta\right) \\
&\equiv \hat{Y}^T \hat{\Sigma}_n^{-1} \left[a_\beta \left(G - \bar{G}\right)\left(G - \bar{G}\right)^T + a_\phi \left(MG - \overline{MG}\right)\left(MG - \overline{MG}\right)^T + a_\gamma \frac{1}{2} BB^T + a_\delta \frac{1}{2} CC^T \right] \hat{\Sigma}_n^{-1} \hat{Y}
\end{aligned} \tag{3}
$$

where a_β, a_ϕ, a_γ and a_δ are scalar constants chosen to minimize the variance of U_ψ. Under the null, U_ψ is distributed as a mixture of chi-square random variables. We use Satterthwaite method [15] to approximate the p values from a scaled χ^2 distribution by matching the first two moments as $U_\psi \sim \kappa \chi^2_\nu$ where $\kappa = \frac{2Var(U_\psi)}{E[U_\psi]}$ and $\nu = \frac{2E[U_\psi]^2}{Var(U_\psi)}$.

Our joint score test will test for the effect of genotype on 1) an overall shift in the gene expression, 2) tissue-specific interaction ($G \times T$), 3) overall methylation ($G \times M$), and 4) tissue-specific methylation ($G \times M \times T$)

3 Individual components of our joint score test statistic

3.1 Score test for the additive genetic effect on the gene expression under the global null

Gene Expression $\longrightarrow Y = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi$

$\uparrow$

Additive genetic effect

The score test for the fixed effect β takes the following form under the global null –

$$
U_\beta = \left(G - \bar{G}\right)^T \Sigma_n^{-1} \hat{Y} \tag{4}
$$

where $\left(G - \bar{G}\right)$ is a vector of mean-centered genotypes for all individuals and $\hat{Y} = \left(Y - J\hat{\alpha} - M\hat{\lambda}\right)$. U_β is a scalar quantity in a linear form and follows a χ^2_1 distribution.

Squaring U_β gives us the following quadratic form, which will be useful while aggregating all the score test statistics.

$$
U_\beta^2 = \hat{Y}^T \Sigma_n^{-1} \left(G - \bar{G}\right)\left(G - \bar{G}\right)^T \Sigma_n^{-1} \hat{Y} \tag{5}
$$

3.2 Variance-component score test for the tissue-specific effect due to genotype on the gene expression under the global null ($G \times T$)

Gene Expression $\longrightarrow Y = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi$

$\uparrow$

Interaction effect between genotype and tissues ($G \times T$)

The score for the variance component $\boldsymbol{\gamma}$ under the global null is –

$$\frac{1}{2}\left\{\hat{Y}^T\Sigma_n^{-1}BB^T\Sigma_n^{-1}\hat{Y} - Tr\left(\Sigma_n^{-1}BB^T\right)\right\} \tag{6}$$

where $\Sigma_n = diag\left(\Sigma, \cdots, \Sigma\right)$ is an $nt \times nt$ block diagonal matrix and $\hat{Y} = \left(Y - J\hat{\alpha} - M\hat{\lambda}\right)$. As the *trace* term does not depend on the data, we use the first term to construct the test statistic.

$$U_\gamma = \frac{1}{2}\hat{Y}^T\Sigma_n^{-1}BB^T\Sigma_n^{-1}\hat{Y} \tag{7}$$

U_γ follows a mixture of chi-square distribution and the p value is approximated using a scaled χ^2 distribution (the Satterthwaite method) by matching the first two moments as $U_\gamma \sim \kappa\chi^2_\nu$ where $\kappa = \frac{2Var(U_\gamma)}{E[U_\gamma]}$ and $\nu = \frac{2E[U_\gamma]^2}{Var(U_\gamma)}$.

3.3 Latent effect (masking effect) of SNP on gene expression via tissue-specific methylation patterns ($G \times M \times T$)

Gene Expression $\longrightarrow \mathrm{Y} = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi$

Cw: **Three-way interaction between methylation, genotype and tissue ($G \times M \times T$)**

The score for the variance component $\boldsymbol{\delta}$ under the global null is –

$$\frac{1}{2}\left\{\hat{Y}^T\Sigma_n^{-1}CC^T\Sigma_n^{-1}\hat{Y} - Tr\left(\Sigma_n^{-1}CC^T\right)\right\} \tag{8}$$

where $\Sigma_n = diag\left(\Sigma, \cdots, \Sigma\right)$ is an $nt \times nt$ block diagonal matrix and $\hat{Y} = \left(Y - J\hat{\alpha} - M\hat{\lambda}\right)$. As the *trace* term does not depend on the data, we use the first term to construct the test statistic.

$$U_\delta = \frac{1}{2}\hat{Y}^T\Sigma_n^{-1}CC^T\Sigma_n^{-1}\hat{Y} \tag{9}$$

U_δ follows a mixture of chi-square distribution, the p value can be approximated using a scaled χ^2 distribution by matching the first two moments as $U_\delta \sim \kappa\chi^2_\nu$ where $\kappa = \frac{2Var(U_\delta)}{E[U_\delta]}$ and $\nu = \frac{2E[U_\delta]^2}{Var(U_\delta)}$.

3.4 The effect of SNP on gene expression via differential methylation patterns under the global null ($G \times M$)

Gene Expression $\longrightarrow \mathrm{Y} = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi$

$MG\phi$: **Interaction effect between methylation and genotype ($G \times M$)**

The score test for the interaction effect ϕ takes the following form under the global null –

$$U_\phi = \left(MG - \overline{MG}\right)^T\Sigma_n^{-1}\hat{Y} \tag{10}$$

where $\Sigma_n = diag(\Sigma, \cdots, \Sigma)$ is an $nt \times nt$ block diagonal matrix and $\hat{Y} = \left(Y - J\hat{\alpha} - M\hat{\lambda}\right)$. U_ϕ is a scalar quantity in a linear form and follows a χ^2_1 distribution. Squaring U_ϕ gives us the following quadratic form, which will be useful while aggregating all the score test statistics.

$$U^2_\phi = \hat{Y}^T \Sigma_n^{-1} \left(MG - \overline{MG}\right) \left(MG - \overline{MG}\right)^T \Sigma_n^{-1} \hat{Y} \tag{11}$$

4 Simulations

General descriptions of the simulations –

- For each tissue, we model the potential genetic association between a target SNP and the expression levels of a target gene at a single locus.
- Methylation data is generated independently from a multivariate normal distribution with a positive semi-definite variance-covariance matrix.
- We have generated one gene and one SNP at a time with the genotypes at each SNP in each individual simulated assuming Hardy-Weinberg equilibrium with varying minor allele frequencies (MAF).
- All the simulations are performed on 5 tissues, 100 observations in each tissue, and genotypes follow common variant frequency (MAF = 0.30).
- For all null simulations, $\theta = \tau = \epsilon = 1$.
- The proportion of variance explained (PVE) by γ and δ are defined as:

$$PVE_\gamma \equiv \left(\frac{\gamma}{\theta + \tau + \epsilon + \gamma + \delta}\right) \qquad PVE_\delta \equiv \left(\frac{\delta}{\theta + \tau + \epsilon + \gamma + \delta}\right)$$

4.1 Evaluating the joint score test

Each simulated dataset was comprised of data from a single locus and a single gene, whose expression is measured across 5 tissues in 100 observations. For a given gene-SNP pair, the genotypes at each SNP in all the individuals were simulated as Binomial(2,0.3), i.e. a minor allele frequency 30% and assuming Hardy-Weinberg equilibrium. Methylation data for 5 tissues was generated from a multivariate normal distribution with zero mean and an identity matrix as a covariance matrix. The gene expression data are generated prospectively, i.e., first genotypes are generated, followed by gene expression according to the following equation –

$$y_{it} = \alpha_t + \beta_i g_i + \lambda_i m_{it} + \phi_i m_{it} g_i + a_i + b_t g_i + c_t m_{it} g_i + d_t m_{it} + \xi_{it} \qquad \xi \overset{i.i.d.}{\sim} N(0, \epsilon) \tag{12}$$

where y_{it} is a vector of gene expression data, α_t is the tissue-specific intercept, β describes the main additive genotypic effect, λ describes the overall effect due to methylation, ϕ describes the interaction effect between the overall methylation and genotype, g is the value of a bi-allelic genotype such that $g \in (0, 1, 2)$ represents the number of copies of the minor allele. The random effect b_t represents tissue-specific effect of the genotype, c_t represents tissue-specific interaction effect between methylation and genotype, d_t represents tissue-specific methylation effect, and a_i is a subject-specific random intercept. We assume that the random effects are independent and that $a_i \sim N_t(0, \tau)$, $b_t \sim N_t(0, \gamma)$, $c_t \sim N_t(0, \delta)$ and $d_t \sim N_t(0, \theta)$.

We use 1,000 data replicates to evaluate type I error and power calculations. Simulations were performed by varying β, the proportion of variance explained by the random effect describing the interaction between genotype and tissue, $PVE_\gamma \equiv \left(\frac{\gamma}{\theta+\tau+\epsilon+\gamma+\delta}\right)$, and the proportion of variance explained by the random effect describing the interaction between genotype, methylation and tissue, $PVE_\delta \equiv \left(\frac{\delta}{\theta+\tau+\epsilon+\gamma+\delta}\right)$. A linear mixed effects model was fit using the package *lme4* [1] in the statistical environment R (R Core Team) [11].

