

METHODOLOGY ARTICLE

Open Access



HT-eQTL: integrative expression quantitative trait loci analysis in a large number of human tissues

Gen Li^{1*} , Dereje Jima², Fred A. Wright^{2,3} and Andrew B. Nobel⁴

Abstract

Background: Expression quantitative trait loci (eQTL) analysis identifies genetic markers associated with the expression of a gene. Most existing eQTL analyses and methods investigate association in a single, readily available tissue, such as blood. Joint analysis of eQTL in multiple tissues has the potential to improve, and expand the scope of, single-tissue analyses. Large-scale collaborative efforts such as the Genotype-Tissue Expression (GTEx) program are currently generating high quality data in a large number of tissues. However, computational constraints limit genome-wide multi-tissue eQTL analysis.

Results: We develop an integrative method under a hierarchical Bayesian framework for eQTL analysis in a large number of tissues. The model fitting procedure is highly scalable, and the computing time is a polynomial function of the number of tissues. Multi-tissue eQTLs are identified through a local false discovery rate approach, which rigorously controls the false discovery rate. Using simulation and GTEx real data studies, we show that the proposed method has superior performance to existing methods in terms of computing time and the power of eQTL discovery.

Conclusions: We provide a scalable method for eQTL analysis in a large number of tissues. The method enables the identification of eQTL with different configurations and facilitates the characterization of tissue specificity.

Keywords: Expression quantitative trait loci, Genotype-tissue expression project, Empirical Bayes, Tissue specific, Local false discovery rate

Background

Expression quantitative trait loci (eQTL) analyses identify single nucleotide polymorphisms (SNPs) that are associated with the expression level of a gene. A gene-SNP pair such that the expression of the gene is associated with the value of the SNP is referred to as an eQTL. One may view eQTL analyses as Genome-Wide Association Studies (GWAS) with multiple molecular phenotypes. Identification of eQTLs is a key step in investigating genetic regulatory pathways. To date, numerous eQTLs have been discovered to be associated with human traits such as height and complex diseases such as Alzheimer's disease and diabetes [1, 2].

With few exceptions, existing eQTL studies have focused on a single tissue; in human studies this tissue is usually blood. An important next step in exploring the genomic regulation of expression is to simultaneously study eQTLs in multiple tissues. Multi-tissue eQTL analysis can strengthen the conclusions of single tissue analyses by borrowing strength across tissues, and can help provide insight into the genomic basis of differences between tissues, as well as the genetic mechanisms of tissue-specific diseases.

Recently, the NIH Common Fund's Genotype-Tissue Expression (GTEx) project has undertaken a large-scale effort to collect and analyze eQTL data in multiple tissues on a growing set of human subjects, and there has been a concomitant development of methods for the analysis of such data. For example, Peterson et al. [3] and Bogomolov et al. [4] developed new error control procedures to control false discovery rates at different

*Correspondence: gl2521@cumc.columbia.edu

¹Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W 168 Street, New York, USA

Full list of author information is available at the end of the article

levels of resolution (e.g., at the SNP level or the gene level) for eQTL analysis. The methods have been used to identify genes whose expression is regulated by SNPs (eGenes), or SNPs that affect the expression levels of multiple genes (eSNPs). However, the methods only concern how to reduce the number of hypotheses in a hierarchical structure, but cannot effectively borrow strength across tissues to enhance eQTL discoveries. Lewin et al. [5], Sul et al. [6] and Han et al. [7] developed regression-based methods via Bayesian multivariate regression and random-effects models. The models accommodate data from multiple tissues simultaneously, and integrate information across tissues for eQTL detection. However, a potential drawback is that they only focus on one gene or gene-SNP pair at a time, and fail to leverage information across different gene-SNP pairs. Flutre et al. [8] and Li et al. [9] developed hierarchical Bayesian models to model summary statistics across multiple tissues. The models capture the marginal distribution of each gene-SNP pair with interpretable parameters, and explicitly characterize heterogeneous eQTL configurations in multiple tissues. However, the model fitting is computationally expensive and cannot scale to a large number of tissues. Recently, Urbut et al. [10] proposed an ad hoc approach based on shrinkage to improve the scalability of the Bayesian models. However, the procedure is subject to overfitting and the model parameters are hard to interpret. Initial analyses and conclusions of the GTEx project are described in [11]. As part of this work, the “Bayesian Model Averaging” method [8] and the MT-eQTL (“MT” stands for multi-tissue) method [9] were applied to 9 human tissues with sample size greater than 80, focusing on local (cis) pairs for which the SNP is within one mega-base (Mb) of the transcription start site (TSS) of the gene. The analysis found that most eQTLs discovered were common across the 9 tissues included in the study, though the effect size may vary from tissue to tissue. In addition, there are a small, but potentially interesting, set of eQTLs that are present only in a subset of tissues, the most common cases being eQTLs that are present in only one tissue, or present in all but one tissue.

