Biostatistics (2017), **0**, 0, *pp.* 1–18
doi:10.1093/biostatistics/MT-eQTL˙supp

An Empirical Bayes Approach for Multiple Tissue eQTL Analysis: Supplementary Materials

GEN LI*

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA

ANDREY A. SHABALIN

Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, VA, USA

IVAN RUSYN

Texas Veterinary Medical Center, Texas A&M University, College Station, TX, USA

FRED A. WRIGHT

Department of Statistics and Biological Sciences, North Carolina State University, Raleigh, NC, USA

ANDREW B. NOBEL

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

gl2521@cumc.columbia.edu

*To whom correspondence should be addressed.

1. Model Fitting and Parameter Estimation

1.1 *Matrix eQTL*

The set of correlations $r_{\lambda k}$ for all transcript-SNP pairs λ and tissues $k = 1, \ldots, K$ can be conveniently calculated using the R package Matrix eQTL by Shabalin (2012). The package is designed for fast eQTL analysis in individual tissues. Matrix eQTL accounts for covariates and can filter transcript-SNP pairs by the distance between their genomic locations. Once Matrix eQTL is applied separately for each tissue, the t-statistics it reports can be transformed into correlations using the simple transformation

$$r_{\lambda k} = \frac{t_{\lambda k}}{\sqrt{d_k + t_{\lambda k}^2}}$$

where d_k is the number of degrees of freedom in the tests for tissue k and is also reported by Matrix eQTL. The set of correlations can then be combined in a single matrix with rows $\mathbf{r}_\lambda$.

1.2 *Modified EM Algorithm*

We wish to estimate the parameter $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$ from the observed z-statistics $\{\mathbf{z}_\lambda : \lambda \in \Lambda\}$, which are computed directly from the sample correlations $r_{\lambda k}$ obtained from Matrix eQTL. In order to make the estimation of θ tractable, we assume that the random vectors $\mathbf{Z}_\lambda$ are independent. The likelihood of the model then has a simple product form, depending only on the unknown parameter θ, and the observed z-statistics $\{\mathbf{z}_\lambda\}$:

$$L(\{\mathbf{z}_\lambda\}|\theta) = \prod_{\lambda \in \Lambda} \sum_{\boldsymbol{\gamma} \in \{0,1\}^K} p_{\boldsymbol{\gamma}} \, f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda \,|\, \theta), \tag{1.1}$$

where $f_{\boldsymbol{\gamma}}(\cdot \,|\, \theta)$ is the probability density function of the $\mathcal{N}_K\left(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma}, \Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\right)$ distribution.

Remark: It is important to note that the parameter θ concerns only the (common) marginal distribution of the random vectors $\mathbf{Z}_\lambda$, and is unaffected by their dependence. The assumption that the random vectors $\mathbf{Z}_\lambda$ are independent facilitates estimation of θ, but does not impose any constraints on the marginal dependence structure of $\mathbf{Z}_\lambda$.

We estimate the parameter θ by seeking to maximize the logarithm of the likelihood (1.1). The log-likelihood is not concave, and there appears to be no closed form solution to the maximization problem. Thus one must to rely on iterative algorithms that produce a sequence of parameters $\theta^{(t)}$ converging to a (local) maximum of the likelihood. A direct approach employing a generic software routine for numerical maximization of the likelihood function would be computationally intensive, as each iteration would require multiple (at least 2^K) calculations of the likelihood function around the estimate obtained at the previous iteration. A much faster convergence can be achieved by applying a modification of Expectation Maximization (EM) algorithm. Details are given below.

We treat the unobserved tissue-specificity information vector $\mathbf{\Gamma}_\lambda \in \{0,1\}^K$ as a latent variable. The joint likelihood of both observed and latent variables is:

$$L(\mathbf{z}, \boldsymbol{\gamma} \,|\, \theta) \;=\; p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z} \,|\, \theta).$$

The EM algorithm operates in an iterative fashion. Let $\theta^{(t)} = (\boldsymbol{\mu}_0^{(t)}, \Delta^{(t)}, \Sigma^{(t)}, \mathbf{p}^{(t)})$ be the estimate of the model parameters after t iterations. The estimate $\theta^{(t+1)}$ is defined by

$$\theta^{(t+1)} \;=\; \arg\max_{\theta} Q(\theta : \theta^{(t)}),$$

where

$$Q(\theta : \theta^{(t)}) = \sum_{\lambda} \mathbb{E}_{\mathbf{\Gamma}_\lambda | \mathbf{z}_\lambda, \theta^{(t)}} \big[\log L(\mathbf{z}_\lambda, \mathbf{\Gamma}_\lambda | \theta) \big].$$

The expectation of the log-likelihood is calculated with respect to the conditional distribution of $\mathbf{\Gamma}_\lambda$ given the observed vector of correlations $\mathbf{z}_\lambda$ and the model parameters $\theta^{(t)}$.

