An Empirical Bayes Approach for Multiple Tissue eQTL Analysis

Gen Li[1], Andrey A. Shabalin[2], Ivan Rusyn[3], Fred A. Wright[4] and Andrew B. Nobel[1]

[1]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill

[2]Department of Pharmacotherapy & Outcomes Science, Virginia Commonwealth University

[3]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill

[4]Bioinformatics Research Center, North Carolina State University

Abstract

Expression quantitative trait loci (eQTL) analyses, which identify genetic markers associated with the expression of a gene, are an important tool in the understanding of diseases in human and other populations. While most eQTL studies to date consider the connection between genetic variation and expression in a single tissue, complex, multi-tissue data sets are now being generated by the GTEx initiative. These data sets have the potential to improve the findings of single tissue analyses by borrowing strength across tissues, and the potential to elucidate the genotypic basis of differences between tissues.

In this paper we introduce and study a multivariate hierarchical Bayesian model (MT-eQTL) for multi-tissue eQTL analysis. MT-eQTL directly models the vector of correlations between expression and genotype across tissues. It explicitly captures patterns of variation in the presence or absence of eQTLs, as well as the heterogeneity of effect sizes across tissues. Moreover, the model is applicable to complex designs in which the set of donors can (i) vary from tissue to tissue, and (ii) exhibit incomplete overlap between tissues. The MT-eQTL model is marginally consistent, in the sense that the model for a subset of tissues can be obtained from the full model via marginalization. Fitting of the MT-eQTL model is carried out via empirical Bayes, using an approximate EM algorithm. Inferences concerning eQTL detection and the configuration of eQTLs across tissues are derived from adaptive thresholding of local false discovery rates, and maximum a-posteriori estimation, respectively. We investigate the MT-eQTL model through a simulation study, and rigorously establish the FDR control of the local FDR testing procedure under mild assumptions appropriate for dependent data.

Contents

1 Introduction

Genetic variation in a population is commonly studied through the analysis of single nucleotide polymorphisms (SNPs), which are variants occurring at specific sites in the genome. Differences among these variants drive primary phenotypic differences between members of the population. For humans these differences range from physical characteristics to disease susceptibility. Mediating the connection between genetic variation and resulting phenotypes are the effects of SNPs on the expression of different genes. Expression quantitative trait locus (eQTL) analysis seeks to identify genetic variants that affect the expression of one or more genes: a gene-SNP pair for which the expression of the gene is associated with the value of the SNP is referred to as an eQTL. Enabled by high-throughput sequencing, eQTL analysis has proven to be an effective approach for the discovery of genomic variants that influence expression, and a potentially useful tool in the study of pathways and networks that underlie disease in human and other populations. For an overview of eQTL analysis and disease mapping, see Cookson et al. (2009), Mackay et al. (2009), Rockman and Kruglyak (2006), and the references therein. Kendziorski and Wang (2006) and Wright et al. (2012) survey existing statistical and computational methods for eQTL analysis, respectively.

To date, most eQTL studies have considered the effects of genetic variation on expression within a single tissue. Nonetheless, these studies have provided enhanced understanding of gene regulation and the etiology of various diseases, *cf.* Franke and Jansen (2009) and Westra et al. (2013). A natural next step in understanding genomic variation of expression is the simultaneous analysis of eQTLs in multiple tissues. Multi-tissue eQTL analysis has the potential to improve the findings of single tissue analyses by borrowing strength across tissues, and to expand the scope of single tissue analyses by addressing more fundamental biological questions about the nature and source of variation between tissues.

In a single tissue eQTL study, the goal is to identify gene-SNP pairs for which the expression of the gene is associated with the SNP genotype. An important feature of multiple tissue studies is that a SNP may be associated with the expression of a gene in some tissues, but not in others. Thus a full multi-tissue analysis must identify

complex patterns of association across multiple tissues. We will refer to an eQTL as 'common' if association is present in all available tissues, and 'tissue-specific' if association is present in at least one tissue, but not all. Until recently, understanding of multi-tissue eQTL relationships was limited by a shortage of true multi-tissue data sets, requiring the assimilation of data or results from different studies (one for each tissue) involving distinct populations, measurement platforms, and analysis protocols, *cf.* Emilsson et al. (2008) and Xia et al. (2012).

Recently, a number of human true multi-tissue eQTL data sets have been collected, for example by Dimas et al. (2009) and Nica et al. (2011), although these contain relatively few tissues. By contrast, the GTEx initiative (Lonsdale et al. (2013)) and related projects are generating eQTL data from dozens of tissues in several hundred individuals, greatly expanding our potential understanding of the variation and specificity of eQTL effects across multiple tissues. The size and complexity of these emerging multi-tissue data sets has created the need to expand existing statistical tools for eQTL analysis.

In this paper we introduce and study a multivariate, hierarchical Bayesian model for the simultaneous analysis of eQTLs in multiple tissues, which we call MT-eQTL. The dimension of the MT-eQTL model is equal to the number of tissues. Importantly, we do not seek to describe the full joint relationship between expression and genotype across tissues. Instead, we directly model the vector $\mathbf{z}$ of Fisher transformed correlations between expression and genotype across tissues, after appropriate scaling to account for different degrees of freedom in each tissue. The entries of $\mathbf{z}$ are z-statistics for testing the association between genotype and expression in each tissue. Working with the test statistics on the transformed scale facilitates modeling and interpretation. The upper panel of Figure 1b shows a density-based scatter plot of the $\mathbf{z}$-vectors from a simulated data set. The lower panel illustrates the results of the MT-eQTL model: vectors close to the origin for which no eQTLs are detected have been removed, resulting in the central white area; detected eQTLs are colored according to whether an eQTL is detected in both tissues (blue points) or a single tissue (red and green points).

The MT-eQTL model can be expressed in an equivalent, mixture form in which each component corresponds to a binary configuration indicating the presence (1)

or absence (0) of an eQTL in each tissue. We adopt an empirical Bayes approach, fitting the MT-eQTL model by maximum likelihood using an EM based algorithm. Throughout we restrict attention to local (sometimes referred to as 'cis') gene-SNP pairs, for which the SNP is within a fixed genomic distance of the coding region of the gene.

We briefly describe some of the key features of the MT-eQTL model. A detailed description is given in Section 2. The model explicitly captures patterns of variation in the presence or absence of eQTLs, as well as the heterogeneity of effect sizes across tissues. In complex multi-tissue data like that from GTEx, the number of samples can vary substantially from tissue to tissue, and the sets of donors for different tissues can exhibit different degrees of overlap. The MT-eQTL model is rich enough to accommodate both of these features. Another important aspect of complex multi-tissue data is that effect sizes in different tissues may be correlated. Correlations in effect sizes arise from biological factors (for example, the underlying relationships among tissues), and are reflected in the correlation structure of the vector $\mathbf{z}$. The correlation structure of $\mathbf{z}$ also reflects experimental factors such as donor overlap among tissues. The MT-eQTL model explicitly accounts for both sources of correlation in an identifiable way. Lastly, the MT-eQTL model has the desirable property of being marginally consistent: roughly speaking, the mixture model for a subset of tissues can be obtained from the full mixture model via marginalization.

Fitting of the MT-eQTL model from the $\mathbf{z}$-vectors of local gene-SNP pairs is carried out via empirical Bayes using an approximate EM algorithm. Fitting is fast enough to accommodate the full analysis of real data sets on a desktop computer. After fitting, the MT-eQTL model provides, for any given $\mathbf{z}$-vector, posterior probabilities for every binary configuration of eQTL absence (0) or presence (1) across tissues. Using the fitted model, we define the local false discovery rate of a gene-SNP pair to be the posterior probability of the zero configuration (no eQTL in any tissue) given its vector of $\mathbf{z}$-statistics. We test for gene-SNP pairs having an eQTL in some tissue by adaptive thresholding of the local false discovery rates. Assessment of tissue specificity can be obtained from the posterior probabilities of non-zero configurations. The procedure is readily generalized to more general hypothesis testing settings.