As GTEx and related projects proceed, data are being collected from an increasing number of subjects, and an increasing number of tissues. In the current GTEx database (v6p), more than 20 tissues have a sample size greater than 150. Existing eQTL analysis methods that can effectively borrow strength across tissues are limited in their ability to perform simultaneous local eQTL analyses in a large number of tissues. Methods like [8] and [9] incorporate and rely on a binary configuration vector, with dimension equal to the number of available tissues, that describes, for each gene-SNP pair, the presence or absence of association in each tissue. The total

number of possible configurations grows exponentially in the number of tissues, making computation, numerical accuracy, and memory management problematic when dealing with large numbers of tissues.

In this paper, we develop an efficient computational tool, called HT-eQTL (“HT” stands for high-tissue), for joint eQTL analysis. The method builds on the hierarchical Bayesian model developed in [9], but the estimation procedure is significantly improved to address scaling issue associated with a large number of tissues. Rather than fitting a full model, HT-eQTL fits models for all pairs of tissues in a parallel fashion, and then synthesizes the resulting pairwise models into a higher order model for all tissues. To do this, we exploit the marginal compatibility of the hierarchical Bayesian model, which is not an obvious property and was proven in [9]. An important innovation is that we employ a multi-Probit model and thresholding to deal with the exponentially growing configuration space. The resulting model and fitting procedure can be efficiently applied to the simultaneous eQTL analysis of 20-25 tissues. Empirical Bayesian methods for controlling false discovery rates in multiple hypothesis testing are developed. We design testing procedures to detect different families of eQTL configurations. We show that the eQTL detection power of HT-eQTL is similar to that of MT-eQTL, and that both outperform the tissue-by-tissue approach, in a simulation study with a moderate number of tissues. We also compare HT-eQTL with the Meta-Tissue method in the analysis of the GTEx v6p data. This analysis shows that the methods have largely concordant results, but that HT-eQTL gains additional power by borrowing strength across tissues.

Methods

In this section we describe the HT-eQTL method, beginning with a review of the hierarchical Bayesian model and the MT-eQTL method in [9], and then describing our proposal on how to fit the Bayesian model in high-tissue settings. Next, we describe a local false discovery rate based method for performing flexible eQTL inference. Finally, we discuss a marginal test and a marginal transformation to check and improve the goodness of fit of the model.

Review: Bayesian hierarchical model and MT-eQTL procedure

Consider a study with n subjects and K tissues. From each subject we have genotype data and measurements of gene expression in a subset of tissues. In many cases, covariate correction will be performed prior to analysis of eQTLs. For $k = 1, \dots, K$, let $n_k \leq n$ denote the number of subjects contributing expression data from tissue k . Let $\lambda = (i, j)$ be the index of a gene-SNP pair consisting of gene i and SNP j , and let Λ be the set of all local (cis)

gene-SNP pairs. For $\lambda = (i, j) \in \Lambda$ and $k = 1, \dots, K$, let $r_\lambda(k)$ denote the (covariate corrected) sample correlation between the expression level of gene i and the number of copies of the minor allele of SNP j in tissue k , and $\rho_\lambda(k)$ be the corresponding population correlation. Define $\mathbf{r}_\lambda = (r_\lambda(1), \dots, r_\lambda(K))$ to be the vector of sample correlations across tissues, and define the vector ρ_λ of population correlations in the same fashion.