Consider the conditional expectation appearing in $Q(\theta : \theta^{(t)})$. Let $p(\boldsymbol{\gamma} \,|\, \theta)$ denote the probability of the configuration $\boldsymbol{\gamma}$ under the probability mass function $\mathbf{p}$ associated with the parameter θ, and define

$$p(\boldsymbol{\gamma} \,|\, \mathbf{z}, \theta) \;=\; \mathbb{P}(\mathbf{\Gamma}_\lambda = \boldsymbol{\gamma} \,|\, \mathbf{z}, \theta) \;=\; \frac{p(\boldsymbol{\gamma} \,|\, \theta) f_{\boldsymbol{\gamma}}(\mathbf{z} \,|\, \theta)}{\sum_{\boldsymbol{\gamma}'} p(\boldsymbol{\gamma}' \,|\, \theta) f_{\boldsymbol{\gamma}'}(\mathbf{z} \,|\, \theta)}$$

The objective function $Q(\theta : \theta^{(t)})$ then has the form

$$Q(\theta : \theta^{(t)}) = \sum_{\lambda} \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \big[\log p(\boldsymbol{\gamma} \,|\, \theta) + \log f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda \,|\, \theta) \big]$$

Maximization of Q with respect to θ leads to the explicit formula

$$p(\boldsymbol{\gamma} \,|\, \theta^{(t+1)}) \;=\; \sum_{\lambda} p(\boldsymbol{\gamma} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \Big/ |\Lambda|$$

where $|\Lambda|$ is the number of gene-SNP pairs under consideration. There appears to be no closed form solution for the iterates of $\boldsymbol{\mu}_0^{(t)}$, $\Sigma^{(t)}$ and $\Delta^{(t)}$. However, in practice, most of the probability mass of $\mathbf{p}$ is concentrated at the two extreme cases $\boldsymbol{\gamma} = \mathbf{0}$ and $\boldsymbol{\gamma} = \mathbf{1}$, reflecting the fact that most transcript-SNP pairs are associated in no tissues or all tissues. Approximating $Q(\cdot)$ by restricting the second sum to $\boldsymbol{\gamma} = 0, 1$ leads to explicit (approximate) estimates of $\boldsymbol{\mu}_0$, Σ and Δ via the following first order conditions:

$$\begin{aligned}
\Delta^{(t+1)} &= \sum_{\lambda} p(\mathbf{0} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \mathbf{z}_\lambda \mathbf{z}_\lambda^T \Big/ \sum_{\lambda} p(\mathbf{0} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \\
\boldsymbol{\mu}_0^{(t+1)} &= \sum_{\lambda} p(\mathbf{1} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \mathbf{z}_\lambda \Big/ \sum_{\lambda} p(\mathbf{1} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) \\
\Sigma^{(t+1)} + \Delta^{(t+1)} &= \sum_{\lambda} p(\mathbf{1} \,|\, \mathbf{z}_\lambda, \theta^{(t)}) (\mathbf{z}_\lambda - \boldsymbol{\mu}_0^{(t+1)})(\mathbf{z}_\lambda - \boldsymbol{\mu}_0^{(t+1)})^T \Big/ \sum_{\lambda} p(\mathbf{1} \,|\, \mathbf{z}_\lambda, \theta^{(t)})
\end{aligned}$$

At some iterations the estimates $\Sigma^{(t+1)}$ may fail to be non-negative definite. In such cases we force $\Sigma^{(t+1)}$ to be non-negative definite by calculating its singular value decomposition and dropping terms with negative coefficients (negative eigenvalues).

Starting with an initial parameter value $\theta^{(0)}$, we perform sequential updates in the manner described above until the change in the likelihood falls below a pre-set threshold. To assess the reliability of the estimate one may run the algorithm multiple times using distinct starting points. In our experiments the algorithm tends to converge to the same estimate regardless of the starting point.

2. Proof of Lemma 2.1

Proof. Let S be a subset of $\{1, \dots, K\}$ with cardinality $|S| = r$. It follows from the defining properties of the multivariate normal distribution that if $\mathbf{U} \sim \mathcal{N}_K(\boldsymbol{\mu}, A)$ then $\mathbf{U}_S \sim \mathcal{N}_r(\boldsymbol{\mu}_S, A_S)$. Therefore we have that

$$\mathbf{Z}_S \;\sim\; \sum_{\boldsymbol{\gamma} \in \{0,1\}^K} p_{\boldsymbol{\gamma}} \, \mathcal{N}_r\big((\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma})_S, \, (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)_S\big) \tag{2.2}$$

Here and in the remainder of the proof we follow the convention that $\boldsymbol{\gamma}$ ranges over $\{0,1\}^K$, and $\boldsymbol{\zeta}$ ranges over $\{0,1\}^r$. Elementary arguments show that

$$(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma})_S \;=\; \boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\gamma}_S \;\text{ and }\; (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)_S \;=\; \Delta_S + \Sigma_S \cdot \boldsymbol{\gamma}_S \boldsymbol{\gamma}_S^T$$

It then follows from (2.2) that

$$\begin{aligned}
\mathbf{Z}_S &\sim \sum_{\boldsymbol{\gamma} \in \{0,1\}^K} p_{\boldsymbol{\gamma}} \, \mathcal{N}_r\big(\boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\gamma}_S, \, \Delta_S + \Sigma_S \cdot \boldsymbol{\gamma}_S \boldsymbol{\gamma}_S^T\big) \\
&= \sum_{\boldsymbol{\zeta} \in \{0,1\}^r} \sum_{\boldsymbol{\gamma} : \boldsymbol{\gamma}_S = \boldsymbol{\zeta}} p_{\boldsymbol{\gamma}} \, \mathcal{N}_r\big(\boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\gamma}_S, \, \Delta_S + \Sigma_S \cdot \boldsymbol{\gamma}_S \boldsymbol{\gamma}_S^T\big) \\
&= \sum_{\boldsymbol{\zeta} \in \{0,1\}^r} \mathcal{N}_r\big(\boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\zeta}, \, \Delta_S + \Sigma_S \cdot \boldsymbol{\zeta}\boldsymbol{\zeta}^T\big) \sum_{\boldsymbol{\gamma} : \boldsymbol{\gamma}_S = \boldsymbol{\zeta}} p_{\boldsymbol{\gamma}} \\
&= \sum_{\boldsymbol{\zeta} \in \{0,1\}^r} p_{\boldsymbol{\zeta},S} \, \mathcal{N}_r\big(\boldsymbol{\mu}_{0,S} \cdot \boldsymbol{\zeta}, \, \Delta_S + \Sigma_S \cdot \boldsymbol{\zeta}\boldsymbol{\zeta}^T\big),
\end{aligned}$$