1.1 Related work

Research on multi-tissue eQTLs is relatively new, with early published work dating from 2007. Most existing multi-tissue analyses extract eQTLs individually from each tissue and then apply post-hoc procedures to assess commonality and specificity. Dimas et al. (2009) and Heinzen et al. (2008) consider the simple pairwise overlap of single tissue eQTL discoveries. Ding et al. (2010) proposed a procedure to measure eQTL overlap that accounts for differences in statistical power between data sets for individual tissues. Fu et al. (2012) proposed a resampling based procedure to assess the tissue-specificity of cis-eQTLs. Bullaughey et al. (2009) examined the gene-SNP associations in five human primary tissues of eQTLs with large effect sizes in lymphoblastoid cell lines. A similar idea is implemented in Nica et al. (2011): given a set of gene-SNP pairs with small p-values in one tissue, the p-values of these same pairs are examined in other tissues to assess enrichment of significant associations. In addition, several meta-analysis based approaches have been applied to integrate eQTL results for different tissues, *cf.* Brown et al. (2013) and Xia et al. (2012).

The papers cited above provide exploratory studies of eQTLs in multiple tissues, or pairwise conditional analysis of eQTLs declared significant in an initial tissue. However, they do not address the *ab-initio* statistical analysis of multi-tissue data in a manner that fully utilizes the data. Gerrits et al. (2009) used an ANOVA model to jointly analyze gene-SNP associations across tissues, with eQTL configurations assigned according to effect sizes in different tissues. Petretto et al. (2010) proposed a sparse Bayesian regression model in which gene expression in different tissues is treated as a multivariate response, and SNPs are treated as predictors; the presence and specificity of eQTLs are captured by a sparse coefficient matrix. Following Wen and Stephens (2011), Flutre et al. (2013) proposed a Bayesian framework for the joint analysis of eQTLs across tissues. They use a linear model to capture gene-SNP association in each tissue, and place a prior distribution on the coefficients subject to a latent indicator of whether or not it is an eQTL. Each of these methods uses permutation based procedures to control and calculate false discovery rates., which is computationally burdensome when dealing with millions of gene-SNP pairs and multiple tissues. In addition, these methods assume that each tissue has samples from an identical set of individuals; as noted above, in many cases the set and number of

donors varies from tissue to tissue.

In recent work, Sul et al. (2013) proposed a"Meta-Tissue" method that combines linear mixed models and meta-analysis. The linear mixed model captures gene-SNP correlations across tissues and accounts for partial overlap among donors. Meta-analysis is used to address detection of eQTLs in multiple tissues, but the model does not use an explicit indicator vector for eQTLs across tissues, making assignment of tissue specificity less straightforward than with other methods. Moreover, their hypothesis testing procedure does not make direct use of the alternative distribution, which may lead to a reduction in statistical power.

1.2 Outline

Specification and fitting of the MT-eQTL empirical Bayes model is described in the next section. Section 3 describes how the MT-eQTL can be applied to multi-tissue inference, including eQTL detection using a simple step-up procedure based on the local false discovery rate, and the determination of eQTL tissue specificity. Theorem 3.2 establishes the asymptotic FDR control of the step-up procedure. Section 4 explores the MT-eQTL model in a 4-tissue simulation study based on an ongoing analysis of data from the GTEx project. The modified EM algorithm used to fit the MT-eQTL model is described in Appendix A. Appendix B contains the proofs of the marginal consistency of the MT-eQTL model. The proof of Theorem 3.2 is given in Appendix C.

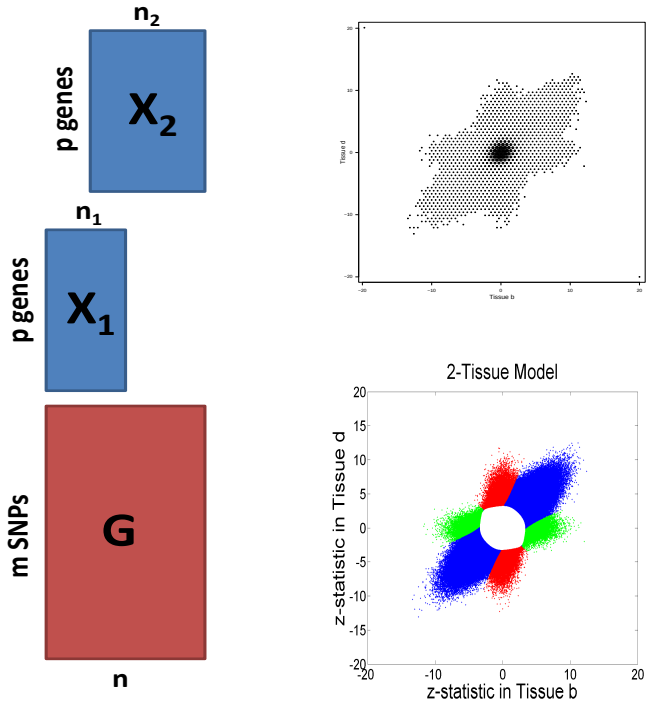
2 The MT-eQTL Model

In this section we describe the MT-eQTL model in detail, beginning with a general description of multi-tissue data, and a detailed account of the multivariate z-statistics on which the model is based.

2.1 Format of Multi-Tissue eQTL Data

The general data format for the multi-tissue eQTL problem is as follows. For each of n donors we have full genotype information, and measurements of gene expression

in at least one of K tissues. We assume that the same array platform is used for measurements of genotype, and similarly for expression.



(a) Raw data format

(b) MT-eQTL input and output

Figure 1: (a) Illustration of the typical data format with two tissues. Genotype data G is available for m SNPs and each of n samples. Expression measurements are available for p genes; sample sets for different tissues may not be the same. (b) Scatter plots of z-vectors from a simulated data set: for all simulated gene-SNP pairs (top), and for significant discoveries with tissue specificity by color (bottom).

Let $\mathbf{G}$ be an $m \times n$ matrix containing the measured genotype of each donor in the study at m genetic loci that are the sites of single nucleotide polymorphisms (SNPs). Each column of $\mathbf{G}$ corresponds to a donor, and each row corresponds to a locus/SNP. The measured transcript levels for tissue k are contained in a $p \times n_k$ matrix $\mathbf{X}_k$, where p is the number of measured transcripts, and $n_k \leq n$ is the number of donors from which samples of tissue k are available. Each column of $\mathbf{X}_k$ has an identifier indicating the donor associated with the measurements in that column. In general, the number of donors n_k can vary widely among tissues, and even if two tissues have

similar numbers of samples, they may have relatively few common donors. The data available for the purposes of multi-tissue eQTL analysis has the form $(\mathbf{G}, \mathbf{X}_1, \ldots, \mathbf{X}_K)$. Figure 1a gives an illustration of the typical data format with two tissues.

2.1.1 Data Preprocessing and Covariate Adjustment

In most cases eQTL analysis is preceded by several preprocessing steps and covariate adjustment. The genotype data matrix $\mathbf{G}$ consists of values 0, 1, and 2, typically coded as the number of minor allele variants; SNPs with too few minor allele instances are often discarded. Expression measurements may be obtained from array-based platforms or from RNA-Seq tag counts. Lowly expressed genes are typically dropped from the analysis.

Genotype and expression data may contain confounding factors. Some confounders, such as gender, are observed, while others are of unknown technical or biological origin. To identify the unknown confounding factors, most studies use principal components, surrogate variables (Leek and Storey, 2007), or PEER cofactors (Stegle et al., 2012) as covariates. We assume that the expression data and genotype data have been residualized for the confounders, so the comparison of these residualized quantities are partial correlations adjusted for covariates. The degrees of freedom lost in fitting the covariates is accounted for in computing the association between expression and genotype.

2.2 Multivariate z-Statistic from Single Tissue Correlations

Denote a measured transcript by $i \in \{1, \ldots, p\}$ and a measured genotype by $j \in \{1, \ldots, m\}$. We focus on a subset Λ of the full index set $\{1, \ldots, p\} \times \{1, \ldots, m\}$ that consists of pairs (i, j) such that SNP j is located within a fixed distance (usually 100 Kilobases or 1 Megabase) of the transcription start site (TSS) of gene i.

Let $\lambda = (i, j)$ be a gene-SNP pair of interest, and let k be a tissue for which measurements of transcript i are available. Let $r_{\lambda k}$ and $\rho_{\lambda k}$ denote, respectively, the sample and population correlation of transcript i and SNP j in tissue k. Note that the sample correlation $r_{\lambda k}$ depends only on the n_k measurements from donors of tissue k. The vector of correlations $\mathbf{r}_\lambda = (r_{\lambda 1}, \ldots, r_{\lambda K})$ captures the association between

the expression of transcript i and the value of genotype j in each of the K tissues. Relationships between different tissues will be reflected in correlations between the entries of $\mathbf{r}_\lambda$. These features make $\mathbf{r}_\lambda$ a natural starting point for a multi-tissue eQTL model.