Let $\mathbf{Z}_\lambda = \mathbf{h}(\mathbf{r}_\lambda) \cdot \mathbf{d}^{1/2}$, where $\mathbf{h}(\cdot)$ is the entrywise Fisher transformation with the effect of variance stabilization, \cdot is the Hadamard product, and \mathbf{d} is a K -vector whose k th component is the number of samples in tissue k minus the number of covariates removed from tissue k minus 3. With proper preprocessing of the gene expression data, the vector \mathbf{Z}_λ is approximately multivariate normal [12] with mean $\mu_\lambda = \mathbf{h}(\rho_\lambda) \cdot \mathbf{d}^{1/2}$ and marginal variance one. In particular, if $\rho_\lambda(k) = 0$ then the k th component of \mathbf{Z}_λ has a standard normal distribution, and can therefore be used as a z-statistic for testing $\rho_\lambda(k) = 0$ vs $\rho_\lambda(k) \neq 0$. Thus we refer to \mathbf{Z}_λ as a z-statistic vector.

The MT-eQTL model introduced in [9] is a Bayesian hierarchical model for the random vector \mathbf{Z}_λ . The model can be expressed in the form of a mixture as

$$\mathbf{Z}_\lambda \sim \sum_{\gamma \in \{0,1\}^K} p(\gamma) \mathcal{N}_K(\boldsymbol{\mu} \cdot \boldsymbol{\gamma}, \boldsymbol{\Delta} + \boldsymbol{\Sigma} \cdot \boldsymbol{\gamma} \boldsymbol{\gamma}'). \quad (1)$$

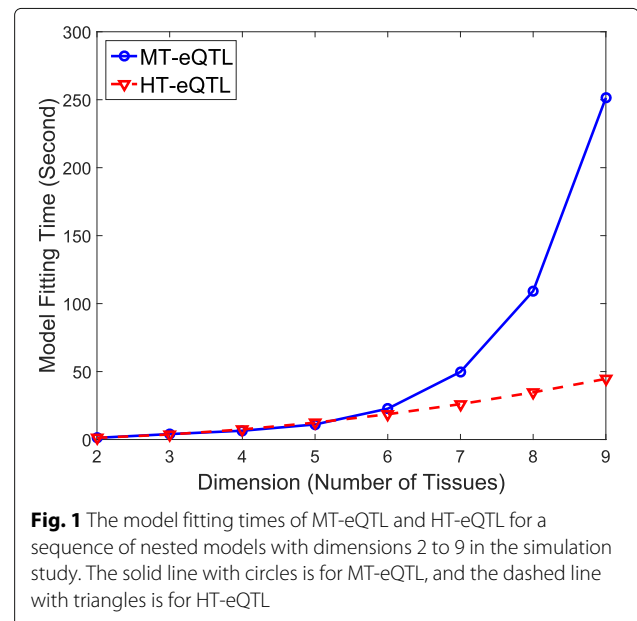
The mixture in (1) is taken over the set $\{0,1\}^K$ of length K binary vectors. Each vector $\boldsymbol{\gamma} \in \{0,1\}^K$ represents a particular configuration of eQTLs across the K available tissues: $\gamma_k = 1$ if the gene-SNP pair indexed by λ is an eQTL in tissue k , and $\gamma_k = 0$ otherwise. We define Hamming class m ($m = 0, \dots, K$) as the set of all binary K -vectors having m ones, which correspond to all configurations in which there is an eQTL in m tissues and no eQTL in $K - m$ tissues. The parameter $p(\boldsymbol{\gamma})$ is the prior probability of the configuration $\boldsymbol{\gamma}$. We collect all the priors in a length- 2^K vector \mathbf{p} . The K -vector $\boldsymbol{\mu}$ characterizes the average true effect size of eQTLs in each tissue. The $K \times K$ correlation matrix $\boldsymbol{\Delta}$ captures the behavior of \mathbf{Z}_λ when no eQTLs are present ($\boldsymbol{\gamma} = \mathbf{0}$): its diagonal entries are 1 due to the variance stabilization caused by the Fisher transformation, and its off-diagonal entries reflect correlations arising from subject overlap between tissues. The $K \times K$ matrix $\boldsymbol{\Sigma}$ captures the covariance structure of non-zero eQTL effect sizes in different tissues. Let $\theta = \{\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}\}$ denote the set of unknown model parameters.