β	ϕ	$PVE_\delta(\%)$	$PVE_\gamma(\%)$	$U_{\beta\|H_0}$	$U_{\phi\|H_0}$	$U_{\gamma\|H_0}$	$U_{\delta\|H_0}$	$U_{\psi\|H_0}$
NO	NO	0	0	0.058	0.045	0.061	0.053	0.06
NO	NO	0	5	0.071	0.055	0.278	0.052	0.161
NO	NO	0	8	0.092	0.064	0.602	0.062	0.427
NO	NO	5	0	0.053	0.151	0.047	0.28	0.173
NO	NO	5	5	0.079	0.153	0.294	0.288	0.325
NO	NO	5	8	0.107	0.143	0.641	0.274	0.571
NO	NO	8	0	0.055	0.251	0.072	0.549	0.383
NO	NO	8	5	0.081	0.255	0.312	0.622	0.585
NO	NO	8	8	0.107	0.263	0.645	0.604	0.734
NO	YES	0	0	0.058	0.883	0.039	0.674	0.171
NO	YES	0	5	0.08	0.884	0.28	0.696	0.385
NO	YES	0	8	0.101	0.888	0.629	0.674	0.616
NO	YES	5	0	0.047	0.825	0.061	0.772	0.381
NO	YES	5	5	0.065	0.844	0.314	0.751	0.525
NO	YES	5	8	0.102	0.834	0.611	0.747	0.725
NO	YES	8	0	0.071	0.762	0.072	0.83	0.573
NO	YES	8	5	0.084	0.76	0.309	0.837	0.677
NO	YES	8	8	0.099	0.719	0.579	0.848	0.826
YES	NO	0	0	0.287	0.05	0.054	0.045	0.208
YES	NO	0	5	0.308	0.058	0.295	0.055	0.357
YES	NO	0	8	0.322	0.059	0.648	0.056	0.588
YES	NO	5	0	0.303	0.127	0.053	0.275	0.355
YES	NO	5	5	0.301	0.147	0.291	0.253	0.484
YES	NO	5	8	0.356	0.116	0.642	0.249	0.689
YES	NO	8	0	0.325	0.279	0.064	0.566	0.585
YES	NO	8	5	0.323	0.268	0.306	0.584	0.68
YES	NO	8	8	0.329	0.284	0.631	0.606	0.823
YES	YES	0	0	0.322	0.916	0.062	0.693	0.421
YES	YES	0	5	0.327	0.88	0.308	0.669	0.552
YES	YES	0	8	0.341	0.89	0.637	0.691	0.74
YES	YES	5	0	0.318	0.81	0.084	0.742	0.57
YES	YES	5	5	0.305	0.809	0.294	0.734	0.666
YES	YES	5	8	0.349	0.802	0.589	0.757	0.809
YES	YES	8	0	0.288	0.761	0.076	0.832	0.705
YES	YES	8	5	0.32	0.767	0.333	0.84	0.816
YES	YES	8	8	0.351	0.737	0.623	0.817	0.869

Table 1: Table comparing the statistical power of the joint score test statistic, U_ψ and the contributions from its main components, U_β, U_γ and U_δ, all under the global null. These data were generated from 1,000 simulations run on 100 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3).

4.2 Comparing tissue-by-tissue approach with our joint score test statistic

A tissue-by-tissue (TBT) analysis fits the following model **Gene ~ CpG + Geno + Geno:CpG** in individual tissues. For tissue t and SNP j, the gene expression is modeled as a function of the genotype and methylation data:

$$Y_t = \alpha_j m_t + \beta_t g_j + \delta_t g_j m_t + \epsilon_t \tag{13}$$

where Y_t is the gene expression g of individuals in tissue t, g_j is the genotype data on SNP j, m_t is methylation information in tissue t.

β	ϕ	$PVE_\delta(\%)$	$PVE_\gamma(\%)$	TBT	U_ψ
NO	NO	0	0	0.053	0.056
NO	NO	0	7	0.097	0.171
NO	NO	0	10	0.222	0.425
NO	NO	7	0	0.18	0.195
NO	NO	7	7	0.202	0.303
NO	NO	7	10	0.368	0.55
NO	NO	10	0	0.426	0.429
NO	NO	10	7	0.453	0.523
NO	NO	10	10	0.554	0.719
NO	YES	0	0	0.325	0.179
NO	YES	0	7	0.396	0.361
NO	YES	0	10	0.522	0.618
NO	YES	7	0	0.519	0.355
NO	YES	7	7	0.584	0.519
NO	YES	7	10	0.66	0.706
NO	YES	10	0	0.669	0.588
NO	YES	10	7	0.706	0.697
NO	YES	10	10	0.779	0.805
YES	NO	0	0	0.143	0.235
YES	NO	0	7	0.248	0.365
YES	NO	0	10	0.392	0.586
YES	NO	7	0	0.288	0.395
YES	NO	7	7	0.393	0.526
YES	NO	7	10	0.512	0.698
YES	NO	10	0	0.514	0.566
YES	NO	10	7	0.589	0.696
YES	NO	10	10	0.683	0.812
YES	YES	0	0	0.48	0.4
YES	YES	0	7	0.549	0.541
YES	YES	0	10	0.678	0.747
YES	YES	7	0	0.641	0.588
YES	YES	7	7	0.686	0.683
YES	YES	7	10	0.754	0.83
YES	YES	10	0	0.772	0.721
YES	YES	10	7	0.765	0.789
YES	YES	10	10	0.847	0.898

Table 2: Table comparing the statistical power of the joint score test statistic, U_ψ and TBT approach. This data were generated from 1,000 simulations run on 100 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3).

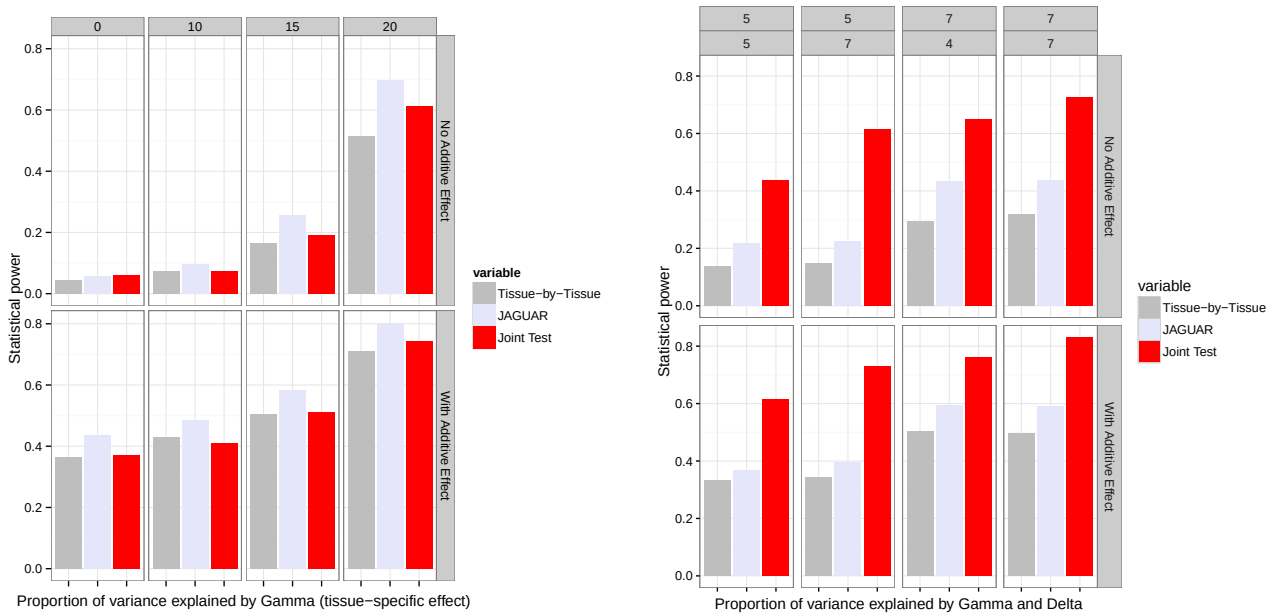
4.3 Testing the significance of the methylation effect in eQTL identification

In the absence of methylation data or when DNA methylation is not effecting gene expression equation 1 can be reduced to -

$$Y = J\alpha + G\beta + Au + Bv + \xi \qquad Y \sim N\left(J\alpha + G\beta, \sigma\right) \tag{14}$$

We ask if including the terms associated with the methylation data to the above equation may lead to any substantial power loss. In order to do this, we keep all the variance components and fixed effects associated with methylation to a zero (i.e. $\lambda = \phi = \delta = \theta = 0$). Statistical power from the joint analysis of genotype and tissue-specific interaction using JAGUAR was compared with our new method. The loss of power is not substantial as seen in Figure 6, which basically tells us that the availability of methylation data will only help

us and not hurt us as much with respect to eQTL identification.



(a) In the absence of methylation effect

(b) In the presence of methylation effect

Figure 4: Comparing TBT, JAGUAR and the joint score test approaches. a) In the absence of methylation data, statistical power from the joint analysis of genotype and tissue-specific interaction using JAGUAR is marginally better than our joint score test. A tissue-by-tissue method is also used for comparison. b) On the other hand, in the presence of DNA methylation effect our method out performs JAGUAR and tissue-by-tissue analyses. The top two rows in the figure indicate PVE_γ and PVE_δ. These data were generated from 1,000 simulations run on 100 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3).

When methylation data is available or when DNA methylation affects gene expression, we ask how JAGUAR compares with our method with respect to statistical power. As expected, our joint score test method does better than JAGUAR and tissue-by-tissue approach.

4.4 Comparing the joint score test with $G \times M \times T$ interaction effect using a variance-component score test

Given that –

Gene Expression $\longrightarrow Y = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi$

Tissue-specific interaction between methylation and genotype

where Y is a nt-dimensional vector of expression levels in t tissues and n individuals, α is a vector of tissue-specific intercepts, G is a nt-dimensional vector of genotypes, β is a fixed effect of genotype across tissue, λ is an overall methylation-specific fixed effect, ϕ is genotype $\times$ methylation interaction effect (fixed effect), $u \sim N\left(0, \tau AA^T\right)$ is a vector of subject-specific random effect, $v \sim N\left(0, \gamma BB^T\right)$ is a vector of tissue-specific

random effects, $w \sim N\left(0, \delta CC^T\right)$ is a vector of tissue-specific random effects that describe the interaction effect between genotype and methylation is a vector of random effects describing the interaction between genotype, methylation and tissue, $x \sim N\left(0, \theta DD^T\right)$ is a vector of tissue-specific random effects describing methylation effects and $\xi \sim N\left(0, \epsilon I_{nt}\right)$. The matrices J, A, B, C, and D are design matrices with B being a function of genotype, C is a function of both genotype and methylation data and finally, D is a function of just the methylation data. J is $nt \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. A is $nt \times nt$ design matrix for the subject-specific intercepts. B is a $nt \times t$ design matrix of stacked genotypes. C is a $nt \times t$ design matrix of stacked (product of) tissue-specific methylation and genotype data. D is $nt \times t$ design matrices of stacked tissue-specific methylation data.