In order to construct a multivariate model for the correlations $\mathbf{r}_\lambda$, it is convenient to work in a Gaussian setting. To this end, let

$$\mathbf{h}(\mathbf{r}_\lambda) \;=\; \big(h(r_{\lambda 1}), \ldots, h(r_{\lambda K})\big)$$

be the vector obtained by applying the Fisher transformation

$$h(r) = \frac{1}{2}\log\Big(\frac{1+r}{1-r}\Big)$$

to each component of $\mathbf{r}_\lambda$. Let

$$\mathbf{d}^{1/2} := (\sqrt{d_1 - 3}, \ldots, \sqrt{d_K - 3})$$

be a scaling vector, where d_k is the degrees of freedom for $\mathbf{X}_k$ and $\mathbf{G}$, equal to n_k minus the number of covariates used to correct genotype and expression for samples in tissue k. Finally, define the vector

$$\mathbf{z}_\lambda \;=\; \mathbf{d}^{1/2} \cdot \mathbf{h}(\mathbf{r}_\lambda) \tag{1}$$

where $\mathbf{u} \cdot \mathbf{v}$ denotes the Hadamard (entry-wise) product of vectors $\mathbf{u}$ and $\mathbf{v}$.

Consider a random vector $\mathbf{Z}_\lambda$ derived in the same fashion as $\mathbf{z}_\lambda$ from random data $(\mathbf{G}, \mathbf{X}_1, \ldots, \mathbf{X}_K)$. We assume that the expression measurements $\mathbf{X}_k$ are approximately normal. Standard arguments for the Fisher transformation (Winterbottom, 1979) show that $h(r_{\lambda k})$ is approximately normal with mean $h(\rho_{\lambda k})$ and variance $(d_k - 3)^{-1}$. By a routine multivariate extension of this fact, $\mathbf{Z}_\lambda$ is approximately normally distributed with mean

$$\boldsymbol{\mu}_\lambda \;=\; \mathbf{d}^{-1/2} \cdot \mathbf{h}(\boldsymbol{\rho}_\lambda).$$

The variance stabilizing property of the Fisher transformation and our choice of scaling ensures that the variance of each entry $Z_{\lambda k}$ of $\mathbf{Z}_\lambda$ is close to one, regardless of $\boldsymbol{\rho}_\lambda$. In particular, if the true correlation $\rho_{\lambda k}$ between transcript i and SNP j for tissue k is zero, then $Z_{\lambda k}$ is approximately standard normal. Thus the k-th entry of the

observed vector $\mathbf{z}_\lambda$ is a z-statistic for testing $\rho_{\lambda k} = 0$ vs. $\rho_{\lambda k} \neq 0$. Importantly, the components of $\mathbf{Z}_\lambda$ need not be independent, even when all the true correlations $\rho_{\lambda k}$ are zero. Capturing this dependence is a key feature of the MT-eQTL model, which is described in detail below.

2.3 Hierarchical Model

Let $\lambda = (i, j)$ be a gene-SNP pair in Λ. MT-eQTL is a multivariate, hierarchical Bayesian model for the random vector $\mathbf{Z}_\lambda$. In detail, we assume that

$$\begin{aligned} \mathbf{Z}_\lambda \,|\, \boldsymbol{\mu}_\lambda &\sim \mathcal{N}_K\left(\boldsymbol{\mu}_\lambda, \Delta\right) && (2) \\ \boldsymbol{\mu}_\lambda &= \boldsymbol{\Gamma}_\lambda \cdot \boldsymbol{\alpha}_\lambda && (3) \\ \boldsymbol{\Gamma}_\lambda &\sim \mathbf{p} \text{ on } \{0,1\}^K && (4) \\ \boldsymbol{\alpha}_\lambda &\sim \mathcal{N}_K(\boldsymbol{\mu}_0, \Sigma), \text{ independent of } \boldsymbol{\Gamma}_\lambda && (5) \end{aligned}$$

The mean vector $\boldsymbol{\mu}_\lambda$ contains the effect sizes for the relationship between transcript i and SNP j in each tissue. The $K \times K$ covariance matrix Δ is constrained to have diagonal entries equal to one, reflecting the variance stabilization of the Fisher transformation, and the scaling in (1). The off-diagonal entries of Δ capture correlations among the entries of $\mathbf{Z}_\lambda$ that are due to commonalities among tissues that arise from the underlying sampling process, for example, correlations resulting from shared donors among a pair of tissues.

We assume that the mean vector $\boldsymbol{\mu}_\lambda$ of $\mathbf{Z}_\lambda$ is equal to the entrywise product of a multinormal random vector $\boldsymbol{\mu}_\lambda$ and a vector $\boldsymbol{\Gamma}_\lambda$ with binary entries. The indicator vector $\boldsymbol{\Gamma}_\lambda$ determines the presence ($\Gamma_{\lambda k} = 0$) or absence ($\Gamma_{\lambda k} = 1$) of an association between transcript i and SNP j in tissues $k = 1, \ldots, K$. The strength of an association, when present, is determined by the corresponding component of $\boldsymbol{\alpha}_\lambda$. The covariance matrix Σ of $\boldsymbol{\alpha}_\lambda$ captures tissue specific variation in effect sizes, and correlations among effect sizes that reflect biological commonalities between tissues. The mean vector $\boldsymbol{\mu}_0$ of $\boldsymbol{\alpha}_\lambda$ captures the average effect sizes across tissues. In practice we usually set $\boldsymbol{\mu}_0 = \mathbf{0}$ because high expression levels of a gene can be associated with either the major or minor allele with roughly equal probability, resulting in average effect sizes to be approximately zero across tissues. We have noticed little effect of

this setting on numerical results. The final parameter of the model is a probability mass function $\mathbf{p}$ on $\{0,1\}^K$ that assigns probabilities to each of the 2^K possible configurations of $\mathbf{\Gamma}_\lambda$. In particular, $p_{\mathbf{0}}$ (i.e., $p_{(0,\cdots,0)}$) is the prior probability that transcript i and SNP j have no association in any tissue.

2.4 Mixture Model

The hierarchical model (2)-(5) describing the distribution of $\mathbf{Z}_\lambda$ is fully specified by $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$, which consists of $2^K + K^2 + K - 1$ real-valued parameters. Estimation of, and inference from, the hierarchical model is based on an equivalent mixture representation that we now discuss.

If $\mathbf{U}$ is distributed as $\mathcal{N}_K(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{\gamma}$ is a fixed vector in $\{0,1\}^K$, then one may readily verify that the entrywise product $\mathbf{U} \cdot \boldsymbol{\gamma}$ is distributed as $\mathcal{N}_K\big(\boldsymbol{\mu} \cdot \boldsymbol{\gamma}, \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\big)$. A straightforward argument then shows that the hierarchical model (2)-(5) is equivalent to a mixture distribution of the form

$$\mathbf{Z}_\lambda \;\sim \sum_{\boldsymbol{\gamma} \in \{0,1\}^K} p_{\boldsymbol{\gamma}}\, \mathcal{N}_K\big(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma},\, \Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\big). \tag{6}$$

We adopt an empirical Bayes approach for performing inference from the model (6). Specifically, the parameters $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$ of the hierarchical model are estimated from the observed z-statistics $\{\mathbf{z}_\lambda : \lambda \in \Lambda\}$ by approximately maximizing a pseudo-likelihood derived from (6); see Appendix A for more details. Beginning with the work of Newton et al. (2001) and Efron et al. (2001), empirical Bayes approaches have been applied to hierarchical models in a number of genetic applications, most notably the study of differential expression and co-expression in gene expression microarrays, cf. Kendziorski et al. (2003), Newton et al. (2004), Smyth et al. (2004) and Efron (2008), and Dawson and Kendziorski (2012).