Under the model (1) the distribution of \mathbf{Z}_λ is a normal mixture with each component corresponding to a specific eQTL configuration. In particular, if $\boldsymbol{\gamma} = \mathbf{0}$ (λ is not an eQTL in any tissue) then $\mathbf{Z}_\lambda \sim \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Delta})$; if $\boldsymbol{\gamma} = \mathbf{1}$ (λ is an eQTL in all tissues) then $\mathbf{Z}_\lambda \sim \mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Delta} + \boldsymbol{\Sigma})$. The true configuration vector for each gene-SNP pair λ can

be viewed as a latent variable. The main goal of a statistical analysis is to obtain the posterior distribution of each latent variable, and to use it to make inferences about eQTL configurations in multiple tissues.

In order to make inference about configuration vectors, we first estimate the model parameters $\theta = \{\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}\}$. In practice it is common to set the average effect size vector $\boldsymbol{\mu}$ to $\mathbf{0}$, as minor alleles are equally likely to be associated with high or low expression, and we assume in what follows that $\boldsymbol{\mu} = \mathbf{0}$. The remaining parameters can be estimated within a maximum pseudo-likelihood framework, where the pseudo-likelihood is defined as the product of the likelihoods of all considered gene-SNP pairs. We note that factorizing the likelihood in this way ignores dependence between adjacent and nearby SNPs arising from linkage disequilibrium. However, our interest is not in the joint behavior of the vectors \mathbf{Z}_λ but in their *marginal* behavior, which is reflected in the mixture (1). In particular, the parameters in Model (1) determine, and are determined by, the marginal distribution of the vectors \mathbf{Z}_λ , and do not depend on joint distribution of the vectors \mathbf{Z}_λ .

A modified EM algorithm was devised in [9] to estimate the parameters from the pseudo-likelihood (see Section A of the Additional file 1). While the method scales linearly with sample size and the number of gene-SNP pairs, its computational time increases exponentially with the number of tissues K (see Fig. 1). For genome-wide studies, it is infeasible to apply the method to data with more than a few tissues. Moreover, the number of configurations grows exponentially with the number of tissues as well, making inference about configurations difficult as



well. Below we introduce a scalable procedure, the HT-eQTL method, to address multi-tissue eQTL analysis in about 20 tissues.

The HT-eQTL method

The original MT-eQTL model has the desirable property of being *marginally compatible*. Let the dimension of the MT-eQTL model be the number of available tissues. Marginal compatibility means that: 1) the marginalization of a K -dimensional model to a subset of L tissues has the same general form as the K -dimensional model; and 2) the corresponding parameters for the L -dimensional model are obtained in the obvious way by restricting the parameters of the K -dimensional model to the subset of L tissues.

Because of marginal compatibility, it is straightforward to obtain a sub-model from a high dimensional model without refitting the MT-eQTL parameters. The HT-eQTL method, which is discussed below, estimates the high dimensional model from the collection of its one- and two-dimensional sub-models. Thus we address the computationally intractable problem of estimating a high dimensional model by considering a manageable number of sub-problems that can be solved efficiently, and in parallel.

In the MT-eQTL model, the covariance matrices Δ and Σ reflect interactions between pairs of tissues, while the probability mass function $p(\cdot)$ captures higher order relationships between tissues. The HT-eQTL model is built from estimates of all one- and two-dimensional sub-models, which can be computed in parallel. In particular, we make use of a Multi-Probit model to approximate the K -th order probability mass function $p(\cdot)$ from the probability mass functions of two-dimensional models. In what follows we denote the estimated parameters of the two-dimensional model for tissue pair (i, j) by

$$p^{ij} = (p_{00}^{ij}, p_{01}^{ij}, p_{10}^{ij}, p_{11}^{ij}), \Delta^{ij} = \begin{pmatrix} 1 & \delta^{ij} \\ \delta^{ij} & 1 \end{pmatrix}, \Sigma^{ij} = \begin{pmatrix} \sigma_{11}^{ij} & \sigma_{12}^{ij} \\ \sigma_{21}^{ij} & \sigma_{22}^{ij} \end{pmatrix}.$$

A description of the two-tissue model fitting procedure can be found in Section A of the Additional file 1.