Parameter of interest is δ; α, β, λ, ϕ, τ, γ, θ and ϵ are nuisance parameters. We test the null hypothesis that $H_0 : \delta = 0$, i.e. there is no effect of genotype on tissue-specific methylation expression pattern. To do so, we compute the efficient scores for δ by projecting off components correlated with the nuisance parameters.

The log-likelihood function of Y conditioned on the genotype is –

$$\ell\left(\Theta; Y\right) = -c - \frac{1}{2}\log \mid \Sigma \mid -\frac{1}{2}\left(Y - J\alpha - G\beta - M\lambda - MG\phi\right)^T \Sigma^{-1}\left(Y - J\alpha - G\beta - M\lambda - MG\phi\right) \tag{15}$$

where Θ represents the vector of all the variance components involved in Σ and c is a constant. Alternatively, under normality, we have

$$Y \sim N(J\alpha + G\beta + M\lambda + MG\phi, \Sigma)$$

The efficient scores evaluated under the null are given by –

$$\frac{1}{2}\left\{\hat{Y}^T \Sigma_n^{-1} CC^T \Sigma_n^{-1} \hat{Y} - Tr\left(\Sigma_n^{-1} CC^T\right)\right\} \tag{16}$$

In order to test the effect of interaction between SNP and tissue-specific methylation data on gene expression data, we test the following hypothesis –

$$H_0 : \delta = 0 \quad H_A : \delta > 0$$

	Hypotheses tested	Parameters of interest	Nuisance parameters
$G \times M \times T$	$H_0 : \delta = 0 \quad H_A : \delta > 0$	δ	α, λ, β, ϕ, γ, τ and ϵ
Joint score test	$H_0 : \beta = \Theta = 0 \quad H_A : \Theta > 0; \beta \neq 0$	β, γ, δ and ϕ	θ, α, λ, τ and ϵ

Table 3: Hypothesis description

A score test statistic can be constructed to test for this effect by projecting off all the components associated with the nuisance parameters.

$$\ddot{U}_\delta = \hat{Y}^T \hat{\Sigma}^{-1} CC^T \hat{\Sigma}^{-1} \hat{Y}$$

where $\hat{\Sigma} = \hat{\tau} AA^T + \hat{\gamma} BB^T + \hat{\theta} DD^T + \hat{\epsilon} I_n$. Under the null, $\ddot{U}_\delta$ is distributed as a mixture of chi-square random variables. We use Satterthwaite method to approximate the p values from a scaled χ^2 distribution by matching the first two moments as $\ddot{U}_\delta \sim \kappa \chi^2_\nu$ where $\kappa = \frac{2Var(\ddot{U}_\delta)}{E[\ddot{U}_\delta]}$ and $\nu = \frac{2E[\ddot{U}_\delta]^2}{Var(\ddot{U}_\delta)}$.

As a comparison, we have also conducted a restricted likelihood ratio test (RLRT) analysis [8], which has the same asymptotic distribution as that of a score test.

β	ϕ	$PVE_\delta(\%)$	$PVE_\gamma(\%)$	$\ddot{U}_\delta$	RLRT	U_ψ
NO	NO	0	0	0.055	0.052	0.051
NO	NO	0	5	0.046	0.04	0.179
NO	NO	0	8	0.044	0.031	0.43
NO	NO	5	0	0.251	0.233	0.154
NO	NO	5	5	0.258	0.238	0.34
NO	NO	5	8	0.262	0.226	0.557
NO	NO	8	0	0.549	0.53	0.421
NO	NO	8	5	0.547	0.539	0.571
NO	NO	8	8	0.528	0.516	0.725
YES	YES	0	0	0.054	0.04	0.462
YES	YES	0	5	0.05	0.042	0.575
YES	YES	0	8	0.064	0.047	0.74
YES	YES	5	0	0.262	0.252	0.559
YES	YES	5	5	0.265	0.242	0.693
YES	YES	5	8	0.252	0.223	0.81
YES	YES	8	0	0.55	0.529	0.71
YES	YES	8	5	0.576	0.565	0.821
YES	YES	8	8	0.564	0.537	0.883

Table 4: Table comparing the statistical power of the joint score test statistic, U_ψ, RLRT and $\ddot{U}_\delta$. This data were generated from 1,000 simulations run on 100 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3).

5 Analysis of Gibbs *et al* adult normal human brain data

5.1 Data description

Fresh frozen tissue samples of the cerebellum (CRBLM), frontal cortex (FCTX), caudal pons (PONS) and temporal cortex (TCTX) were obtained from 150 neuropathologically normal samples [7]. Genotyping was performed using Infinium HumanHap550 beadchips (Illumina) to assay genotypes for 561,466 SNPs, from the cerebellum tissue samples. CpG methylation status was determined using HumanMethylation27 BeadChips (Illumina), which measure methylation at 27,578 CpG dinucleotides at 14,495 genes. Profiling of 22,184 mRNA transcripts was performed using HumanRef-8 Expression BeadChips (Illumina) The datasets are publicly available (GEO Accession Number: **GSE15745**; dbGAP Study Accession: **phs000249.v1.p1**).

Accession ID	Repository	Data type	Platform	Number of probes
GSE15745	GEO	Gene expression data	Illumina humanRef-8 v2.0 expression beadchip	22,184
GSE15745	GEO	Methylation data	Illumina HumanMethylation27 BeadChip	27,578
phs000249.v1.p1	dbGaP	Genotype data	HumanHap550v3.0 7	561,466

Table 5: *A description of brain data*

5.2 Data analysis design

We performed data analyses that focused on *cis* candidate regions.

- The proximity of an eQTL to the transcription start site of a gene does not exceed 100 kilobase up- and down-stream of the transcription start site of a gene (*cis*-SNP).
- We picked CpG islands that are less than 1.5 kilobase up- and down-stream of the transcription start site of the same gene.

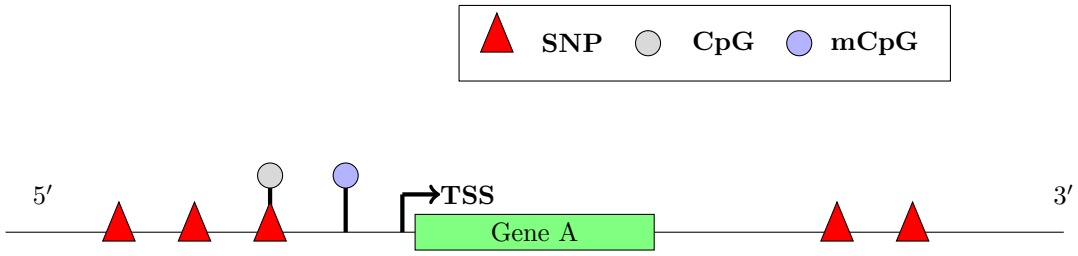


Figure 5: An illustration of the analysis design. The red triangles indicate SNPs and the circles, CpG sites (gray = unmethylated; blue = methylated). CpG sites that are at least 1.5 Kb from the transcription start site (TSS) of a gene were picked for the analysis. All the SNPs that were picked did not exceed 100 kilobase up- and down-stream of the transcription start site of a gene (*cis*-SNPs).

Each mRNA-CpG pair will be tested for a strong association with every *cis*-SNP thus identifying significantly associated mRNA-CpG-SNP triplets.

5.3 Data Preprocessing

5.3.1 Gene Expression data

Gene expression on four brain regions are publicly available (Gene Expression Omnibus (GEO) Accession Number: GSE15745) as non-normalized data, which were individually quantile normalized. All the genes expression probes on sex chromosomes X and Y were removed from the analysis. Each gene expression probe was then adjusted for the biological and methodological covariates such as tissue bank, gender, hybridization batch and numeric covariates such as post-mortem interval (PMI) and age in order to remove any associated confounding effects using the following linear model –

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + PC_1 + \ldots + PC_{10} + \epsilon$$

where Y is the gene expression data, $X_1 \ldots X_n$ represent the aforementioned biological and methodological covariates while $PC_1 \ldots PC_{10}$ are the first 10 principal components obtained from the original gene expression data. Following a previous analyses [5], we removed the global variation in expression among tissues by using the residual expression for each probe in each tissue after removing the first 10 principal components along with the covariates. These residuals then were used in the downstream analyses.

5.3.2 Genotype data

The genotype data is recoded into a SNP matrix of values 0, 1 and 2 representing minor allele counts. Samples with African (GSM394931 in CRBLM, GSM395081 in FCTX, GSM395226 in PONS and GSM395374 in TCTX) and Asian (GSM394121 in CRBLM, GSM394263 in FCTX, GSM394405 in PONS and GSM394566 in TCTX) ancestry were removed from the analysis. These SNPs were filtered on the missing-ness of the individual data and the SNP data (excluded SNPs with missing data), followed by MAF (included SNPs with MAF ≥ 0.05)and

Hardy-Weinberg equilibrium (HWE; p-values < 0.001) in the same order using PLINK [14] software. We ended with 400,097 SNPs after preprocessing.