which is the desired expression for distribution of $\mathbf{Z}_S$. □

3. Proof of Theorem 3.2

3.1 *Continuity and Monotonicity of $F(t)$*

Lemma 3.1 Let U be a bounded, non-negative random variable. For $t \geqslant 0$ define

$$G(t) \;=\; \mathbb{E}[\,U \,|\, U \leqslant t\,] \;=\; \frac{\mathbb{E}[\,U\, \mathbb{I}(U \leqslant t)\,]}{\mathbb{P}(U \leqslant t)}. \tag{3.3}$$

Then the following hold:

1. G is non-decreasing and right continuous;

2. If $\mathbb{P}(U = t) = 0$ then G is continuous at t;

3. If $\mathbb{P}(a < U < b) > 0$ for each $0 < a < b < L$ then G is strictly increasing on $(0, L)$.

Proof. To show that G is non-decreasing it suffices to show that $G(t + \delta) - G(t) \geqslant 0$ for each fixed $t \geqslant 0$ and $\delta > 0$. If $G(t) = 0$ then the result is immediate as the function G is non-negative. If $G(t)$ is positive, then

$$\begin{aligned} G(t+\delta) - G(t) &= \frac{\mathbb{E}[\,U\,\mathbb{I}(U \leqslant t+\delta)\,]}{\mathbb{P}(U \leqslant t+\delta)} \;-\; \frac{\mathbb{E}[\,U\,\mathbb{I}(U \leqslant t)\,]}{\mathbb{P}(U \leqslant t)} \\ &= \frac{\mathbb{E}[\,U\,\mathbb{I}(U \leqslant t+\delta)\,]\,\mathbb{P}(U \leqslant t) \;-\; \mathbb{E}[\,U\,\mathbb{I}(U \leqslant t)\,]\,\mathbb{P}(U \leqslant t+\delta)}{\mathbb{P}(U \leqslant t+\delta)\,\mathbb{P}(U \leqslant t)}. \end{aligned}$$

By elementary arguments the numerator of the last fraction can be expressed as

$$\begin{aligned} &\mathbb{E}[\,U\,\mathbb{I}(t < U \leqslant t+\delta)\,]\,\mathbb{P}(U \leqslant t) \;-\; \mathbb{E}[\,U\,\mathbb{I}(U \leqslant t)\,]\,\mathbb{P}(t < U \leqslant t+\delta) \\ &\geqslant t\,\mathbb{P}(t < U \leqslant t+\delta)\,\mathbb{P}(U \leqslant t) \;-\; t\,\mathbb{P}(U \leqslant t)\,\mathbb{P}(t < U \leqslant t+\delta) \qquad (3.4) \\ &= 0. \end{aligned}$$

Thus G is non-decreasing. Right continuity of G follows by applying the monotone convergence theorem to the numerator and denominator in (3.3). If $\mathbb{P}(U = t) = 0$ then continuity of G at t follows from the dominated convergence theorem in a similar fashion. Finally, if $\mathbb{P}(t < U < t+\delta) > 0$ then the inequality in (3.4) is strict, and the final claim follows by considering $t \in [0, L)$ and $\delta > 0$ such that $t + \delta < L$. □

Lemma 3.2 For $i = 0, \ldots, m$ let f_i be the density of the d-variate normal distribution $\mathcal{N}_d(\mu_i, \Sigma_i)$ and let $c_1, \ldots, c_m$ be positive constants. If at least one of $f_1, \ldots, f_m$ is not equal to f_0, then

$$m_d\Big(\Big\{x : f_0(x) = \textstyle\sum_{j=1}^{m} c_j\, f_j(x)\Big\}\Big) = 0$$

where $m_d(\cdot)$ denotes Lebesgue measure on $\mathbb{R}^d$.

Proof. Define $h(x) = f_0(x) - \sum_{j=1}^m c_j\, f_j(x)$ and let $A = \{x : h(x) = 0\}$. As h is continuous, A is a closed subset of $\mathbb{R}^d$. We establish the result by way of contradiction. Consider first the case in which $d = 1$ and $h(x) = 0$ for each $x \in \mathbb{R}$. By an easy argument, we can assume that the densities f_i, $i = 0, 1, \ldots, m$ are distinct and that $m \geqslant 1$. Let μ_i and σ_i be, respectively, the mean and variance of the distribution specified by the density f_i. Let (σ_j, μ_j) be the largest element, under the usual lexicographic order, of the set $\{(\sigma_i, \mu_i) : 0 \leqslant i \leqslant m\}$. Considering the limit of $h(x)/f_j(x)$ as x tends to infinity, we conclude that $c_j = 0$ if $j \neq 0$ or $1 = 0$ if $j = 0$. In either case we obtain a contradiction, and therefore $h(x)$ cannot be identically equal to zero.