Assemble Δ : For each tissue pair (i, j) where $1 \leq i < j \leq K$, the corresponding off-diagonal value of Δ is denoted by δ_{ij} . An asymptotically consistent estimate of δ_{ij} is the off-diagonal value of Δ^{ij} , which is the null covariance matrix for the two-dimensional model for tissue pair (i, j) . Making this substitution for each $i < j$ and placing ones along the diagonal yields the proposed estimate of Δ (i.e., $\hat{\Delta}$). In practice, since each Δ^{ij} is typically estimated from a large number of gene-SNP pairs, $\hat{\Delta}$ is very close to Δ with negligible variability. If $\hat{\Delta}$ is not positive definite (which did not occur in our numerical studies), we set the negative eigenvalues of $\hat{\Delta}$ by 0, and rescale it to be a correlation matrix.

Assemble Σ : To estimate the covariance matrix $\Sigma = \{\sigma_{ij}\}$, we decompose it into the diagonal values, which are tissue-specific variances, and the corresponding correlation matrix. For each diagonal entry σ_{kk} ($k = 1, \dots, K$), there are $K - 1$ estimates, namely $\sigma_{22}^{1k}, \dots, \sigma_{22}^{(k-1)k}, \sigma_{11}^{k(k+1)}, \dots, \sigma_{11}^{kK}$. In practice, the distribution of z-statistics is usually heavy-tailed, inflating the pairwise estimates of the variance. As a remedy, we propose to use the minimum of the $K - 1$ estimates as the estimate of σ_{kk} to compensate the inflation effect. The induced correlation matrix from Σ is estimated in the same way as Δ . In particular, we start with a matrix having ones along the diagonal and off-diagonal entries $\sigma_{12}^{ij} / \sqrt{\sigma_{11}^{ij} \sigma_{22}^{ij}}$. We then obtain the closest positive semi-definite matrix by setting negative eigenvalues to zero, and rescale the resulting matrix to be a correlation matrix. Combining the correlation matrix with the diagonal variance terms, we obtain the estimate $\hat{\Sigma}$.

The Multi-Probit Model for \mathbf{p} : Existing multi-tissue eQTL studies [9, 11] support several broad conclusions about eQTL configurations across tissues. Researchers found that most gene-SNP pairs are not an eQTL in *any* tissue (Hamming class 0) or were an eQTL in *all* tissues (Hamming class K). With larger sample sizes and a larger number of tissues (thus providing increased power to detect cross-tissue sharing), we expect these two Hamming classes to predominate.

In general, the probability mass functions obtained from two-dimensional models will not determine a unique probability mass function on the full K -dimensional model. Here we make use of a multi-Probit model through which we equate the values of the estimated probability mass function with integrals of a multivariate normal probability density. In particular, for each tissue pair (i, j) , we select thresholds $\tau_1^{ij}, \tau_2^{ij} \in \mathbb{R}$ and a correlation $\omega^{ij} \in (0, 1)$ so that if (W_i, W_j) are bivariate normal with mean zero, variance one, and correlation ω^{ij} then

$$\Pr \left\{ \mathbb{I}(W_i \geq \tau_1^{ij}) = u \text{ and } \mathbb{I}(W_j \geq \tau_2^{ij}) = v \right\} = p^{ij}(u, v)$$

for each $u, v \in \{0, 1\}$. Here $\mathbb{I}(A)$ is the indicator function of A , and $p^{ij}(\cdot)$ is the estimated probability mass function for the pair (i, j) .

Beginning with a symmetric matrix having diagonal values 1 and off-diagonal values equal to ω^{ij} , we define a correlation matrix Ω following the procedure used to define $\hat{\Delta}$. Let $\phi_K(\cdot)$ be the probability density function of the corresponding K -variate normal distribution $\mathcal{N}_K(\mathbf{0}, \Omega)$. For each tissue j , we define an aggregate threshold τ^j to be the minimum of τ_1^{ij} ($i < j$) and τ_2^{ji} ($j < i$). Here we use the minimum because pairwise models may occasionally overestimate the null prior probability $p^{ij}(0, 0)$. Subsequently, for each configuration $\gamma \in \{0, 1\}^K$, we define the probability

$$p(\gamma) = \int_{I_1} \cdots \int_{I_K} \phi_K(\mathbf{x}) d\mathbf{x}$$

where I_k is equal to $(-\infty, \tau^k]$ if $\gamma_k = 0$, and (τ^k, ∞) , if $\gamma_k = 1$. Consequently, we obtain the estimate of probability mass function \mathbf{p} for the K -dimensional model.