5.3.3 Methylation data

Methylation data was obtained as a "series matrix file" from GEO. The methylation data consisted of Beta-values, which represent the ratio of methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities) [4]. Therefore, Beta value for an i^{th} interrogated CpG site is –

$$Beta_i = \frac{max(y_{i,methyl}, 0)}{max(y_{i,methyl}, 0) + max(y_{i,unmethyl}, 0) + const}$$

where $y_{i,methyl}$ and $y_{i,unmethyl}$ are the intensities measured by the i^{th} methylation and unmethylated probes, respectively. Beta values range between 0 and 1. An initial assessment of the methylation data included PCA verification of the presence of any potential confounding variables. The first two principal components plotted as a scattered plot show no visible variation between the frontal and temporal cortices as was observed by the authors [7]. The methylation data was later adjusted for the biological and methodological covariates such as tissue bank, gender, hybridization batch and numeric covariates such as post-mortem interval (PMI) and age in order to remove any associated confounding effects using the following linear model –

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

where Y is the methylation expression data and $X_1 \ldots X_n$ represent the aforementioned biological and methodological covariates. The residual methylation expression was later used in the subsequent downstream analyses.

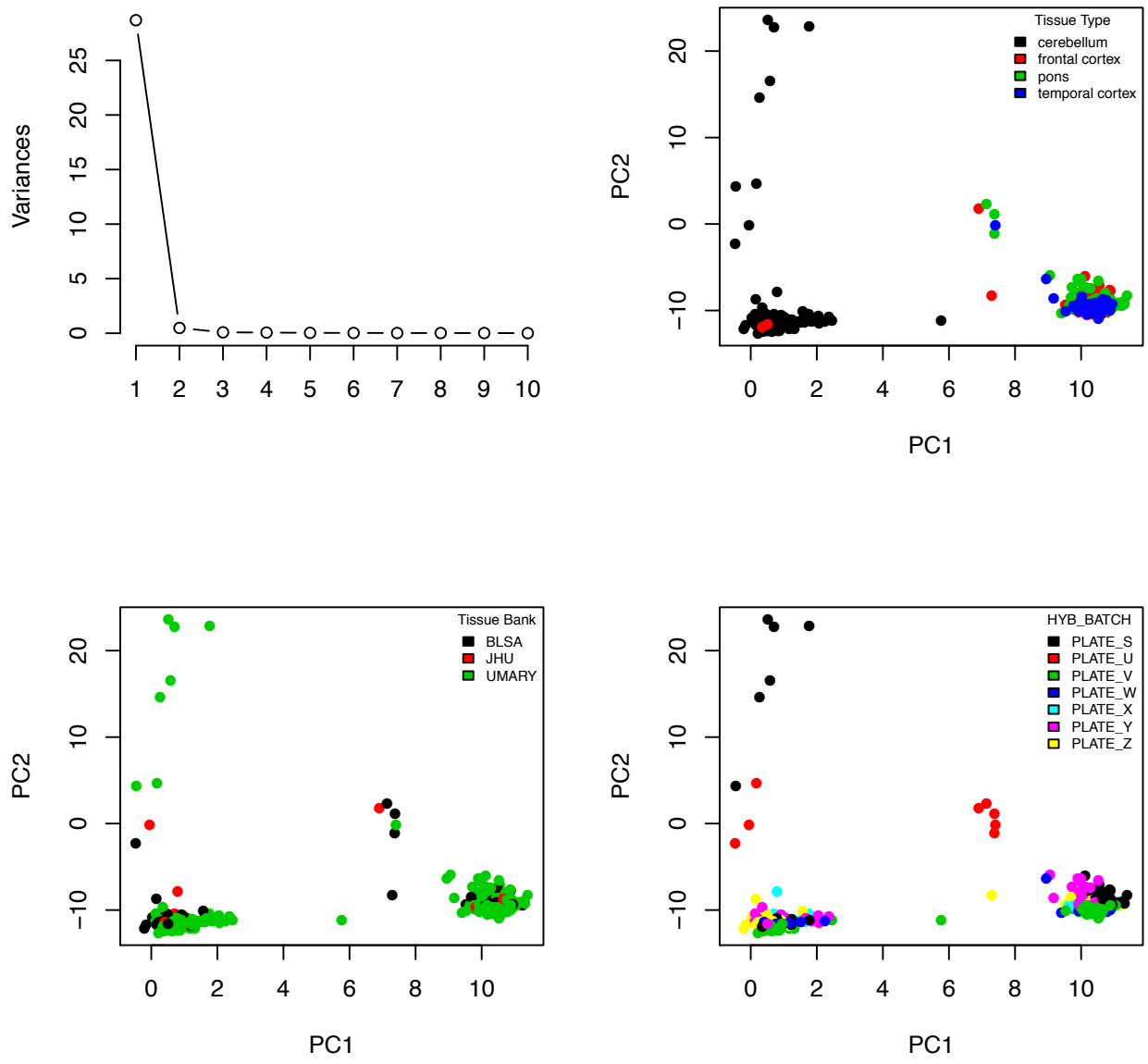


Figure 6: PCA plots exploring the presence of any biological or methodological variation using the first two principal components of the unadjusted methylation data.

5.4 Results

Three different types of analyses were performed –

- A region-by-region eQTL and mQTL analyses
- A region-by-region analysis using a linear model

- Joint analysis using our model

5.4.1 Region-by-Region analysis: Separate eQTL and mQTL analysis

Currently, this is the most commonly used approach in associating CpG sites, SNPs and genes. Gibbs *et al* have first selected every pairing of CpG methylation sites and mRNA transcripts or genes (CpG site is less than 1 Mb from the transcription start site of a gene) and both the CpG site or the mRNA transcript had a significant *cis*-eQTL. mRNA-CpG-SNP triplets were then identified based on the significant correlations between the *cis*-SNP and either mRNA transcript or the CpG methylation site. We have employed the same strategy.

A region-by-region eQTL and mQTL analysis was performed using the following additive model –

$$Y = \beta_0 + \beta_1 G + \epsilon$$

where Y is either gene expression or CpG methylation expression data and G represents genotypes encoded as allele dosage. In order to correct for the number of traits being tested, the p values obtained from the above model were adjusted for multiple hypothesis using an optimized FDR approach [17]. Q values were estimated from each set of p values (originated from each region-by-region analysis) and minimum q value for a given mRNA-SNP or CpG-SNP pair across all the brain regions was computed, which indicates the presence of a statistically significant pair in at least one brain region. The number of significant associations were then assessed at 5% FDR (p value $\leq \frac{0.05}{4}$ where 4 is the number of brain regions).

Using the above process, we identified a total of 6,930 mRNA-SNP pairs and a total of 3,820 CpG-SNP pairs significant in at least one region of the brain, which constituted a total of 1,911 mRNA-CpG-SNP triplets significant in at least one brain region.

	CRBLM	FCTX	PONS	TCTX	Common	Unique
mRNA - *cis*SNP pairs	4,353	4,450	3,134	3,796	4,366	2,564
CpG - *cis*SNP pairs	798	2,322	1,753	2,922	2,309	1,511

Table 6: A region-by-region *cis* analysis of Gibbs *et al* data

The number of mRNA-SNP pairs reported in Table 4 are very close to what was reported by the authors. However, there is some discrepancy in the number of CpG-SNP pairs. This could be due to different preprocessing methods employed by the authors for methylation data.

Triplets	Gene	Chromosome	CRBLM	FCTX	PONS	TCTX	SNP Location	CpG Start Location	CpG End Location
ILMN_1702822 - cg22402007 - rs1187323	NTRK2	9	FALSE	TRUE	TRUE	TRUE	86473236	86472369	86475721
ILMN_1714067 - cg22402007 - rs1187323	NTRK2	9	FALSE	TRUE	TRUE	TRUE	86473236	86472369	86475721
ILMN_1702822 - cg22402007 - rs3758317	NTRK2	9	FALSE	FALSE	TRUE	FALSE	86472435	86472369	86475721
ILMN_1707585 - cg10726357 - rs3802492	TSCOT	9	TRUE	FALSE	FALSE	TRUE	114692657	114692124	114693426
ILMN_1752199 - cg07596401 - rs2459214	LHPP	10	TRUE	TRUE	TRUE	TRUE	126140138	126139604	126141042
ILMN_1810199 - cg12640109 - rs2246920	ASRGL1	11	FALSE	TRUE	FALSE	TRUE	61861513	61860899	61862330
ILMN_1793476 - cg16245261 - rs1488864	PRKCDBP	11	TRUE	FALSE	FALSE	TRUE	6298905	6297418	6298977
ILMN_1793476 - cg05628549 - rs1488864	PRKCDBP	11	FALSE	FALSE	FALSE	TRUE	6298905	6297418	6298977
ILMN_1811692 - cg07844572 - rs2584624	FTSJ3	17	FALSE	TRUE	FALSE	FALSE	59258278	59257457	59259153
ILMN_1735910 - cg08341874 - rs4790706	VMO1	17	FALSE	TRUE	TRUE	TRUE	4636312	4634929	4636534
ILMN_1794490 - cg06851207 - rs8107491	FLJ10781	19	FALSE	TRUE	TRUE	TRUE	51667007	51666325	51667046
ILMN_1716921 - cg17132967 - rs8108275	ZNF83	19	FALSE	FALSE	FALSE	TRUE	57833378	57832897	57834129

Table 7: A list of meQTL as identified by the TBT method. TRUE indicates statistical significance (q value $\leq$ 0.05) whereas FALSE indicates otherwise. Each triplet consists of a mRNA-CpG-SNP combination.

We found 12 SNPs located in the CpG island, called meQTL (Table 5). These meQTL change the methylation status of the CpG site and thus regulate gene expression.

5.4.2 Region-by-Region analysis: Identifying triplets using a linear model

In a region-by-region analysis, we apply the following regression model to identify mRNA-CpG-SNP triplets –

$$Y = G\beta + M\lambda + MG\phi + \epsilon$$

where Y is the gene expression data, G represents genotypes encoded as allele dosage, M represents CpG methylation expression data and MG represents the interaction effect between a SNP and a CpG site. In order to correct for the number of traits being tested, the p values obtained from the above model were adjusted for multiple hypothesis using an optimized FDR approach [17]. Q values were estimated from each set of p values (originated from each region-by-region analysis) and minimum q value for a given mRNA-SNP or CpG-SNP pair across all the brain regions was computed, which indicates the presence of a statistically significant pair in at least one brain region. The number of significant associations were then assessed at 5% FDR (p value $\leq \frac{0.05}{4}$ where 4 is the number of brain regions).