The mixture model (6) is readily interpretable. Each component of the model corresponds to a unique configuration $\boldsymbol{\gamma}$, or equivalently, a unique pattern of tissue specificity. The model component corresponding to $\boldsymbol{\gamma} = \mathbf{0}$ represents the case in which there are no eQTLs in any tissue, and has associated (null) distribution $\mathcal{N}_K(\mathbf{0}, \Delta)$. The model component corresponding to $\boldsymbol{\gamma} = \mathbf{1}$ represents the case in which there are eQTLs in every tissue, and has associated distribution $\mathcal{N}_K(\boldsymbol{\mu}_0, \Delta + \Sigma)$. Other values

of $\boldsymbol{\gamma}$ represent intermediate cases in which there are eQTLs in some tissues (those with $\gamma_k = 1$) and not in others (those with $\gamma_k = 0$).

2.5 Marginal Consistency

In eQTL studies with multiple tissues, it is likely that some subsets of the tissues are of particular interest. From the point of view of model fitting and model interpretation, it is desirable if the model for any subset of tissues is consistent with the full model in the sense that it can be obtained from the full model (or any model on a superset of tissues) via marginalization. We refer to this property as *marginal consistency*.

To elaborate, let $S \subseteq \{1, \ldots, K\}$ be a subset of r tissues, with $1 \leq r \leq K$. The mixture model (6) has two important compatibility properties: (i) the marginalization of the full model to S has the same general form as the model derived from S alone; and (ii) the parameters of the marginal model are obtained by restricting the parameters of the full model to S. The following definition and lemma makes these statements precise. A proof of the lemma is given in the appendix.

Definition: Let $S \subseteq \{1, \ldots, K\}$ with cardinality $|S| = r$. For each vector $\mathbf{u} \in \mathbb{R}^K$ let $\mathbf{u}_S = (u_k : k \in S) \in \mathbb{R}^r$ be the vector obtained by restricting $\mathbf{u}$ to the entries in S. Similarly, for each matrix $A \in \mathbb{R}^{K \times K}$ let $A_S = \{a_{kl} : k, l \in S\}$ be the $r \times r$ matrix obtained by retaining only the rows and columns with indices in S. Note that if A is non-negative (positive) definite, then A_S is non-negative (positive) definite as well.

Lemma 2.1. *If $\mathbf{Z} \in \mathbb{R}^K$ be a random vector having the mixture distribution (6), then*

$$\mathbf{Z}_S \;\sim\; \sum_{\boldsymbol{\zeta} \in \{0,1\}^r} p_{S,\boldsymbol{\zeta}}\, \mathcal{N}_r\big(\boldsymbol{\mu}_{0S} \cdot \boldsymbol{\zeta},\, \Delta_S + \Sigma_S \cdot \boldsymbol{\zeta}\boldsymbol{\zeta}^T\big) \tag{7}$$

where $(p_{S,\mathbf{0}}, \cdots, p_{S,\mathbf{1}})$ is the probability mass function on $\{0,1\}^r$ obtained by marginalizing $\mathbf{p}$ to S, i.e., $p_{S,\boldsymbol{\zeta}} = \sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}_S = \boldsymbol{\zeta}} p_{\boldsymbol{\gamma}}$.

Remark: Suppose that the parameters $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$ of the full mixture model (6) are estimated from the z-statistic vectors $\{\mathbf{z}_\lambda : \lambda \in \Lambda\}$, and let $S \subseteq \{1, \ldots, K\}$ be a set containing r tissues. Lemma 2.1 describes the model θ_S obtained by marginalizing the full model to the tissue set S.

3 Multi-Tissue eQTL Inference

Once fit, the mixture model (6) provides the basis for inference about eQTLs across tissues. In practice, we expect that θ will be well-estimated due to the large number of available gene-SNP pairs; we therefore regard θ as fixed and known. For data sets with small sample sizes, approximate standard errors can be obtained from the likelihood via the observed information matrix.

In most applications the covariance matrix Δ will be positive definite, and we assume this is the case here. With this assumption, the distribution $\mathcal{N}_K\big(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma},\, \Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\big)$ associated with the configuration $\boldsymbol{\gamma} \in \{0,1\}^K$ has a density, which we denote by $f_{\boldsymbol{\gamma}}$. Thus under the mixture model (6) the random vector $\mathbf{Z}_\lambda$ has density

$$f(\mathbf{z}) \;=\; \sum_{\boldsymbol{\gamma}} p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z}) \qquad \mathbf{z} \in \mathbb{R}^K. \tag{8}$$

In view of this expression and the hierarchical model (2)-(5), one may regard $\mathbf{Z}_\lambda$ as one element of a jointly distributed pair $(\boldsymbol{\Gamma}_\lambda, \mathbf{Z}_\lambda)$, where

$$\boldsymbol{\Gamma}_\lambda \sim \mathbf{p} \;\text{ and }\; \mathbf{Z}_\lambda \,|\, \boldsymbol{\Gamma}_\lambda \;\sim\; f_{\boldsymbol{\gamma}}. \tag{9}$$

We carry out multi-tissue eQTL analysis based on the posterior distribution of the configuration $\boldsymbol{\Gamma}_\lambda$ given the observed vector of z-statistics $\mathbf{z}_\lambda$. Two inference problems are of central interest to us. The first is eQTL detection, in all tissues and in a subset of tissues. The second is assessing the tissue specificity of eQTLs in transcript-SNP pairs where an eQTL is present in at least one tissue.

3.1 Detection of eQTLs Using the Local False Discovery Rate

A primary goal of multi-tissue analysis is testing each transcript-SNP pair for the presence of an eQTL in at least one tissue. This can be formulated as a multiple testing problem, namely testing

$$\mathrm{H}_{0,\lambda} : \boldsymbol{\Gamma}_\lambda = \mathbf{0} \;\text{ versus }\; \mathrm{H}_{1,\lambda} : \boldsymbol{\Gamma}_\lambda \neq \mathbf{0} \;\text{ for }\; \lambda \in \Lambda. \tag{10}$$

For $\lambda = (i,j) \in \Lambda$ the null hypothesis $H_{0,\lambda}$ asserts that there is no eQTL between transcript i and SNP j in any tissue, while the alternative $H_{1,\lambda}$ asserts that there is an eQTL between i and j in at least one tissue.

The null hypotheses can also be expressed in the form $\mathrm{H}_{0,\lambda} : \mathbf{Z}_\lambda \sim \mathcal{N}_K\big(\mathbf{0},\, \Delta\big)$. It is possible to derive a p-value for $\mathbf{z}_\lambda$ directly from the null distribution, and then control the overall false discovery rate in (10) using a step-up procedure like that of Benjamini and Hochberg (1995). However, this type of analysis ignores relevant information about the distribution of $\mathbf{Z}_\lambda$ under the alternative that is contained in the mixture model.

We address the multiple testing problem (10) using the local false discovery rate introduced by Efron et al. (2001) in the context of an empirical Bayes analysis of differential expression in microarrays. Other applications of the local false discovery rate to genomic problems can be found in Newton et al. (2004), Efron (2007), and Efron (2008). To simplify notation in what follows, let $(\mathbf{\Gamma}, \mathbf{Z})$ denote a generic pair distributed as $(\mathbf{\Gamma}_\lambda, \mathbf{Z}_\lambda)$.

Definition: The *local false discovery rate* of an observed z-statistic vector $\mathbf{z}$ under the model (6) is defined by

$$\eta(\mathbf{z}) \;:=\; \mathbb{P}(\mathbf{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} = \mathbf{z}) \;=\; \frac{p_{\mathbf{0}} f_{\mathbf{0}}(\mathbf{z})}{f(\mathbf{z})}. \tag{11}$$

Let $\alpha \in (0, 1)$ be a target false discover rate (FDR) for the multiple testing problem (10). Vectors $\mathbf{z}$ for which the local false discovery rate $\eta(\mathbf{z})$ is small provide evidence for the alternative $\mathbf{\Gamma} \neq \mathbf{0}$. We carry out testing of gene-SNP pairs using a simple step-up procedure that is applied to the running average of the ordered local false discover rates. The procedure, which is described below, appears in essentially the same form in Newton et al. (2004), Sun and Cai (2007), and Cai and Sun (2009).

Local FDR Step-Up Procedure: Target FDR $= \alpha$

1. Given: Observed z-statistic vectors $\{\mathbf{z}_\lambda : \lambda \in \Lambda\}$.

2. Enumerate the elements of Λ as $\lambda_1, \ldots, \lambda_N$ so that $\eta(\mathbf{z}_{\lambda_1}) \leq \cdots \leq \eta(\mathbf{z}_{\lambda_N})$.