The remainder of the proof proceeds by induction on d. Consider first the case $d = 1$. Note that $h(x)$ is an analytic function of the real variable x. If $m_1(A) > 0$ then there exists $M < \infty$ such that $m_1(A \cap [-M, M]) > 0$. In particular, there are infinitely many points of A in the compact set $[-M, M]$. Thus A has a limit point x_0, and $h(x_0) = 0$ as A is closed. As the zeros of a non-zero analytic function are necessarily isolated, it follows that $h(x)$ is identically zero. This contradicts the argument given above, and we conclude that $m_1(A) = 0$.

Assume now that the lemma holds for dimensions $1, \ldots, d-1$, and consider the general case of dimension d. Suppose that $m_d(A) > 0$. By Fubini's theorem, there exist a Borel measurable set $B \subset \mathbb{R}$ such that (i) $m_1(B) > 0$ and (ii) for every $x_d \in B$ the section

$$A(x_d) = \{x_1^{d-1} : (x_1^{d-1}, x_d) \in A\} \subseteq \mathbb{R}^{d-1}$$

has $(d-1)$-dimensional Lebesgue measure greater than zero. (Here x_1^{d-1} denotes the ordered sequence $x_1, \ldots, x_{d-1}$.) Note that $h(x) = 0$ can be written in the equivalent form

$$0 \;=\; f_0(x_1^{d-1} \,|\, x_d)\, f_0(x_d) \;-\; \sum_{j=1}^m c_j\, f_j(x_1^{d-1} \,|\, x_d)\, f_j(x_d) \qquad x \in A \tag{3.5}$$

where $f_j(x_1^{d-1} \,|\, x_d)$ denotes the conditional density of x_1^{d-1} given x_d under f_j, and $f_j(x_d)$ denotes the marginal density of x_d under f_j. If for each $x_d \in B$ the conditional densities $f_j(x_1^{d-1} \,|\, x_d)$ are equal on $A(x_d)$ then (3.5) becomes

$$0 \;=\; f_0(x_d) \;-\; \sum_{j=1}^{m} c_j\, f_j(x_d) \qquad x_d \in B,$$

which contradicts the induction hypothesis. Suppose then that for some $x_d \in B$ the conditional densities $f_j(x_1^{d-1} \,|\, x_d)$ are not all equal on $A(x_d)$. Then equation (3.5) becomes

$$0 \;=\; f_0(x_1^{d-1} \,|\, x_d) \;-\; \sum_{j=1}^{m} c_j'\, f_j(x_1^{d-1} \,|\, x_d) \qquad x_1^{d-1} \in A(x_d)$$

where $c_j' = c_j\, f_j(x_d)/f_0(x_d)$. Our assumption regarding the conditional densities ensures that $f_j(x_1^{d-1} \,|\, x_d)$ is different from $f_0(x_1^{d-1} \,|\, x_d)$ for some $j \geqslant 1$, again contradicting the induction hypothesis. This completes the proof. □

Lemma 3.3 Let $\eta(\mathbf{z})$ be local false discovery rate defined as

$$\eta(\mathbf{z}) \;:=\; \mathbb{P}(\boldsymbol{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} = \mathbf{z}) \;=\; \frac{p_{\mathbf{0}} f_{\mathbf{0}}(\mathbf{z})}{f(\mathbf{z})}.$$

and assume that every diagonal entry of Σ is positive. Then the following hold.

1. $\inf_{\mathbf{z}\in\mathbb{R}^d} \eta(\mathbf{z}) = 0$.

2. For every $c \geqslant 0$ the Lebesgue measure of the set $\{\mathbf{z} : \eta(\mathbf{z}) = c\}$ in $\mathbb{R}^K$ is zero.

Proof. **Proof of 1:** As $\eta(z)$ is always positive, it is enough to show that there exists $\mathbf{z} \in \mathbb{R}^d$ and $\boldsymbol{\gamma} \in \{0,1\}^K$ such that $f_{\mathbf{0}}(b\mathbf{z})/f_{\boldsymbol{\gamma}}(b\mathbf{z}) \to 0$ as $b \to \infty$. From the exponential form of the multivariate normal densities, it can be seen that the last relation will hold if the matrix $\Delta^{-1} - (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1}$ has an eigenvalue greater than zero.

Let $\mathbf{x}_0$ be an eigenvector of the matrix Δ corresponding to the smallest eigenvalue $\lambda_{\min}(\Delta)$ (which is positive by assumption). Assume without loss of generality that $||\mathbf{x}_0|| = 1$. Using the variational formula for eigenvalues, and the relationship between the eigenvalues of a matrix and those of its inverse, we find that

$$\begin{aligned}
\lambda_{\max}(\Delta^{-1} - (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1}) &= \max_{z:||z||=1} z^T(\Delta^{-1} - (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1})z \\
&\geqslant \max_{z:||z||=1} z^T \Delta^{-1} z \; - \; \max_{z:||z||=1} z^T(\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1} z \\
&= \lambda_{\max}(\Delta^{-1}) \; - \; \lambda_{\max}((\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)^{-1}) \\
&= \lambda_{\min}(\Delta) \; - \; \lambda_{\min}(\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T) \\
&\geqslant \mathbf{x}_0^T \Delta \mathbf{x}_0 \; - \; \mathbf{x}_0^T(\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)\mathbf{x}_0 \\
&= \mathbf{x}_0^T(\Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)\mathbf{x}_0
\end{aligned}$$

Let $1 \leqslant i \leqslant K$ be any index for which $x_{0,i} \neq 0$. If $\boldsymbol{\gamma}$ is the binary K-vector having a 1 in position i and all other entries equal to 0, then it is easy to see that the last expression above is $\sigma_{ii}\, x_{0,i}^2$, which is positive.