	CRBLM	FCTX	PONS	TCTX	Common	Unique
mRNA - CpG - SNP	1,943	2,675	1,846	3,225	2,790	435

Table 8: Triplets identified using a region-by-region *cis* analysis of Gibbs *et al* data

Using the above process, we identified a total of 3,225 mRNA-CpG-SNP triplets, out of which 19 are meQTL (SNP is located at a CpG site or within a CpG island).

Triplets	Gene	CRBLM	FCTX	PONS	TCTX	Chromosome	SNP Location	CpG Start Location	CpG End Location
ILMN_1808552 - cg02620769 - rs3809147	CCDC65	FALSE	TRUE	FALSE	TRUE	12	47584326	47583729	47584457
ILMN_1790659 - cg13174197 - rs1108842	GNL3	TRUE	TRUE	TRUE	TRUE	3	52695120	52694379	52695574
ILMN_1776077 - cg09503975 - rs606458	SF1	TRUE	TRUE	TRUE	TRUE	11	64302967	64301584	64303340
ILMN_1776077 - cg11295902 - rs606458	SF1	TRUE	TRUE	TRUE	TRUE	11	64302967	64301584	64303340
ILMN_1742808 - cg09503975 - rs606458	SF1	TRUE	TRUE	FALSE	TRUE	11	64302967	64301584	64303340
ILMN_1744835 - cg10413342 - rs629426	MRPL21	TRUE	TRUE	TRUE	TRUE	11	68427680	68427282	68429195
ILMN_1654250 - cg10413342 - rs629426	MRPL21	TRUE	TRUE	TRUE	TRUE	11	68427680	68427282	68429195
ILMN_1811692 - cg11296363 - rs2584624	FTSJ3	TRUE	FALSE	TRUE	TRUE	17	59258278	59257457	59259153
ILMN_1811692 - cg07844572 - rs2584624	FTSJ3	TRUE	FALSE	TRUE	TRUE	17	59258278	59257457	59259153
ILMN_1738223 - cg10865119 - rs3807067	C6ORF208	FALSE	TRUE	FALSE	TRUE	6	169932125	169931539	169933083
ILMN_1694502 - cg24269846 - rs2277339	PRIM1	FALSE	FALSE	FALSE	TRUE	12	55432336	55431390	55433256
ILMN_1694502 - cg19847271 - rs2277339	PRIM1	FALSE	FALSE	FALSE	TRUE	12	55432336	55431390	55433256
ILMN_1737050 - cg15305343 - rs10489769	NSUN4	FALSE	FALSE	FALSE	TRUE	1	46579290	46578232	46579801
ILMN_1814650 - cg22307649 - rs4938619	TRAPPC4	TRUE	TRUE	FALSE	TRUE	11	118395295	118393819	118395567
ILMN_1814650 - cg11054882 - rs4938619	TRAPPC4	TRUE	TRUE	FALSE	TRUE	11	118395295	118393819	118395567
ILMN_1776384 - cg08261177 - rs3760669	ALDH16A1	TRUE	FALSE	TRUE	TRUE	19	54648102	54647579	54648890
ILMN_1722309 - cg04621255 - rs2997922	ENDOG	TRUE	TRUE	FALSE	TRUE	9	130620565	130619231	130621779
ILMN_1722309 - cg00228799 - rs2997922	ENDOG	TRUE	TRUE	TRUE	TRUE	9	130620565	130619231	130621779
ILMN_1771697 - cg24484103 - rs245111	VRK3	FALSE	TRUE	TRUE	TRUE	19	55220709	55219869	55221636

Table 9: meQTL identified using a region-by-region *cis* analysis of Gibbs *et al* data. TRUE indicates statistical significance (q value ≤ 0.05) whereas FALSE indicates otherwise. Each triplet consists of a mRNA-CpG-SNP combination.

5.4.3 Joint analysis

We applied our joint model –

$$Y = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi$$

where Y is the gene expression data, G represents genotypes encoded as allele dosage, M represents CpG methylation expression data and MG represents the interaction effect between a SNP and a CpG site. If T indicates tissue, the $G \times T$ effect is measured by the random effect $v \sim N(0, \gamma)$, the $G \times T \times M$ effect is measured by $w \sim N(0, \delta)$. In order to correct for the number of traits being tested, the p values obtained from the above model were adjusted for multiple hypothesis using an optimized FDR approach [17]. Q values were estimated from each set of p values and we computed the minimum q value for mRNA-CpG-SNP triplets. The number of significant associations were then assessed at 5% FDR (p value ≤ 0.05).

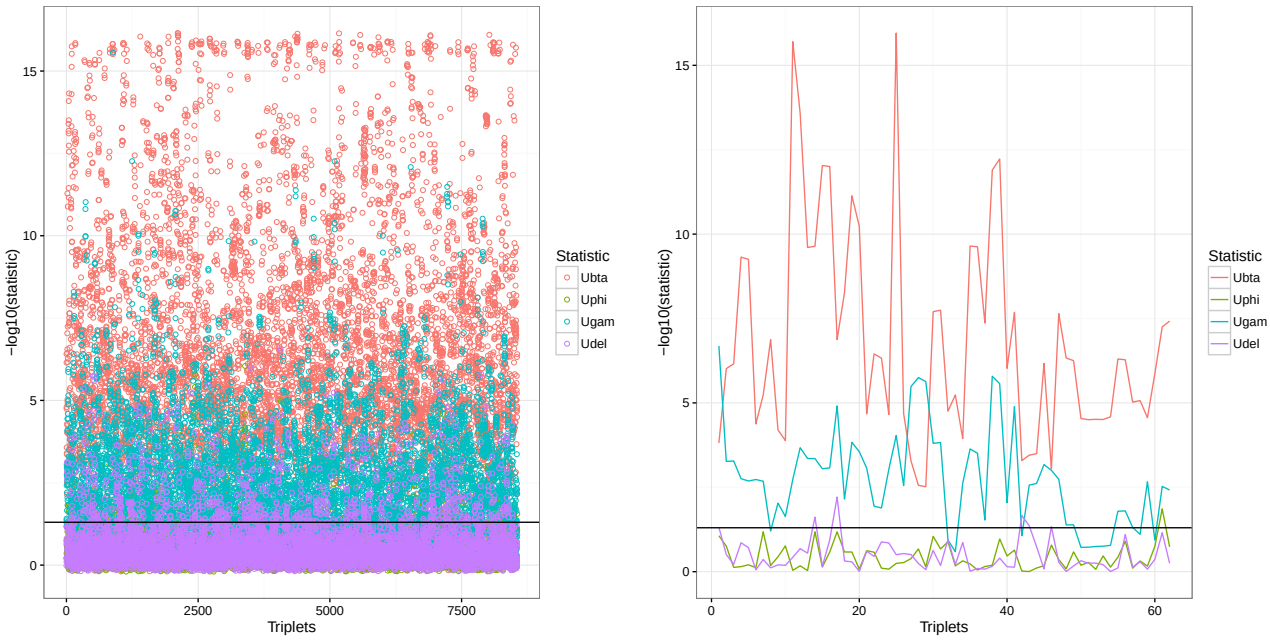
Triplets	Gene	Chromosome	SNP Location	CpG Start Location	CpG End Location
ILMN_1808552 - cg02620769 - rs3809147	CCDC65	12	47584326	47583729	47584457
ILMN_1814871 - cg11187508 - rs12242503	CWF19L1	10	102018457	102017052	102018485
ILMN_1814871 - cg20925178 - rs12242503	CWF19L1	10	102018457	102017052	102018485
ILMN_1814650 - cg11054882 - rs4938619	TRAPPC4	11	118395295	118393819	118395567
ILMN_1814650 - cg22307649 - rs4938619	TRAPPC4	11	118395295	118393819	118395567
ILMN_1794490 - cg06851207 - rs8107491	FLJ10781	19	51667007	51666325	51667046
ILMN_1794490 - cg17296166 - rs8107491	FLJ10781	19	51667007	51666325	51667046
ILMN_1776384 - cg08261177 - rs3760669	ALDH16A1	19	54648102	54647579	54648890
ILMN_1664587 - cg14147105 - rs2303680	FAM125A	19	17392450	17391317	17392695
ILMN_1664587 - cg02217814 - rs2303680	FAM125A	19	17392450	17391317	17392695
ILMN_1744835 - cg10413342 - rs629426	MRPL21	11	68427680	68427282	68429195
ILMN_1654250 - cg10413342 - rs629426	MRPL21	11	68427680	68427282	68429195
ILMN_1742808 - cg11295902 - rs606458	SF1	11	64302967	64301584	64303340
ILMN_1742808 - cg09503975 - rs606458	SF1	11	64302967	64301584	64303340
ILMN_1776077 - cg11295902 - rs606458	SF1	11	64302967	64301584	64303340
ILMN_1776077 - cg09503975 - rs606458	SF1	11	64302967	64301584	64303340
ILMN_1737050 - cg15305343 - rs10489769	NSUN4	1	46579290	46578232	46579801
ILMN_1793598 - cg20869710 - rs2956114	APIP	11	34894389	34893732	34895095
ILMN_1722309 - cg00228799 - rs2997922	ENDOG	9	130620565	130619231	130621779
ILMN_1722309 - cg04621255 - rs2997922	ENDOG	9	130620565	130619231	130621779
ILMN_1718070 - cg02612026 - rs1052571	CASP9	1	15723200	15722905	15724492
ILMN_1718070 - cg02612026 - rs4645989	CASP9	1	15722930	15722905	15724492
ILMN_1718070 - cg21449655 - rs4645989	CASP9	1	15722930	15722905	15724492
ILMN_1718070 - cg21449655 - rs1052571	CASP9	1	15723200	15722905	15724492
ILMN_1790659 - cg13174197 - rs1108842	GNL3	3	52695120	52694379	52695574
ILMN_1806106 - cg13174197 - rs1108842	GNL3	3	52695120	52694379	52695574
ILMN_1799743 - cg15560112 - rs3810276	MYBPC2	19	55626751	55626345	55627632
ILMN_1763207 - cg06150468 - rs2221593	SNFT	1	210940054	210938926	210941332
ILMN_1763207 - cg21421701 - rs2221593	SNFT	1	210940054	210938926	210941332
ILMN_1676631 - cg13352495 - rs231622	CCNU	5	54564347	54562888	54566115
ILMN_1676631 - cg08514736 - rs231622	CCNU	5	54564347	54562888	54566115
ILMN_1759184 - cg14093125 - rs2247289	C19ORF48	19	55999904	55998833	56000331
ILMN_1696852 - cg19282452 - rs7256487	FLJ46230	19	19716654	19716515	19716869
ILMN_1721723 - cg03192551 - rs8079946	SLC25A39	17	39757234	39756526	39760418
ILMN_1744068 - cg05026033 - rs2305451	ELP3	8	28006589	28006001	28006978
ILMN_1744068 - cg18249244 - rs2305451	ELP3	8	28006589	28006001	28006978
ILMN_1771697 - cg24484103 - rs245111	VRK3	19	55220709	55219869	55221636
ILMN_1811692 - cg07844572 - rs2584624	FTSJ3	17	59258278	59257457	59259153
ILMN_1811692 - cg11296363 - rs2584624	FTSJ3	17	59258278	59257457	59259153
ILMN_1810199 - cg12640109 - rs2246920	ASRGL1	11	61861513	61860899	61862330
ILMN_1738223 - cg10865119 - rs3807067	C6ORF208	6	169932125	169931539	169933083
ILMN_1673518 - cg22982528 - rs4624474	BRWD1	21	39608197	39606107	39608381