3. Reject hypotheses $\mathrm{H}_{0,\lambda_1}, \ldots, \mathrm{H}_{0,\lambda_L}$ where L is the largest integer such that $L^{-1} \sum_{l=1}^{L} \eta(\mathbf{z}_{\lambda_l}) \leq \alpha$.

In order to better understand the local FDR step-up procedure, and to assess its performance, it is useful to express the procedure in an equivalent form. As noted by

Efron et al. (2001), the false discovery rate associated with a rejection region $R \subseteq \mathbb{R}^k$ for the multiple testing problem (10) is given by $\mathbb{P}(\mathbf{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} \in R)$. They establish the following elementary fact, which exhibits a connection between the false discovery rate and the local false discovery rate.

Proposition 3.1. *If $R \subseteq \mathbb{R}^k$ is such that $\mathbb{P}(\mathbf{Z} \in R) > 0$, then $\mathbb{P}(\mathbf{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} \in R) = \mathbb{E}(\eta(\mathbf{Z}) \,|\, \mathbf{Z} \in R)$.*

As noted above, vectors $\mathbf{z}$ for which $\eta(\mathbf{z})$ is small provide evidence against $\mathbf{\Gamma} = \mathbf{0}$, so it is natural to reject $\mathrm{H}_{0,\lambda}$ when $\eta(\mathbf{z}_\lambda)$ falls below an appropriate threshold. Consider rejection regions of the form $R(t) = \{\mathbf{z} : \eta(\mathbf{z}) \leq t\}$ for $t \in (0, 1)$. Given a target false discovery rate α, we wish to find t such that $\alpha = \mathbb{P}(\mathbf{\Gamma} = \mathbf{0} \,|\, \mathbf{Z} \in R(t))$. By Proposition 3.1 this is equivalent to finding $t \in (0, 1)$ such that $F(t) = \alpha$, where

$$F(t) \;:=\; \mathbb{E}(\eta(\mathbf{Z}) \,|\, \eta(\mathbf{Z}) \leq t) \;=\; \frac{\mathbb{E}[\eta(\mathbf{Z})\, \mathbb{I}(\eta(\mathbf{Z}) \leq t)]}{\mathbb{P}(\eta(\mathbf{Z}) \leq t)}. \tag{12}$$

The empirical analog of $F(t)$ is the ratio

$$\hat{F}(t) \;=\; \frac{\sum_{\lambda \in \Lambda} \eta(\mathbf{z}_\lambda)\, \mathbb{I}(\eta(\mathbf{z}_\lambda) \leq t)}{\sum_{\lambda \in \Lambda} \mathbb{I}(\eta(\mathbf{z}_\lambda) \leq t)},$$

which depends only on $\eta(\cdot)$ and the observed vectors $\{\mathbf{z}_\lambda\}$. It is easy to see that the local FDR step-up procedure is equivalent to the rule

$$\text{Reject } \mathrm{H}_{0,\lambda} \text{ if and only if } \eta(\mathbf{z}_\lambda) \leq \sup\{t : \hat{F}(t) \leq \alpha\}. \tag{13}$$

We show in Proposition C.4 that $F(t)$ is strictly increasing and continuous. Thus if $F(t)$ and $\hat{F}(t)$ were equal, the local FDR step-up procedure and the idealized threshold procedure would coincide. In general, $F(t)$ and $\hat{F}(t)$ will be different, but multiplying the numerator and denominator of $\hat{F}(t)$ by $|\Lambda|^{-1}$ it is evident that the two functions will be close if $|\Lambda|$ is large and the dependence among the observed z-vectors is not extreme. Asymptotic control of the false discovery rate by the local FDR step-up procedure is established in Theorem 3.2 below.

Let $\Lambda^* \subseteq \mathbb{N} \times \mathbb{N}$ be an infinite index set, and let $\Lambda_1, \Lambda_2, \ldots \subseteq \Lambda^*$ be a sequence of finite subsets of Λ^*. Let $\alpha \in (0, 1)$ be a target FDR that is less than the maximum value of $\eta(\mathbf{z})$. For each $n \geq 1$ let $\{(\mathbf{\Gamma}_\lambda, \mathbf{Z}_\lambda) : \lambda \in \Lambda_n\}$ be jointly distributed pairs having the same distribution as $(\mathbf{\Gamma}, \mathbf{Z})$. In order to assess the performance of the local

FDR step-up procedure on the observed z-statistic vectors $\{\mathbf{Z}_\lambda : \lambda \in \Lambda_n\}$ we consider the equivalent rule (13), which rejects $H_{0,\lambda}$ when $\eta(\mathbf{Z}_\lambda) \leq \hat{\theta}_n = \sup\{t : \hat{F}_n(t) \leq \alpha\}$ where

$$\hat{F}_n(t) \;=\; \frac{\sum_{\lambda\in\Lambda_n} \eta(\mathbf{Z}_\lambda)\, \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leq t)}{\sum_{\lambda\in\Lambda_n} \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leq t)} \qquad 0 < t < 1.$$

The number of false discoveries and total discoveries for the local FDR step-up procedure are equal, respectively, to

$$M_n \;=\; \sum_{\lambda\in\Lambda_n} \mathbb{I}(\mathbf{\Gamma}_\lambda = 0)\, \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leq \hat{\theta}_n) \quad \text{and} \quad N_n \;=\; \sum_{\lambda\in\Lambda_n} \mathbb{I}(\eta(\mathbf{Z}_\lambda) \leq \hat{\theta}_n).$$

Theorem 3.2. *Let $(\mathbf{\Gamma}, \mathbf{Z})$ have joint distribution given by the mixture model (9) with parameters $(\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$. Assume that Δ is positive definite and that the diagonal entries of Σ are positive. If $\hat{F}_n(t) \to F(t)$ in probability for each $t \in (0,1)$ then $\mathbb{E}M_n/\mathbb{E}N_n \to \alpha$ as n tends to infinity.*

The proof of Theorem 3.2 is given in Appendix C below. The ratio of expectations $\mathbb{E}M_n/\mathbb{E}N_n$ is sometimes referred to as the marginal false discovery rate (m-FDR). Sun and Cai (2007) and Cai and Sun (2009) established optimality properties and m-FDR control of several local FDR based testing procedures, including the step-up procedure used here, under independence and monotonicity assumptions. However, these assumptions are typically violated in the setting of interest to us here. The monotonicity assumption, which in the present case involves the relationship between the distributions of the local FDR $\eta(\mathbf{Z}_\lambda)$ under $H_{0,\lambda}$ and $H_{1,\lambda}$, does not appear to hold. Moreover, in eQTL data there are typically significant correlations between nearby SNPs (linkage disequilibrium), leading to to complex, non-stationary correlations between the gene-SNP based vectors $\mathbf{Z}_\lambda$.

Theorem 3.2 makes no explicit assumptions on the joint distribution of the vectors $\mathbf{Z}_\lambda$; instead it relies on the relatively weak condition that $\hat{F}_n(t) \to F(t)$ in probability. This condition holds, for example, under the (very mild) assumption that the variance of the numerator and denominator of $\hat{F}_n(t)$ is equal to $o(|\Lambda_n|^2)$. While strong correlations between nearby SNPs will be present, gene-SNP pairs that are well separated will have little or no correlation, so the variance decay assumption is reasonable in practice. When the variance decay assumption holds, the conclusion of the theorem may be strengthened to $M_n/N_n = \alpha + o_P(1)$.

In regards to the proof of Theorem 3.2, the assumption that Δ be positive definite is only needed to ensure the existence of the the densities f_γ; the assumption that the diagonal entries of Σ are positive is reasonable in practice, but can likely be weakened. The proof makes use of the properties of the multivariate normal, specifically the normality of conditional distributions and the fact that normal densities are analytic functions, but could likely be extended to more general exponential families with additional work.