Proof of 2: This follows immediately from Lemma 3.2

Proposition 3.4 The function $F(t)$ defined as

$$F(t) \;:=\; \mathbb{E}(\eta(\mathbf{Z}) \,|\, \eta(\mathbf{Z}) \leqslant t) \;=\; \frac{\mathbb{E}[\eta(\mathbf{Z})\, \mathbb{I}(\eta(\mathbf{Z}) \leqslant t)]}{\mathbb{P}(\eta(\mathbf{Z}) \leqslant t)}.$$

is continuous and strictly increasing on the interval $(0, L_\eta)$, where $L_\eta = \sup_{\mathbf{z} \in \mathbb{R}^d} \eta(\mathbf{z}) < 1$.

Proof: Note that $F(t)$ is of the form $g(t)$ in (3.3) with $U = \eta(\mathbf{Z})$. Part 2 of Lemma 3.3 establishes that $\mathbb{P}(\eta(bZ) = t) = 0$, and continuity of F then follows from Lemma 3.1. For $0 < a < b < L_\eta$ we have

$$\mathbb{P}(a < \eta(\mathbf{Z}) < b) \;=\; \mathbb{P}(\eta(\mathbf{Z}) \in (a,b)) \;=\; \mathbb{P}(\mathbf{Z} \in \eta^{-1}(a,b)).$$

As $\eta(\mathbf{z})$ is continuous $\eta^{-1}(a,b)$ is an open subset of $\mathbb{R}^d$. Moreover, $\eta^{-1}(a,b)$ is non-empty by Part 1 of Lemma 3.3. Thus $\mathbb{P}(a < \eta(\mathbf{Z}) < b) > 0$ as the density f of $\mathbf{Z}$ is positive on $\mathbb{R}^d$. Continuity of $F(t)$ then follows from Lemma 3.1. □

3.2 *Proof of Theorem 3.2*

Lemma 3.5 Let $G_1, G_2, \ldots : [0,1] \to \mathbb{R}$ be non-decreasing functions. For fixed $\alpha \in (0, L_\eta)$ define $\theta_n = \sup\{t : G_n(t) \leqslant \alpha\}$ and let $\theta \in (0,1)$ be the unique number such that $F(\theta) = \alpha$. If $G_n(t) \to F(t)$ for each t in a dense subset T of $[0,1]$ then $\theta_n \to \theta$.

Proof. Suppose by way of contradiction that $|\theta_n - \theta| \not\to 0$. Then there exists $\delta_1, \delta_2 > 0$ such that $\{\theta - \delta_1, \theta + \delta_2\} \subseteq T$ and an infinite subsequence n_k of $1, 2, \ldots$ such that either $\theta_{n_k} \leqslant \theta - 2\delta_1$ for each $k \geqslant 1$ or $\theta_{n_k} \geqslant \theta + 2\delta_2$ for each $k \geqslant 1$. In the first case, the definition of θ_n and the monotonicity of G_n imply

$$\alpha \ \leqslant \ G_{n_k}(\theta_{n_k} + \delta_1) \ \leqslant \ G_{n_k}(\theta - \delta_1)$$

Taking limits as $k \to \infty$ we find $\alpha \ \leqslant \ F(\theta - \delta_1) < \alpha$ as F is strictly increasing, which is a contradiction. In the second case, a similar argument shows that

$$\alpha \ \geqslant \ G_{n_k}(\theta_{n_k} - \delta_2) \ \geqslant \ G_{n_k}(\theta + \delta_2).$$

Taking limits as $k \to \infty$ yields $\alpha \ \geqslant \ F(\theta + \delta_2) > \alpha$, which is again a contradiction. This concludes the proof. □

Proof of Theorem 3.2: Let $\hat{\theta}_n = \sup\{t : \hat{F}_n(t) \leqslant \alpha\}$ and let θ be the unique number such that $F(\theta) = \alpha$. We claim that $\hat{\theta}_n \to \theta$ in probability. To show this, assume to the contrary that there exists $\delta > 0$ and a subsequence n_k such that

$$\mathbb{P}\big(|\hat{\theta}_{n_k} - \theta| > \delta\big) > \delta \quad \text{for each} \quad k \geqslant 1. \tag{3.6}$$

Let T be any countable, dense subset of $[0,1]$. Our assumptions imply that $\hat{F}_n(t) \to F(t)$ in probability for each $t \in T$. By a standard diagonalization argument, there exists a subsequence m_k of n_k such that $\hat{F}_{m_k}(t) \to F(t)$ with probability one for each $t \in T$. It then follows from Lemma 3.5 that $\hat{\theta}_{m_k} \to \theta$ with probability one, which contradicts (3.6).