Continued...

Triplets	Gene	Chromosome	SNP Location	CpG Start Location	CpG End Location
ILMN_1674498 - cg01758870 - rs12700457	C7ORF46	7	23687245	23685962	23687263
ILMN_1674498 - cg08707078 - rs12700457	C7ORF46	7	23687245	23685962	23687263
ILMN_1713605 - cg18619831 - rs1200353	RPAP1	15	39623526	39623254	39624429
ILMN_1790461 - cg11406695 - rs3818532	C6ORF125	6	33787785	33786897	33787849
ILMN_1694502 - cg24269846 - rs2277339	PRIM1	12	55432336	55431390	55433256
ILMN_1756326 - cg05465755 - rs3211663	CKS2	9	91115908	91115007	91116607
ILMN_1756326 - cg03712708 - rs3211663	CKS2	9	91115908	91115007	91116607
ILMN_1730818 - cg26005082 - rs11878617	C19ORF30	19	4720740	4720438	4720853
ILMN_1730818 - cg02479575 - rs11878617	C19ORF30	19	4720740	4720438	4720853
ILMN_1730818 - cg03996793 - rs11878617	C19ORF30	19	4720740	4720438	4720853
ILMN_1730818 - cg04640886 - rs11878617	C19ORF30	19	4720740	4720438	4720853
ILMN_1730818 - cg01400401 - rs11878617	C19ORF30	19	4720740	4720438	4720853
ILMN_1680348 - cg26885858 - rs1976403	NBPF3	1	21639040	21638732	21639906
ILMN_1680348 - cg22190291 - rs1976403	NBPF3	1	21639040	21638732	21639906
ILMN_1804735 - cg09622447 - rs1788484	CBS	21	43370010	43367042	43370279
ILMN_1804735 - cg22633722 - rs1788484	CBS	21	43370010	43367042	43370279
ILMN_1726809 - cg03046445 - rs3809140	BHLHB3	12	26169711	26168959	26171095
ILMN_1787415 - cg06027949 - rs12542216	SNX16	8	82916578	82915812	82917588
ILMN_1793724 - cg17746675 - rs440746	C3ORF31	3	11863556	11862686	11863678
ILMN_1793724 - cg03169527 - rs440746	C3ORF31	3	11863556	11862686	11863678

Table 10: A list of 62 meQTL as identified by our joint score test approach. Triplets of mRNA-CpG-SNP in bold were identified by a region-by-region analysis. A total of 18 (out of 19) triplets were picked up by our method.

When applied our joint model on Gibbs *et al* data, we identified 8,555 statistically significant mRNA - CpG - SNP triplets with a 78% overlap with the number of triplets identified by the region-by-region analysis. A total of 62 meQTL are identified using our method. These include 18 out of 19 meQTL (95%) identified by a region-by-region analysis.



(a) 8,555 statistically significant triplets

(b) 62 meQTL

Figure 7: Triplets with expanded statistics. All the values are -log10 unadjusted p values of individual statistics that comprise our joint score test statistic.

≤ 0.05	U_β	U_ϕ	U_γ	U_δ
FALSE	565 (6.6%)	7918 (92.5%)	2064 (24.1%)	7522 (87.9%)
TRUE	7990 (93.4%)	637 (7.4%)	6491 (75.9%)	1033 (1.2%)

Table 11: The distribution of the different types of effects as measured by our joint score test statistic. U_β is a measure of the main additive genetic effect. U_ϕ is a measure of $G \times M$ effect. U_γ measures $G \times T$ effect while U_δ is a measure of $G \times M \times T$.

Majority of these significant associations are driven by the additive genetic effect and the tissue-specific interaction with the genotype (Table 9).

6 Mathematical derivations

$$\mathbf{Y} = J\alpha + G\beta + Au + Bv + Cw + Dx + \xi \tag{17}$$

where –

- $\mathbf{Y}$ is $nt \times 1$ vector, and $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{it})^T$
- $\boldsymbol{\xi} = \left(\xi_i^T, \cdots, \xi_n^T\right)^T$ is a $nt \times 1$ vector

- $\boldsymbol{J} = (I_t, \cdots, I_t)^T$ is a $nt \times t$ matrix
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_t)^T$ is a vector of length t
- $\mathbf{G}$ is $nt \times 1$ vector and $G_i = (G_{i1}, G_{i2}, \cdots, G_{it})^T$
- $\mathbf{A}$ is $nt \times n$ model matrix with each column represented by a combination of 1_t^T and **zeros** indicating an individual's intercept.
- $\mathbf{u}$ is a vector of individual-specific random effects of length n.
- $\mathbf{B}$= $(g_i I_t, \cdots, g_n I_t)^T$ is a $nt \times t$ matrix
- $\mathbf{v}$= $(v_1, v_2, \cdots, v_t)^T$ is a vector of length t and $v \sim N_t(0, \gamma)$
- $\mathbf{C}$= $(g_i M_t, \cdots, g_n M_t)^T$ is a $nt \times t$ matrix
- $\mathbf{w}$= $(w_1, w_2, \cdots, w_t)^T$ is a vector of length t and $w_t \sim N_t(0, \delta)$
- $\mathbf{D}$= $(M_t, \cdots, M_t)^T$ is a $nt \times t$ matrix
- $\mathbf{x}$= $(x_1, x_2, \cdots, x_t)^T$ is a vector of length t and $x_t \sim N_t(0, \theta)$

Parameters of interest –

$$\Theta = \{\gamma, \delta\} \quad and \quad \beta \quad and \quad \phi$$

We define global null as –

$$\boxed{H_0 : \beta = \phi = \Theta = 0; \qquad H_A : \beta \neq 0; \phi \neq 0; \quad \Theta > 0}$$

From equation 1, the likelihood function of Y conditioned on the genotype can be written as –

$$L_n(\Theta; Y) = \prod_{i=1}^{n} (2\pi)^{-\frac{n}{2}} \mid \Sigma_i \mid^{-\frac{1}{2}} exp\left[-\frac{1}{2}(Y - J\alpha - G\beta - M\lambda - MG\phi)^T \Sigma_i^{-1} (Y - J\alpha - G\beta - M\lambda - MG\phi)\right] \tag{18}$$

The marginal log-likelihood function derived from the above equation is –

$$\ell_n(\Theta; Y) = -c - \frac{1}{2}\log \mid \Sigma \mid -\frac{1}{2}(Y - J\alpha - G\beta - M\lambda - MG\phi)^T \Sigma^{-1} (Y - J\alpha - G\beta - M\lambda - MG\phi) \tag{19}$$

In other words, the marginal distribution of the above model can be represented as –

$$Y \sim N(J\alpha + G\beta + M\lambda + MG\phi, \Sigma)$$

where Σ is the total variance for an individual. Variance for a single individual looks like a $t \times t$ matrix of ϵ, τ, γ, δ, ϕ and θ where t is the number of tissues. In a matrix format, for all individuals the total variance can be represented as –

$$\boxed{\Sigma_n = \epsilon I_n + \tau AA^T + \gamma BB^T + \delta CC^T + \theta DD^T}$$

While AA^T forms a block diagonal structure whereas BB^T, CC^T and DD^T do not. Model matrix B is a function of genotype while C is a function of both genotype and methylation data. D is a function of just the methylation data.