3.2 Analysis for Subsets of Tissues

In some problems, a subset $S \subseteq \{1, \ldots, K\}$ of the available tissues may be of primary interest. The multiple testing framework described above can be adapted to the tissues in S in two primary ways. The first is to construct a model based only on the tissues in S and use the resulting local FDR to identify multi-tissue eQTLs. However, this approach does not make use of the available data from tissues outside S and as such it does not borrow strength from commonalities among tissues. As an alternative, one may use the *marginal local FDR* for S, defined by

$$\eta_S(\mathbf{z}) \;:=\; \mathbb{P}(\mathbf{\Gamma}_S = \mathbf{0} \,|\, \mathbf{Z} = \mathbf{z}) \;=\; \frac{\sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}_S=\mathbf{0}} p_{\boldsymbol{\gamma}} f_{\boldsymbol{\gamma}}(\mathbf{z})}{f(\mathbf{z})}. \tag{14}$$

Here $\mathbf{\Gamma}_S$ and $\boldsymbol{\gamma}_S$ denote, respectively, the restriction of the vectors $\mathbf{\Gamma}$ and $\boldsymbol{\gamma}$ to the tissues in S, while $p_{\boldsymbol{\gamma}}$, $f_{\boldsymbol{\gamma}}$ and f correspond to the full model (6). We emphasize that the marginal local FDR $\eta_S(\mathbf{z})$ is a function of the complete vector of z-statistics, and therefore depends on the fitted model for the full set of tissues.

3.3 Assessments of Tissue Specificity

Testing gene-SNP pairs is typically the first step in multi-tissue eQTL analysis. Rejection of $\mathrm{H}_{0,\lambda}$ is based on evidence that λ is an eQTL in at least one of the available tissues. More detailed statements about the pattern of eQTLs across tissues can be made using information about the full configuration vector $\mathbf{\Gamma}_\lambda$. If the hypothesis $\mathrm{H}_{0,\lambda}$ is rejected, a natural estimate of $\mathbf{\Gamma}_\lambda$ is the maximum a-posteriori (MAP) configuration defined by

$$\widehat{\boldsymbol{\gamma}}_\lambda \;=\; \underset{\boldsymbol{\gamma}\in\{0,1\}^K\setminus\mathbf{0}}{\arg\max}\; p(\boldsymbol{\gamma} \,|\, \mathbf{z}_\lambda) \;=\; \underset{\boldsymbol{\gamma}\in\{0,1\}^K\setminus\mathbf{0}}{\arg\max}\; p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda).$$

The MAP rule is investigated in the simulation section below. As an alternative, one may compute the marginal posterior probability of an eQTL in each tissue k, namely

$$p(\mathbf{\Gamma}_{\lambda,k} = 1|\mathbf{z}_\lambda) \;=\; \sum_{\boldsymbol{\gamma}:\gamma_k=1} p(\boldsymbol{\gamma}|\mathbf{z}_\lambda) \;=\; \sum_{\boldsymbol{\gamma}:\gamma_k=1} p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda)/f(\mathbf{z}_\lambda),$$

and declare an eQTL in tissue k if this marginal probability exceeds a predefined threshold. Both MAP and thresholding of the marginal posterior extend to subsets of tissues.

3.4 Testing a Family Configurations

The goal of the multiple testing problem (10) is to determine whether the configuration $\mathbf{\Gamma}_\lambda$ of a gene-SNP pair is equal to $\mathbf{0}$ or belongs to the complementary set $\{0,1\}^K \setminus \{\mathbf{0}\}$. More generally, one may test membership of $\mathbf{\Gamma}_\lambda$ in any fixed subset $T \subseteq \{0,1\}^K$ of configurations. The associated testing problem can be written as

$$\mathrm{H}_{0,\lambda}^T : \mathbf{\Gamma}_\lambda \in T^c \;\text{ versus }\; \mathrm{H}_{1,\lambda}^T : \mathbf{\Gamma}_\lambda \in T, \qquad \lambda \in \Lambda. \tag{15}$$

A test statistic for (15) can be obtained by marginalizing the full local FDR (11), which yields

$$\eta_T(\mathbf{z}) \;:=\; \mathbb{P}(\mathbf{\Gamma} \in T^c \,|\, \mathbf{Z} = \mathbf{z}) \;=\; \frac{\sum_{\boldsymbol{\gamma}:\boldsymbol{\gamma}\in T^c} p_{\boldsymbol{\gamma}} f_{\boldsymbol{\gamma}}(\mathbf{z})}{f(\mathbf{z})}.$$

The local FDR step-up procedure can then be applied to the values $\{\eta_T(\mathbf{z}_\lambda)\}$ in order to control the overall FDR in (15).

4 Simulation Study

In this section, we illustrate MT-eQTL with a simulation study. The basis of our model and inference procedure is the collection of z-statistic vectors derived from the observed genotype and transcript data. Thus we directly simulate the z-statistic vectors themselves. Further details and results are described below.

4.1 Simulation Setting

We simulate vectors $\mathbf{z}_\lambda$ independently from the mixture model (6) using parameters $\theta = (\boldsymbol{\mu}_0, \Delta, \Sigma, \mathbf{p})$ that are obtained from an ongoing eQTL analysis of data with

$K = 4$ tissues, which we denote by a, b, c, and d, and roughly 10^7 local gene-SNP pairs. Sample sizes, sample overlap, and degrees of freedom after covariate correction are given in Table 1.

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

For simplicity, in the simulations and model fitting we set $\boldsymbol{\mu}_0$ to zero. The generating parameters Δ and Σ based on the real data analysis are as follows:

$$\Delta = \begin{pmatrix} 1.0000 & 0.1347 & 0.0805 & 0.1089 \\ 0.1347 & 1.0000 & 0.1204 & 0.1794 \\ 0.0805 & 0.1204 & 1.0000 & 0.1288 \\ 0.1089 & 0.1794 & 0.1288 & 1.0000 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 6.5699 & 5.3098 & 4.4683 & 4.7126 \\ 5.3098 & 5.9752 & 4.7906 & 5.5778 \\ 4.4683 & 4.7906 & 5.5263 & 4.6493 \\ 4.7126 & 5.5778 & 4.6493 & 6.0178 \end{pmatrix}.$$

The generating parameter $\mathbf{p}$ can be found in Table 3. We simulated $|\Lambda| = 10^7$ vectors $\mathbf{z}_\lambda$ from (6) in a two-step fashion: first drawing $\boldsymbol{\gamma} \in \{0,1\}^4$ from $\mathbf{p}$, and then drawing $\mathbf{z}_\lambda$ from $f_{\boldsymbol{\gamma}}(\mathbf{z})$. 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-tissue model, as well as all possible 1-, 2-, and 3-tissue models. We terminated EM updates when the difference between log likelihoods in two consecutive iterations was less than 0.01. The number of iterations until convergence of the EM procedure did not vary greatly from model to model. For all models, the average number of iterations is 80, with numbers varying from 40 to 132. The running time of the EM procedure depends on the number of tissues in the model, ranging from about 1 second per iteration for the 1-tissue models

to about 40 seconds per iteration for full 4-tissue model. Fitting of the 4-tissue model based on the simulated data took slightly more than one hour; fitting the real data (about 10^7 transcript-SNP pairs) required 1.5 hours. On real and simulated data, fitting of the mixture model is extremely fast.

As expected, the parameters estimated from the simulated data are very close to those used to generate the data. For the 4-tissue 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%. As for the probability mass vector $\mathbf{p}$, thirteen of sixteen entries had relative error less than 1%; the relative errors of the remaining entries were 1.45%, 1.66% and 4.31%. These results confirm that the approximate EM procedure works well on the simulated data.

In order to investigate the marginal consistency of the estimated parameters, we marginalized the estimated parameters from the 4-tissue model, and compared them with parameters estimated directly from lower dimensional models. The average relative differences of the parameters $\mathbf{p}$, Σ, and Δ for 1-tissue, 2-tissue, and 3-tissue models are summarized in Table 2.

	1-tissue	2-tissue	3-tissue
$\mathbf{p}$	9.8E-4	2.4E-3	4.1E-3
Σ	1.7E-3	1.1E-3	5.8E-4
Δ	0	1.4E-3	5.2E-4

Table 2: The average relative differences of $\mathbf{p}$, Σ, and Δ between the marginalization of the 4-tissue model and the direct estimations from all 1-, 2-, and 3-tissue models.

4.3 Results

We first applied the adaptive thresholding procedure to the full 4-tissue model to identify gene-SNP pairs that are eQTLs in at least one tissue. Throughout we used an overall FDR threshold $\alpha = 0.05$. For all models considered in the simulation, the true false discovery rates are always slightly below 0.05.

Table 3 shows results from the 4-tissue model with MAP estimates of the con-

figuration $\boldsymbol{\gamma}$. The "TS-config" column enumerates the 16 possible configurations according to the tissues in which eQTLs are present. The "True" column shows the true numbers of transcript-SNP pairs with the specified configuration in the simulated data. The "Discoveries" column shows the number of transcript-SNP pairs in the simulation estimated to have the specified configuration. The "Intersection" column shows cardinality of the intersection of true and discovered transcript-SNP pairs with the specified configuration. The "Proportion" column gives the proportion of true discoveries.