In order to establish the theorem, it will be convenient to work with version of M_n and N_n in which the data-dependent threshold $\hat{\theta}_n$ is replaced by the limiting value θ. Define

$$\tilde{M}_n \;=\; \sum_{\lambda \in \Lambda_n} \mathbb{I}(\mathbf{\Gamma}_\lambda = 0)\, \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leqslant \theta) \quad \text{and} \quad \tilde{N}_n \;=\; \sum_{\lambda \in \Lambda_n} \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leqslant \theta)$$

Note that $\mathbb{E}\tilde{N}_n = |\Lambda_n| \cdot \mathbb{P}(\eta(\mathbf{Z}) \leqslant \theta)$. By an elementary conditioning argument,

$$\begin{aligned} \mathbb{E}\tilde{M}_n &= \sum_{\lambda \in \Lambda_n} \mathbb{E}\Big\{\mathbb{P}(\mathbf{\Gamma}_\lambda = 0 \,|\, \mathbf{Z}_\lambda)\, \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leqslant t_n(\alpha))\Big\} \\ &= \sum_{\lambda \in \Lambda_n} \mathbb{E}\Big\{\eta(\mathbf{Z}_\lambda)\, \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leqslant t_n(\alpha))\Big\} \\ &= |\Lambda_n| \cdot \mathbb{E}[\eta(\mathbf{Z})\, \mathbb{I}(\eta(\mathbf{Z}) \leqslant t)]. \end{aligned}$$

For each $\delta > 0$,

$$\begin{aligned} \mathbb{E}|\tilde{N}_n - N_n| &\leqslant \sum_{\lambda \in \Lambda_n} \mathbb{P}(\eta(\mathbf{Z}_\lambda) \in [\hat{\theta}_n, \theta] \cup [\theta, \hat{\theta}_n]) \\ &\leqslant |\Lambda_n| \, \Big[\mathbb{P}\big(\eta(\mathbf{Z}) \in (\theta - \delta, \theta + \delta)\big) + \mathbb{P}\big(|\hat{\theta}_n - \theta| \geqslant \delta\big)\Big]. \end{aligned}$$

As $\hat{\theta}_n \to \theta$ in probability and the distribution of $\eta(\mathbf{Z})$ has no point masses, the last inequality implies that $\mathbb{E}|\tilde{N}_n - N_n| = |\Lambda_n| \cdot o(1)$. A similar argument shows that $\mathbb{E}|\tilde{M}_n - M_n| = |\Lambda_n| \cdot o(1)$. Thus as n tends to infinity,

$$\begin{aligned} \frac{\mathbb{E}M_n}{\mathbb{E}N_n} &= \frac{\mathbb{E}\tilde{M}_n + |\Lambda_n| \cdot o(1)}{\mathbb{E}\tilde{N}_n + |\Lambda_n| \cdot o(1)} \\ &= \frac{\mathbb{E}[\eta(\mathbf{Z})\, \mathbb{I}(\eta(\mathbf{Z}) \leqslant \theta)] + o(1)}{\mathbb{P}(\eta(\mathbf{Z}) \leqslant \theta) + o(1)} \\ &\to \frac{\mathbb{E}[\eta(\mathbf{Z})\, \mathbb{I}(\eta(\mathbf{Z}) \leqslant \theta)]}{\mathbb{P}(\eta(\mathbf{Z}) \leqslant \theta)} \;=\; F(\theta) \;=\; \alpha. \end{aligned}$$

This completes the proof of the theorem.

4. Simulation Study

In this section, we examine the performance of MT-eQTL through a simulation study with $K = 4$ tissues. As the basis of the model and subsequent inferences is the collection of z-statistic vectors derived from the observed genotype and transcript data, we directly simulate the z-statistics.

4.1 *Simulation Setting*

We simulate 10 million vectors $\mathbf{z}_\lambda$ independently from the MT-eQTL model using parameters $\theta = (\Delta, \Sigma, \mathbf{p}, \boldsymbol{\mu}_0)$ obtained from eQTL analysis of data from the GTEx initiative (we consider the tissues blood, lung, muscle, and thyroid, which we denote by a, b, c, and d, respectively). Sample sizes, sample overlap, and degrees of freedom after covariate correction are given in Table S1. The true model parameters are given in Table S2. Note that the average effect size parameter $\boldsymbol{\mu}_0$ is set to be zero in data generation and model fitting for simplicity.We remark that allowing $\boldsymbol{\mu}_0$ to be free has little effect on the numerical results.

We simulated each vector $\mathbf{z}_\lambda$ in a two-step fashion: first drawing $\boldsymbol{\gamma} \in \{0,1\}^4$ from $\mathbf{p}$, and then drawing $\mathbf{z}_\lambda$ from $f_{\boldsymbol{\gamma}}(\mathbf{z})$ given $\boldsymbol{\gamma}$. Access to the true configurations $\boldsymbol{\gamma}$ enables us to assess false discovery rates associated with inferences from the fitted model.

4.2 *Model Fit*

The approximate EM procedure was used to fit the full 4-dimensional model, as well as all possible 1-, 2-, and 3-dimensional models. We terminated EM updates when the difference between log likelihoods in two consecutive iterations was less than 0.01. The average number of iterations until convergence of the EM procedure was 80. The running time of the EM procedure depended on the number of tissues in the model, ranging from about 1 second per iteration for the 1-dimensional models to about 40 seconds per iteration for full 4-dimensional model on a standard desktop PC. Fitting of the 4-dimensional model based on the simulated data took slightly more than one hour.

As expected, the parameters estimated from the simulated data are very close to those used to generate the data. For the 4-dimensional model, the relative error of each entry of Σ is less than 0.3%, while the relative error for each entry of Δ is less than 0.7%. For the probability mass vector $\mathbf{p}$, thirteen of sixteen entries had relative error less than 1%, with the remaining relative errors equal to 1.45%, 1.66% and 4.31%. These results confirm that the approximate EM procedure works well on the simulated data.

4.3 *Results*

We apply the adaptive thresholding procedure to the local FDRs and detect eQTLs with different configurations. In particular, we attempt to identify eQTLs in at least one tissue, in a single tissue, and in all tissues. The corresponding null configurations are shown in the second column of Table S3. In all studies, we fixed the nominal FDR threshold at $\alpha = 0.05$. Table S3 contains the number of true alternative cases, the number of total discoveries, the number of overlaps (i.e., true positives), the true positive rate (TPR), and the FDR in each study.