6.1 Score function

$$\frac{\partial \ell_i}{\partial \alpha} = \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i - \phi_j m_i g_i \right)$$

$$\frac{\partial \ell_i}{\partial \beta} = \mathbf{1}_t g_i \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i - \phi_j m_i g_i \right)$$

$$\frac{\partial \ell_i}{\partial \lambda} = \mathbf{1}_t m_i \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i - \phi_j m_i g_i \right)$$

$$\frac{\partial \ell_i}{\partial \phi} = m_i g_i \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i - \phi_j m_i g_i \right)$$

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \tau} &= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \tau} \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \tau} \right) \right\} \\
&= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} A_i A_i^T \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} A_i A_i^T \right) \right\}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \gamma} &= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \gamma} \Sigma_i^{-1} \left(Y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \gamma} \right) \right\} \\
&= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1} \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} B_i B_i^T \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} B_i B_i^T \right) \right\}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \delta} &= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \delta} \Sigma_i^{-1} \left(Y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \delta} \right) \right\} \\
&= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1} \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} C_i C_i^T \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} C_i C_i^T \right) \right\}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \theta} &= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \theta} \Sigma_i^{-1} \left(Y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \theta} \right) \right\} \\
&= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1} \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} E_i E_i^T \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} E_i E_i^T \right) \right\}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \epsilon} &= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \epsilon} \Sigma_i^{-1} \left(y_{i,j} - \alpha_{j,t} - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma^{-1} \frac{\partial V_i}{\partial \epsilon} \right) \right\} \\
&= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-1} I \Sigma_i^{-1} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} I \right) \right\} \\
&= \frac{1}{2} \left\{ \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right)^T \Sigma_i^{-2} \left(y_{i,j} - \alpha_j - \mathbf{1}_t \beta_j g_i - \lambda_j m_i \right) - \mathrm{Tr} \left(\Sigma_i^{-1} \right) \right\}
\end{aligned}$$

6.2 Information matrix

$$
\begin{aligned}
I_{\beta\alpha} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\beta\partial\alpha}\right] \\
&= \Sigma_i^{-1} E[g_i] \\
&= \Sigma_i^{-1} \sum_{i=1}^{n} g_i
\end{aligned}
$$

$$
\begin{aligned}
I_{\beta\lambda} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\beta\partial\lambda}\right] \\
&= \Sigma_i^{-1} \sum_{i=1}^{n} m_i g_i
\end{aligned}
$$

$$
\begin{aligned}
I_{\alpha\alpha} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\alpha\partial\alpha^T}\right] \\
&= \Sigma_i^{-1}
\end{aligned}
$$

$$
\begin{aligned}
I_{\alpha\lambda} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\alpha\partial\lambda}\right] \\
&= \Sigma_i^{-1} \sum_{i=1}^{n} m_i
\end{aligned}
$$

$$
\begin{aligned}
I_{\lambda\lambda} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\lambda\partial\lambda^T}\right] \\
&= \Sigma_i^{-1} \sum_{i=1}^{n} g_i m_i^2
\end{aligned}
$$

$$
\begin{aligned}
I_{\gamma\tau} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\gamma\partial\tau}\right] \\
&= -\left(-\frac{1}{2}\operatorname{Tr}\left(V^{-1}\frac{\partial\Sigma_i}{\partial\gamma}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\tau}\right)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}B_i B_i^T \Sigma_i^{-1} A_i A_i^T\right)
\end{aligned}
$$

$$
\begin{aligned}
I_{\gamma\epsilon} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\gamma\partial\epsilon}\right] \\
&= -\left(-\frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\gamma}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\epsilon}\right)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}B_i B_i^T \Sigma_i^{-1} I\right)
\end{aligned}
\tag{20}
$$

$$
\begin{aligned}
I_{\delta\tau} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\delta\partial\tau}\right] \\
&= -\left(-\frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}\frac{\partial\Sigma_i}{\partial\delta}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\tau}\right)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}C_iC_i^T\Sigma_i^{-1}A_iA_i^T\right)
\end{aligned}
$$

$$
\begin{aligned}
I_{\delta\epsilon} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\gamma\partial\epsilon}\right] \\
&= -\left(-\frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\delta}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\epsilon}\right)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}C_iC_i^T\Sigma_i^{-1}I\right)
\end{aligned}
\tag{21}
$$

$$
\begin{aligned}
I_{\phi\tau} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\phi\partial\tau}\right] \\
&= -\left(-\frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}\frac{\partial\Sigma_i}{\partial\gamma}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\tau}\right)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}D_iD_i^T\Sigma_i^{-1}A_iA_i^T\right)
\end{aligned}
$$

$$
\begin{aligned}
I_{\phi\epsilon} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\gamma\partial\epsilon}\right] \\
&= -\left(-\frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\gamma}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\epsilon}\right)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}D_iD_i^T\Sigma_i^{-1}I\right)
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
I_{\tau\tau} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\tau\partial\tau^T}\right] \\
&= -\left(-\frac{1}{2}\operatorname{Tr}\left(V_i^{-1}\frac{\partial\Sigma_i}{\partial\tau}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\tau}\right)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}A_iA_i^T\Sigma_i^{-1}A_iA_i^T\right)
\end{aligned}
\tag{23}
$$

$$
\begin{aligned}
I_{\tau\epsilon} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\tau\partial\epsilon}\right] \\
&= -\left(-\frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\tau}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\epsilon}\right)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma_i^{-1}A_iA_i^T\Sigma_i^{-1}\right)
\end{aligned}
\tag{24}
$$

$$
\begin{aligned}
I_{\epsilon\epsilon} &= -E_{Y|G=g}\left[\frac{\partial^2 \ell_i}{\partial\epsilon\partial\epsilon}\right] \\
&= -\left(-\frac{1}{2}\,\mathrm{Tr}\left(\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\epsilon}\Sigma_i^{-1}\frac{\partial\Sigma_i}{\partial\epsilon}\right)\right) \\
&= \frac{1}{2}\,\mathrm{Tr}\left(\Sigma_i^{-1} I \Sigma_i^{-1} I\right) \\
&= \frac{1}{2}\,\mathrm{Tr}\left(\Sigma_i^{-2}\right)
\end{aligned}
\tag{25}
$$

6.3 Joint score test

Let the parameters of interest be $\psi = (\beta, \phi, \gamma, \delta)^T$ and the nuisance parameters be $\eta = (\alpha, \lambda, \tau, \theta, \epsilon)^T$. The following is constructed under the null (H_0) –

$$
U_\psi = \begin{bmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \phi} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \delta} \end{bmatrix} - \begin{bmatrix} I_{\beta\alpha} & I_{\beta\lambda} & I_{\beta\tau} & I_{\beta\theta} & I_{\beta\epsilon} \\ I_{\phi\alpha} & I_{\phi\lambda} & I_{\phi\tau} & I_{\phi\theta} & I_{\phi\epsilon} \\ I_{\gamma\alpha} & I_{\gamma\lambda} & I_{\gamma\tau} & I_{\gamma\theta} & I_{\gamma\epsilon} \\ I_{\delta\alpha} & I_{\delta\lambda} & I_{\delta\tau} & I_{\delta\theta} & I_{\delta\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\lambda} & I_{\alpha\tau} & I_{\alpha\theta} & I_{\alpha\epsilon} \\ I_{\lambda\alpha} & I_{\lambda\lambda} & I_{\lambda\tau} & I_{\lambda\theta} & I_{\lambda\epsilon} \\ I_{\tau\alpha} & I_{\tau\lambda} & I_{\tau\tau} & I_{\tau\theta} & I_{\tau\epsilon} \\ I_{\theta\alpha} & I_{\theta\lambda} & I_{\theta\tau} & I_{\theta\theta} & I_{\theta\epsilon} \\ I_{\epsilon\alpha} & I_{\epsilon\lambda} & I_{\epsilon\tau} & I_{\epsilon\theta} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \lambda} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \theta} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}
$$

$$
U_\psi = \begin{bmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \phi} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \delta} \end{bmatrix} - \begin{bmatrix} I_{\beta\alpha} & I_{\beta\lambda} & 0 & 0 & 0 \\ I_{\phi\alpha} & I_{\phi\lambda} & 0 & 0 & 0 \\ 0 & 0 & I_{\gamma\tau} & I_{\gamma\theta} & I_{\gamma\epsilon} \\ 0 & 0 & I_{\delta\tau} & I_{\delta\theta} & I_{\delta\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\lambda} & 0 & 0 & 0 \\ I_{\lambda\alpha} & I_{\lambda\lambda} & 0 & 0 & 0 \\ 0 & 0 & I_{\tau\tau} & I_{\tau\theta} & I_{\tau\epsilon} \\ 0 & 0 & I_{\theta\tau} & I_{\theta\theta} & I_{\theta\epsilon} \\ 0 & 0 & I_{\epsilon\tau} & I_{\epsilon\theta} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \lambda} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \theta} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}
$$

Under the null, we can show that –

$$
\begin{aligned}
U_\beta &= \left(G - \bar{G}\right)^T \Sigma_n^{-1} \hat{Y} \\
U_\phi &= \left(MG - \overline{MG}\right)^T \Sigma_n^{-1} \hat{Y} \\
U_\gamma &= \frac{1}{2}\left\{\hat{Y}^T \Sigma_n^{-1} B_i B_i^T \Sigma_n^{-1} \hat{Y} - Tr\left(\Sigma_n^{-1} B_i B_i^T\right)\right\} \\
U_\delta &= \frac{1}{2}\left\{\hat{Y}^T \Sigma_n^{-1} C_i C_i^T \Sigma_n^{-1} \hat{Y} - Tr\left(\Sigma_n^{-1} C_i C_i^T\right)\right\}
\end{aligned}
$$