TS-config	100*$\mathbf{p}$	True	Discoveries	Intersection	Proportion
0	77.24	7720693	8961544	7669320	0.86
a	1.96	196868	52070	33128	0.64
b	1.04	103866	23786	17070	0.72
c	1.88	189859	45253	28738	0.64
d	2.05	202925	53716	37600	0.70
a-b	0.29	29516	4592	3035	0.66
a-c	0.08	7835	446	313	0.70
a-d	0.09	9507	1280	870	0.68
b-c	0.10	9552	1448	903	0.62
b-d	0.33	32552	5196	2997	0.58
c-d	0.37	36738	6382	4294	0.67
a-b-c	0.19	19022	1730	1258	0.73
a-b-d	0.86	85418	9115	6194	0.68
a-c-d	0.09	8614	951	731	0.77
b-c-d	1.08	107405	14031	9445	0.67
a-b-c-d	12.34	1239630	818460	640847	0.78

Table 3: eQTL analysis results from the 4-tissue model for the simulation data.

For each configuration, only a modest fraction (about 1/4) of the true eQTLs with that configuration are detected by the local FDR procedure. This does not imply that the local FDR procedure is under-powered, but instead reflects features of the data generation process that we believe are representative of real data. In detail, the multi-tissue z-statistics of each gene-SNP pair are generated from a mixture multivariate

Gaussian distribution centered at zero. As a result, the majority of alternative gene-SNP pairs have z-statistics near zero; these z-statistics are not readily distinguishable from those generated under the null.

For most configurations the proportion of true discoveries relative to total discoveries (the "Proportion" column) are above 60 percent. This is relatively high, given that distinguishing between nearby configurations (those with 1's in all but one of same positions) as well as the null configuration can be difficult.

Vectors of z-statistics can serve as a useful, and potentially powerful, tool for visualizing and interpreting the results of a multi-tissue eQTL analysis. Figure 1b shows scatter plots of z-statistics for tissue b and tissue d, while Figure 2 shows scatter plots of z-statistics for tissue a and tissue c. The black and white plot shows the density of the observed z-statistic vectors, while the companion plot shows the results of inference based on the fitted two-dimensional MT-eQTL model. In the companion plot, z-statistic vectors deemed not to be significant are omitted, leading to the white space at the center of the plot. The remaining points (corresponding to eQTLs) are colored according to their assessed tissue specificity: green represents the configuration $(1, 0)$ in which there is an eQTL in tissue 1 but not tissue 2; red represents the configuration $(0, 1)$ in which there is an eQTL in tissue 2 but not tissue 1; and blue represents the configuration $(1, 1)$ in which there is an eQTL in both tissues.

The overall shape of each plot is a tilted ellipse, with extreme values along the main diagonal and, to a lesser extent, along the coordinate axes. As expected, significant points close to one of the coordinate axes show evidence for an eQTL in a single tissue (tissue specific eQTL), while those along the positive diagonal show evidence for eQTLs in both tissues (common eQTL). In all other pairs of tissues (not shown), we observe similar results. In Figure 2, we also observe some discoveries along the anti-diagonal. For anti-diagonal pairs there is significant correlation between genotype and expression in each tissue, but the correlation is positive in one tissue, and negative in the other. The model reasonably identifies these points as common eQTLs. This behavior also appears in the analysis of real data, cf. Fu et al. (2012). Better statistical and biological understanding of eQTLs with different effects directions in different tissues is the subject of ongoing research.

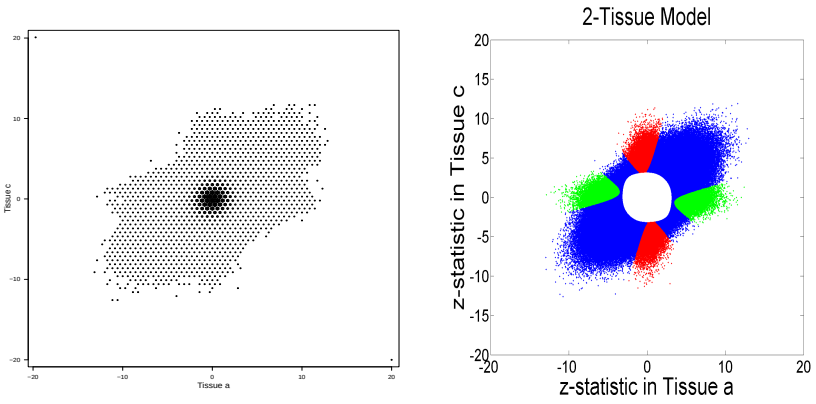


Figure 2: Scatter plots of z-statistics for tissue a and tissue c. Density-based scatter plot for all gene-SNP pairs (left), and significant eQTL discoveries with tissue specificity assessments from the fitted two-dimensional MT-eQTL model (right).

In order to assess how the use of multiple tissues increases statistical power in the context of the simulation, we fit models for tissue sets $\{a\}$, $\{a, b\}$, $\{a, b, c\}$, and $\{a, b, c, d\}$ and only focus on eQTL detection in tissue a. In each case we applied the adaptive thresholding procedure to the marginal local FDR defined in (14). Figure 3 shows that the number of discoveries for tissue a increases steadily with the number of auxiliary tissues. The realized false discovery rates were all controlled at the specified 0.05 level.

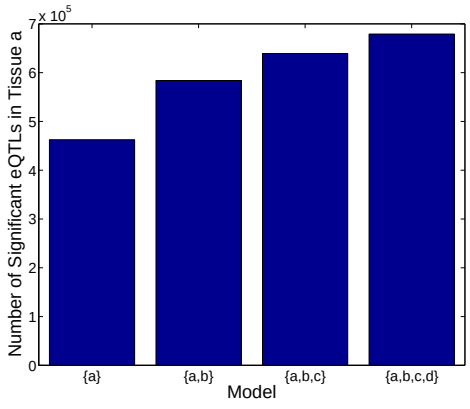


Figure 3: The number of significant discoveries in tissue a from the model for $\{a\}$, $\{a, b\}$, $\{a, b, c\}$, and $\{a, b, c, d\}$ respectively.

To better visualize gains in eQTL discoveries with the use of auxiliary tissues, we

also fit models for tissue sets $\{a, b\}$, $\{a, b, c\}$, and $\{a, b, c, d\}$, and detected eQTLs in tissue a and tissue b. Figure 4 shows the scatter plots of significant eQTL discoveries in the two tissues with tissue specificity assessments from different models. Clearly, borrowing strength from auxiliary tissues, higher dimensional models discover more eQTL gene-SNP pairs with moderate effects in tissue a and tissue b. The numbers of discoveries from the 2-tissue, 3-tissue, and 4-tissue models are 707151, 789951, and 863253 respectively. All realized FDRs are within .0015 of the nominal .05 threshold.

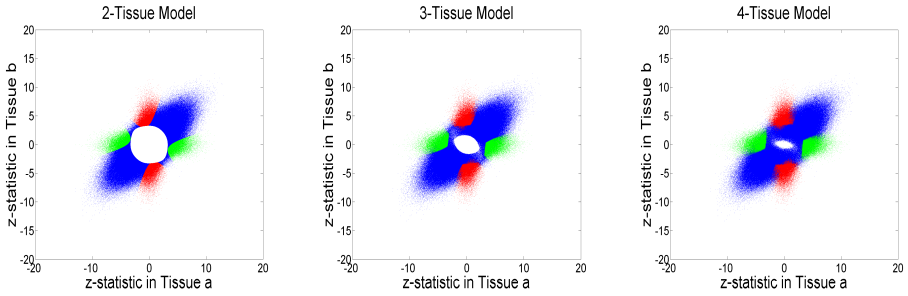


Figure 4: Scatter plots of significant discoveries in tissues a and tissue b from models for $\{a, b\}$, $\{a, b, c\}$, and $\{a, b, c, d\}$. Tissue specificities are assessed through the MAP rule.

Acknowledgements:

This research was supported by NIH Grants MH090936-01 and MH101819-01, and by NSF Grants DMS 0907177 and DMS 1310002.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 289–300.

Brown, C. D., Mangravite, L. M., and Engelhardt, B. E. (2013). Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genetics* **9,** e1003649.