In all studies, the observed FDRs are strictly below the nominal level of 5%; and the TPRs are around 40%. These TPRs are considered relatively high because many alternative cases may have modest to small effect sizes in the simulated data, and are not readily distinguishable from the null cases. This behavior is representative of real data where signals are not always highly identifiable. The TPR for testing cross-tissue eQTL is the lowest among all cases because it is the most challenging problem: a cross-tissue eQTL may be easily mistaken as other eQTL configurations (e.g., 1110, 1101, 1011, 0111, etc). Nonetheless, the TPR for such case is still reasonably high, which demonstrates the efficacy of the proposed method. In addition, we emphasize that our method is flexible enough to detect eQTL of different configurations. Numerical results indicate that the method has great power in identifying significant association between a gene and a SNP in a single tissue or in any tissue.

In order to assess how the use of auxiliary tissues increases statistical power of detecting eQTL in a target tissue, we fit a series of nested MT-eQTL models for tissue sets {a}, {a,b}, {a,b,c}, and {a,b,c,d} and only focused on eQTL detection in tissue a. In each study, we fixed the target set $S = \{\text{a}\}$, and applied the adaptive thresholding procedure to the marginal local FDR defined in Section 3.3 in the main article. We set the nominal FDR at the level of 0.05 for all studies. As a result, the discoveries from different studies are comparable, as they are all detected gene-SNP pairs with eQTL in tissue a at the FDR of 0.05. Table S4 shows the TPRs and FDRs in different studies. The number of true alternative cases is 1,596,410. The TPRs increase steadily with the number of auxiliary tissues considered in the analysis, while the FDRs are all controlled at the nominal level. The result indicates that by borrowing information from auxiliary tissues, the model gains power of detecting eQTL in a target tissue without inflating the FDR. Similar results hold for more sophisticated hypothesis testings.

5. GTEx Estimations

The sample information of the GTEx pilot data is provided in Figure S1. The estimated model parameters Δ and Σ for the GTEx data are given below. The tissues are ordered alphabetically. The parameter $\boldsymbol{\mu}_0$ was set to zero. The estimated mass function $\mathbf{p}$ (prior probabilities for 512 configurations) is provided in a separate text file (SuppC-p.txt) due to space limitations.

$$\Delta = \begin{pmatrix} 1.0000 & 0.1704 & 0.0923 & 0.1010 & 0.1390 & 0.1409 & 0.1687 & 0.1415 & 0.1441 \\ 0.1704 & 1.0000 & 0.0960 & 0.1179 & 0.1518 & 0.1460 & 0.1942 & 0.1336 & 0.1491 \\ 0.0923 & 0.0960 & 1.0000 & 0.0779 & 0.1312 & 0.0780 & 0.1007 & 0.0890 & 0.1032 \\ 0.1010 & 0.1179 & 0.0779 & 1.0000 & 0.1268 & 0.1192 & 0.1093 & 0.0893 & 0.1247 \\ 0.1390 & 0.1518 & 0.1312 & 0.1268 & 1.0000 & 0.1188 & 0.1543 & 0.1220 & 0.1767 \\ 0.1409 & 0.1460 & 0.0780 & 0.1192 & 0.1188 & 1.0000 & 0.1366 & 0.1095 & 0.1258 \\ 0.1687 & 0.1942 & 0.1007 & 0.1093 & 0.1543 & 0.1366 & 1.0000 & 0.1372 & 0.1477 \\ 0.1415 & 0.1336 & 0.0890 & 0.0893 & 0.1220 & 0.1095 & 0.1372 & 1.0000 & 0.1097 \\ 0.1441 & 0.1491 & 0.1032 & 0.1247 & 0.1767 & 0.1258 & 0.1477 & 0.1097 & 1.0000 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 4.2692 & 4.5320 & 4.1062 & 3.2993 & 4.6078 & 4.0864 & 4.2076 & 3.9694 & 4.4595 \\ 4.5320 & 5.4178 & 4.4545 & 3.6526 & 5.0411 & 4.5731 & 4.6975 & 4.3167 & 5.0072 \\ 4.1062 & 4.4545 & 6.1588 & 3.3196 & 5.0385 & 4.2452 & 4.0646 & 4.0090 & 4.5213 \\ 3.2993 & 3.6526 & 3.3196 & 3.2123 & 3.7223 & 3.6852 & 3.3418 & 3.1225 & 3.7332 \\ 4.6078 & 5.0411 & 5.0385 & 3.7223 & 5.5488 & 4.5088 & 4.6816 & 4.5263 & 5.2369 \\ 4.0864 & 4.5731 & 4.2452 & 3.6852 & 4.5088 & 5.1569 & 4.0399 & 3.9304 & 4.3674 \\ 4.2076 & 4.6975 & 4.0646 & 3.3418 & 4.6816 & 4.0399 & 4.5993 & 4.0265 & 4.6699 \\ 3.9694 & 4.3167 & 4.0090 & 3.1225 & 4.5263 & 3.9304 & 4.0265 & 4.3420 & 4.4163 \\ 4.4595 & 5.0072 & 4.5213 & 3.7332 & 5.2369 & 4.3674 & 4.6699 & 4.4163 & 5.6492 \end{pmatrix}.$$

The fitting times of MT-eQTL and the Meta-Tissue method (Sul *and others*, 2013) on a sequence of sub-models of different dimensions based on alphabetically ordered tissues are presented in Table S5. (We note that fitting sub-models of MT-eQTL is unnecessary in practice, as one can obtain them through marginalization of the full model.) The use of configuration vectors in MT-eQTL makes analysis results more interpretable, but it also makes the runtime of MT-eQTL sensitive to the number of tissues. Nevertheless, our method is computationally efficient when the number of tissues is moderate. On the other hand, the Meta-Tissue method is less restricted by the number of tissues, but it is more sensitive to the total number of gene-SNP pairs. The runtime for Meta-Tissue may quickly become impractical when there are too many gene-SNP pairs.