6.4 Variance component score test for the $G \times M \times T$ effect

Let the parameters of interest be $\psi = (\delta)^T$ and the nuisance parameters be $\eta = (\alpha, \beta, \phi, \gamma, \lambda, \tau, \theta, \epsilon)^T$. The following is constructed under the null (H_0) –

$$\ddot{U}_\delta = \left[\frac{\partial l}{\partial \delta}\right] - \begin{bmatrix} I_{\delta\alpha} & I_{\delta\beta} & I_{\delta\phi} & I_{\delta\lambda} & I_{\delta\tau} & I_{\delta\gamma} & I_{\delta\theta} & I_{\delta\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\beta} & I_{\alpha\phi} & I_{\alpha\lambda} & I_{\alpha\tau} & I_{\alpha\gamma} & I_{\alpha\theta} & I_{\alpha\epsilon} \\ I_{\beta\alpha} & I_{\beta\beta} & I_{\beta\phi} & I_{\beta\lambda} & I_{\beta\tau} & I_{\beta\gamma} & I_{\beta\theta} & I_{\beta\epsilon} \\ I_{\phi\alpha} & I_{\phi\beta} & I_{\phi\phi} & I_{\phi\lambda} & I_{\phi\tau} & I_{\phi\gamma} & I_{\phi\theta} & I_{\phi\epsilon} \\ I_{\lambda\alpha} & I_{\lambda\beta} & I_{\lambda\phi} & I_{\lambda\lambda} & I_{\lambda\tau} & I_{\lambda\gamma} & I_{\lambda\theta} & I_{\lambda\epsilon} \\ I_{\tau\alpha} & I_{\tau\beta} & I_{\tau\phi} & I_{\tau\lambda} & I_{\tau\tau} & I_{\tau\gamma} & I_{\tau\theta} & I_{\tau\epsilon} \\ I_{\gamma\alpha} & I_{\gamma\beta} & I_{\gamma\phi} & I_{\gamma\lambda} & I_{\gamma\tau} & I_{\gamma\gamma} & I_{\gamma\theta} & I_{\gamma\epsilon} \\ I_{\theta\alpha} & I_{\theta\beta} & I_{\theta\phi} & I_{\theta\lambda} & I_{\theta\tau} & I_{\theta\gamma} & I_{\theta\theta} & I_{\theta\epsilon} \\ I_{\epsilon\alpha} & I_{\epsilon\beta} & I_{\epsilon\phi} & I_{\epsilon\lambda} & I_{\epsilon\tau} & I_{\epsilon\gamma} & I_{\epsilon\theta} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \phi} \\ \frac{\partial l}{\partial \lambda} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \theta} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}$$

$$\ddot{U}_\delta = \left[\frac{\partial l}{\partial \delta}\right] - \begin{bmatrix} 0 & 0 & 0 & 0 & I_{\delta\tau} & I_{\delta\gamma} & I_{\delta\theta} & I_{\delta\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\beta} & I_{\alpha\phi} & I_{\alpha\lambda} & 0 & 0 & 0 & 0 \\ I_{\beta\alpha} & I_{\beta\beta} & I_{\beta\phi} & I_{\beta\lambda} & 0 & 0 & 0 & 0 \\ I_{\phi\alpha} & I_{\phi\beta} & I_{\phi\phi} & I_{\phi\lambda} & 0 & 0 & 0 & 0 \\ I_{\lambda\alpha} & I_{\lambda\beta} & I_{\lambda\phi} & I_{\lambda\lambda} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{\tau\tau} & I_{\tau\gamma} & I_{\tau\theta} & I_{\tau\epsilon} \\ 0 & 0 & 0 & 0 & I_{\gamma\tau} & I_{\gamma\gamma} & I_{\gamma\theta} & I_{\gamma\epsilon} \\ 0 & 0 & 0 & 0 & I_{\theta\tau} & I_{\theta\gamma} & I_{\theta\theta} & I_{\theta\epsilon} \\ 0 & 0 & 0 & 0 & I_{\epsilon\tau} & I_{\epsilon\gamma} & I_{\epsilon\theta} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \phi} \\ \frac{\partial l}{\partial \lambda} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \theta} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}$$

From above, we can show that –

$$\ddot{U}_\delta = \frac{1}{2}\left\{\hat{Y}^T \Sigma_n^{-1} C_i C_i^T \Sigma_n^{-1} \hat{Y} - Tr\left(\Sigma_n^{-1} C_i C_i^T\right)\right\}$$

where $\hat{\Sigma} = \hat{\tau} AA^T + \hat{\gamma} BB^T + \hat{\theta} DD^T$ and $\hat{Y} = \left(Y - J\hat{\alpha} - G\hat{\beta} - M\hat{\lambda} - MG\hat{\phi}\right)$

6.5 Optimal weights to minimize the variance of U_ψ

Let $a = (a_\beta, a_\phi, a_\gamma, a_\delta)^T$, $U_\psi = \left(U_\beta^2, U_\phi^2, U_\gamma, U_\delta\right)$, and $V_\psi = Var\left(U_\psi\right)$. We want to find the minimum variance linear combination $a^T V_\psi$, subject to the constraint that $a_\beta + a_\gamma + a_\delta + a_\phi = 1$ or $a^T \mathbf{1} = 1$. Specifically, we wish to minimize $a^T V_\psi a$ where V_ψ is –

$$V_\psi = \begin{bmatrix} Var(U_\beta^2) & Cov(U_\beta^2, U_\phi^2) & Cov(U_\beta^2, U_\gamma) & Cov(U_\beta^2, U_\delta) \\ Cov(U_\beta^2, U_\phi^2) & Var(U_\phi^2) & Cov(U_\phi^2, U_\gamma) & Cov(U_\phi^2, U_\delta) \\ Cov(U_\beta^2, U_\gamma) & Cov(U_\phi^2, U_\gamma) & Var(U_\gamma) & Cov(U_\gamma, U_\delta) \\ Cov(U_\beta^2, U_\delta) & Cov(U_\phi^2, U_\delta) & Cov(U_\gamma, U_\delta) & Var(U_\delta) \end{bmatrix} = \begin{bmatrix} V_{\beta\beta} & V_{\beta\phi} & V_{\beta\gamma} & V_{\beta\delta} \\ V_{\beta\phi} & V_{\phi\phi} & V_{\phi\gamma} & V_{\phi\delta} \\ V_{\beta\gamma} & V_{\phi\gamma} & V_{\gamma\gamma} & V_{\gamma\delta} \\ V_{\beta\delta} & V_\phi\delta & V_{\gamma\delta} & V_{\delta\delta} \end{bmatrix}$$

Using Lagrangian multipliers to perform constrained optimization, we see that –

$$\mathcal{L}\left(a|\lambda\right) = a^T V_\psi a - \lambda\left(a^T \mathbf{1} - 1\right)$$

where $\mathbf{1} = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T$ and $\lambda > 0$.

$$\frac{\partial}{\partial\left(a^T, \lambda\right)} = \left(a^T V_\psi a - \lambda\left(a^T \mathbf{1} - 1\right)\right) = 0$$

From the above equations, we have the following system of equations–

$$2V_\psi a - \lambda \mathbf{1} = 0 \qquad\qquad a^T\mathbf{1} = \mathbf{1}^T a = 1$$

$$a = \frac{\lambda}{2} V_\psi^{-1}\mathbf{1}$$

and

$$1 = a\mathbf{1}^T = \frac{\lambda}{2}\mathbf{1}^T V_\psi^{-1}\mathbf{1}$$

so that,

$$\lambda = \frac{2}{\mathbf{1}^T V_\psi^{-1}\mathbf{1}}$$

This gives our optimal weights –

$$a = \frac{V_\psi^{-1}\mathbf{1}}{\mathbf{1}^T V_\psi^{-1}\mathbf{1}}$$

References

[1] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. R package version 1.1-7.

[2] J.T. Bell, A.A. Pai, J.K. Pickrell, D.J. Gaffney, R Pique-Regi, J.F. Degner, Y Gilad, and J.K. Pritchard. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biology*, 12(R10), 2011.

[3] A.M Deaton and A Bird. Cpg islands and the regulation of transcription. *Genes and Development*, 2011.

[4] P Du, X Zhang, C Huang, N Jafari, W.A. Kibbe, L Hou, and S.M. Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11, 2010.

[5] J Fu, M.G.M. Wolfs, P Deelen, H Westra, and et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genetics*, 8(1), 2012.

[6] P. K Geyer, M. M Green, and V. G Corces. Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in drosophila. *EMBO J.*, 9(2247-2256), 1990.

[7] J.R. Gibbs, M.P. van der Brug, D.G. Hernandez, B.J. Traynor, M.A. Nalls, S-L Lai, S Arepally, A Dillman, I.P. Rafferty, J Troncoso, R Johnson, H.R. Zielke, L Ferrucci, D.L. Longo, M.R. Cookson, and A.B Singleton. Abundant quantitative trait loci exist for dna methylation and gene expression in human brain. *Plos Genet*, 6(5), 2010.

[8] S Greven, C.M. Crainiceanu, H Kuchenhoff, and A Peters. Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17(4), 2008.

[9] M Gutierrez-Arcelus, H Ongen, T Lappalainen, S.B. Montgomery, A Buil, A Yurovsky, J Bryois, I Padioleau, L Romano, A Planchon, E Falconnet, D Biesler, M Gagnebin, T Giger, C Borel, A Letourneau, P Makrythanasis, M Guipponi, C Gehrig, S.E. Antonarakis, and E.T. Dermitzakis. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genetics*, 2015.

[10] A Hellman and A Chess. Extensive sequence-influenced dna methylation polymorphism in the human genome. *Epigenetics Chromatin*, 24(3):1, 2010.

[11] R Ihaka and R.C. Gentleman. A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[12] M Lemire, S Zaidi, M Ban, and B et al Ge. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nature Communications*, 6(6326), 2014.

[13] C-T Ong and V.G. Corces. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12, 2011.

[14] S Purcell, B Neale, K Todd-Brown, and et al. PLINK: a tool-set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 2007.

[15] F.E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946.

[16] R Shoemaker, J Deng, W Wang, and K Zhang. Allele-specific methylation is prevalent and is contributed by cpg-snps in the human genome. *Genome Res.*, 20:883–889, 2010.

[17] J.D. Storey and R Tibshirani. Statistical significance for genome-wide experiments. *PNAS*, 2003.

[18] T Swift-Scanlan, C.T. Smith, S.A. Bardowell, and C.A. Boettiger. Comprehensive interrogation of cpg island methylation in the gene encoding comt, a key estrogen and catecholamine regulator. *BMC Medical Genomics*, 2014.

[19] J.R Wagner, S Busche, B Ge, T Kwan, T Pastinen, and M Blanchette. The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology*, 15, 2014.

[20] C Wrzodek, F Büchel, G Hinselmann, J Eichner, F Mittag, and A Zell. Linking the epigenome to the genome: Correlation of different features to dna methylation of cpg islands. *Plos ONE*, 7(4), 2012.