Bullaughey, K., Chavarria, C. I., Coop, G., and Gilad, Y. (2009). Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Human Molecular Genetics* **18,** 4296–4303.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *Journal of the American Statistical Association* **104,**.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10,** 184–194.

Dawson, J. A. and Kendziorski, C. (2012). An empirical bayesian approach for identifying differential coexpression in high-throughput experiments. *Biometrics* **68,** 455–465.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Arcelus, M. G., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325,** 1246–1250.

Ding, J., Gudjonsson, J. E., Liang, L., Stuart, P. E., Li, Y., Chen, W., Weichenthal, M., Ellinghaus, E., Franke, A., Cookson, W., et al. (2010). Gene expression in skin and lymphoblastoid cells: refined statistical method reveals extensive overlap in cis-eQTL signals. *The American Journal of Human Genetics* **87,** 779–789.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics* **35,** 1351–1377.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* pages 1–22.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association* **96,** 1151–1160.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics* **9,** e1003486.

Franke, L. and Jansen, R. C. (2009). eqtl analysis in humans. In *Cardiovascular Genomics*, pages 311–328. Springer.

Fu, J., Wolfs, M. G., Deelen, P., Westra, H.-J., Fehrmann, R. S., te Meerman, G. J., Buurman, W. A., Rensen, S. S., Groen, H. J., Weersma, R. K., et al. (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genetics* **8,** e1002431.

Gerrits, A., Li, Y., Tesson, B. M., Bystrykh, L. V., Weersing, E., Ausema, A., Dontje, B., Wang, X., Breitling, R., Jansen, R. C., et al. (2009). Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genetics* **5,** e1000692.

Heinzen, E. L., Ge, D., Cronin, K. D., Maia, J. M., Shianna, K. V., Gabriel, W. N., Welsh-Bohmer, K. A., Hulette, C. M., Denny, T. N., and Goldstein, D. B. (2008). Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biology* **6,** e1000001.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22,** 3899–3914.

Kendziorski, C. and Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome* **17,** 509–517.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3,** e161.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45,** 580–585.

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10,** 565–577.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K.-W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational biology* **8,** 37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics* **7,** e1002003.

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., et al. (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Computational Biology* **6,** e1000737.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics* **7,** 862–872.

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

Smyth, G. K. et al. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3,** 3.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols* **7,** 500–507.

Sul, J. H., Han, B., Ye, C., Choi, T., and Eskin, E. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genetics* **9,** e1003491.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* **102,** 901–912.

Wen, X. and Stephens, M. (2011). Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *arXiv Preprint arXiv:1111.1210* .

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., et al. (2013). Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature genetics* **45,** 1238–1243.

Winterbottom, A. (1979). A note on the derivation of fisher's transformation of the correlation coefficient. *The American Statistician* **33,** 142–143.

Wright, F. A., Shabalin, A. A., and Rusyn, I. (2012). Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics* **13,** 343–352.

Xia, K., Shabalin, A. A., Huang, S., Madar, V., Zhou, Y.-H., Wang, W., Zou, F., Sun, W., Sullivan, P. F., and Wright, F. A. (2012). seeQTL: a searchable database for human eQTLs. *Bioinformatics* **28,** 451–452.

A Model Fitting and Parameter Estimation

A.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 which is defined in Section 2.2 and is also reported by Matrix eQTL. The set of correlations can then be combined in a single matrix with rows $\mathbf{r}_\lambda$.

A.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 $\{\mathbf{z}_\lambda\}$:

$$L(\{\mathbf{z}_\lambda\}|\theta) = \prod_{\lambda \in \Lambda} \sum_{\boldsymbol{\gamma} \in \{0,1\}^K} p_{\boldsymbol{\gamma}}\, f_{\boldsymbol{\gamma}}(\mathbf{z}_\lambda \,|\, \theta), \tag{16}$$

where $f_{\boldsymbol{\gamma}}(\cdot \,|\, \theta)$ is the probability density function of the $\mathcal{N}_K\big(\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma}, \Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T\big)$ 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 (16). 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.

B Proof of Lemma 2.1

Proof. Let S be a subset of $\{1, \ldots, K\}$ 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 $\mathbf{U}_S \sim \mathcal{N}_r(\boldsymbol{\mu}_S, A_S)$. It therefore follows from (6) that

$$\mathbf{Z}_S \;\sim \sum_{\boldsymbol{\gamma} \in \{0,1\}^K} p_{\boldsymbol{\gamma}} \, \mathcal{N}_r\big((\boldsymbol{\mu}_0 \cdot \boldsymbol{\gamma})_S, \, (\Delta + \Sigma \cdot \boldsymbol{\gamma}\boldsymbol{\gamma}^T)_S\big) \tag{17}$$

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 (17) 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 $\mathbf{Z}_S$. □

C Proof of Theorem 3.2

C.1 Continuity and Monotonicity of $F(t)$

Lemma C.1. *Let U be a bounded, non-negative random variable. For $t \geq 0$ define*

$$G(t) \;=\; \mathbb{E}[\,U \,|\, U \leq t\,] \;=\; \frac{\mathbb{E}[\,U\, \mathbb{I}(U \leq t)\,]}{\mathbb{P}(U \leq t)}. \tag{18}$$

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 $G(t+\delta) - G(t) \geq 0$ for each fixed $t \geq 0$ 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

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

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

$$\begin{aligned} &\mathbb{E}[\, U\, \mathbb{I}(t < U \leq t+\delta)\,]\, \mathbb{P}(U \leq t) \;-\; \mathbb{E}[\, U\, \mathbb{I}(U \leq t)\,]\, \mathbb{P}(t < U \leq t+\delta) \\ &\quad\geq\; t\, \mathbb{P}(t < U \leq t+\delta)\, \mathbb{P}(U \leq t) \;-\; t\, \mathbb{P}(U \leq t)\, \mathbb{P}(t < U \leq t+\delta) \qquad (19) \\ &\quad=\; 0. \end{aligned}$$

Thus G is non-decreasing. Right continuity of G follows by applying the monotone convergence theorem to the numerator and denominator in (18). 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 (19) is strict, and the final claim follows by considering $t \in [0, L)$ and $\delta > 0$ such that $t + \delta < L$. □

Lemma C.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 $m \geq 1$. 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 $\{(\sigma_i, \mu_i) : 0 \leq i \leq m\}$. 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

$$0 \;=\; f_0(x_1^{d-1} \,|\, x_d)\, f_0(x_d) \;-\; \sum_{j=1}^{m} c_j\, f_j(x_1^{d-1} \,|\, x_d)\, f_j(x_d) \qquad x \in A \tag{20}$$

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 (20) 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 (20) 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 $j \geq 1$, again contradicting the induction hypothesis. This completes the proof. □

Lemma C.3. *Let $\eta(\mathbf{z})$ be defined as in (11) 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 $c \geq 0$ 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

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

Let $1 \leq i \leq K$ 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 C.2

Proposition C.4. *The function $F(t)$ defined in (12) 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 (18) with $U = \eta(\mathbf{Z})$. Part 2 of Lemma C.3 establishes that $\mathbb{P}(\eta(bZ) = t) = 0$, and continuity of F then follows from Lemma C.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 C.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 C.1. □

C.2 Proof of Theorem 3.2

Lemma C.5. *Let $G_1, G_2, \ldots : [0,1] \to \mathbb{R}$ be non-decreasing functions. For fixed $\alpha \in (0, L_\eta)$ define $\theta_n = \sup\{t : G_n(t) \leq \alpha\}$ 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 $\theta_{n_k} \leq \theta - 2\delta_1$ for each $k \geq 1$ or $\theta_{n_k} \geq \theta + 2\delta_2$ for each $k \geq 1$. In the first case, the definition of θ_n and the monotonicity of G_n imply

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

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

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

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

Proof of Theorem 3.2:

Proof. Let $\hat{\theta}_n = \sup\{t : \hat{F}_n(t) \leq \alpha\}$ 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

$$\mathbb{P}\big(|\hat{\theta}_{n_k} - \theta| > \delta\big) > \delta \quad \text{for each} \quad k \geq 1. \tag{21}$$

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 C.5 that $\hat{\theta}_{m_k} \to \theta$ with probability one, which contradicts (21).

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

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

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

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

For each $\delta > 0$,

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

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,

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

This completes the proof of the theorem. □