References

Shabalin, Andrey A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**(10), 1353–1358.

Sul, Jae Hoon, Han, Buhm, Ye, Chun, Choi, Ted and Eskin, Eleazar. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genetics* **9**(6), e1003491.

[*Received XXXXX; revised XXXXX; accepted for publication XXXXX*]

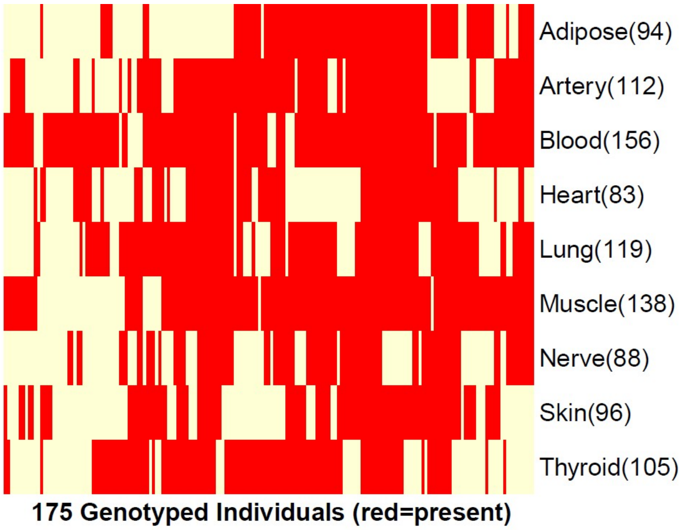

Fig. S1: Sample information of the GTEx data. Each column represents a genotyped individual with expression measurements in at least one tissue; each row corresponds to a tissue. Red means the individual is a donor of the corresponding tissue.

Table S1: Sample sizes (diagonal), sample overlap (off-diagonal), and degrees of freedom for different tissues in the simulation.

	a	b	c	d	Degree of Freedom
a	156	104	122	90	137
b		119	100	84	100
c			138	88	119
d				105	86

Table S2: The true generating model parameters $(\Delta, \Sigma, \mathbf{p})$ for the simulation study. The prior probabilities are provided for all possible eQTL configurations represented by 4-digit 0/1 sequences: 0 means no eQTL and 1 indicates the presence of eQTL in a tissue.

(a) Δ

	a	b	c	d
a	1.0000	0.1347	0.0805	0.1089
b	0.1347	1.0000	0.1204	0.1794
c	0.0805	0.1204	1.0000	0.1288
d	0.1089	0.1794	0.1288	1.0000

(b) Σ

	a	b	c	d
a	6.5699	5.3098	4.4683	4.7126
b	5.3098	5.9752	4.7906	5.5778
c	4.4683	4.7906	5.5263	4.6493
d	4.7126	5.5778	4.6493	6.0178

(c) $\mathbf{p}$

Config (abcd)	0000	0001	0010	0011	0100	0101	0110	0111
Prior	0.7721	0.0202	0.0190	0.0037	0.0104	0.0033	0.0010	0.0107
Config (abcd)	1000	1001	1010	1011	1100	1101	1110	1111
Prior	0.0196	0.0010	0.0008	0.0009	0.0029	0.0085	0.0019	0.1240

Table S3: A variety of eQTL detection inferences with the MT-eQTL model in the 4-tissue simulation study. From top to bottom, we aim to identify eQTLs: 1) in at least one tissue; 2) in tissue a (the null consists of all configurations with 0 in the first position); 3) in tissue b; 4) in tissue c; 5) in tissue d; 6) in all 4 tissues.

Index	Null Config	# Alternative Cases	# Discoveries	# Overlaps	TPR	FDR
1)	0000	2,279,307	1,038,456	987,083	.4331	.0495
2)	0***	1,596,410	679,207	645,700	.4045	.0493
3)	*0**	1,626,961	746,265	709,258	.4359	.0496
4)	**0*	1,618,655	663,367	630,502	.3895	.0495
5)	***0	1,722,789	770,722	732,600	.4252	.0495
6)	all but 1111	1,239,630	417,867	397,341	.3205	.0491

Table S4: The TPRs and FDRs of *detecting eQTL in tissue a* using the MT-eQTL model on tissue sets {a}, {a,b}, {a,b,c}, and {a,b,c,d}.

	{a}	{a,b}	{a,b,c}	{a,b,c,d}
TPR	.2753	.3475	.3806	.4045
FDR	.0500	.0499	.0496	.0493

Table S5: Approximate fitting times for k-dimensional MT-eQTL models and Meta-Tissue methods using the GTEx data.

Time	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$
MT-eQTL	$<$ 1 min	15 min	30 min	1 hr	2.5hr	6hr	11hr	16 hr	24 hr
Meta-Tissue	130 min	165 min	165 min	3 hr	3 hr	3.25 hr	3.5 hr	4 hr	5 hr