Threshold $p(\cdot)$: In practice, many of the 2^K possible configurations will have estimated probabilities close to zero. In order to further reduce the number of configurations, we set the threshold for the prior probabilities to be 10^{-5} , and truncate those values below the threshold to be zero. The remaining probabilities are rescaled to have total mass one. As a result, the total number of configurations with non-zero probabilities is dramatically reduced to a manageable level for subsequent inferences.

Inferences

The first, and often primary, goal of eQTL analysis in multiple tissues is to detect which gene-SNP pairs are an eQTL in some tissue. Subsequent testing may seek to identify gene-SNP pairs that are an eQTL in a specific tissue, and pairs that are an eQTL in some, but not all, tissues. As the model (1) is fit with large number of gene-SNP pairs, we ignore the estimation error associated with the model parameters and treat the estimated values as fixed and true for the purposes of subsequent inference.

The mixture model (1) may be expressed in an equivalent, hierarchical form, in which for each gene-SNP pair λ , there is a latent random vector $\Gamma_\lambda \in \Gamma$ indicating whether or not that pair is an eQTL in each of the K tissues. The prior distribution of Γ_λ is characterized by the probabilistic mass function $p(\cdot)$. In the hierarchical model, given that $\Gamma_\lambda = \gamma$, the random z-statistic vector \mathbf{Z}_λ has distribution $\mathcal{N}_K(\mathbf{0}, \mathbf{\Delta} + \mathbf{\Sigma} \cdot \gamma\gamma')$. The posterior distribution of Γ_λ given the observed vector \mathbf{z}_λ can be used to test eQTL configurations for the gene-SNP pair λ .

Detection of eQTLs with specified configurations can be formulated as a multiple testing problem, and addressed through the use of local false discovery rates derived from the posterior distribution of gene-SNP pairs. Suppose that we are interested in identifying gene-SNP pairs with eQTL configurations in a set $S \subseteq \{0, 1\}^K$. This can be cast as a multiple testing problem

$$H_{0,\lambda} : \Gamma_\lambda \in S^c \text{ versus } H_{1,\lambda} : \Gamma_\lambda \in S$$

where $\lambda \in \Lambda$. Rejecting the null hypothesis for a gene-SNP pair λ indicates that λ is likely to have an eQTL configuration in S . There are several families S of particular interest, corresponding to different configurations of interest:

- Testing for the presence of an eQTL in any tissue:
 $S = \{\gamma : \gamma \neq \mathbf{0}\}$
- Testing for presence of a tissue-specific eQTL, i.e., an eQTL in some, but not all, tissues:
 $S = \{\gamma : \gamma \neq \mathbf{0}, \gamma \neq \mathbf{1}\}$

- Testing for presence of an eQTL in tissue k only:
 $S = \{\gamma : \gamma_k = 1\}$
- Testing for presence of a common eQTL, i.e., an eQTL in all tissues: $S = \{\mathbf{1}\}$.

To carry out multiple testing under the hierarchical Bayesian model, we make use of the local false discovery rate (lfdr) for the set S , which is defined as the posterior probability that the configuration Γ lies in S^c given the observed z-statistics vector \mathbf{z} . The local false discovery rate was introduced by [13] in the context of an empirical Bayes analysis of differential expression in microarrays. Other applications can be found in [14–16]. Formally, the lfdr for $S \subseteq \{0, 1\}^K$ is defined by

$$\eta_S(\mathbf{z}) := \Pr(\Gamma \in S^c \mid \mathbf{Z} = \mathbf{z}) = \frac{\sum_{\gamma \in S^c} p(\gamma) f_\gamma(\mathbf{z})}{\sum_{\gamma \in \{0,1\}^K} p(\gamma) f_\gamma(\mathbf{z})}, \quad (2)$$

where $f_\gamma(\mathbf{z})$ is the pdf of $\mathcal{N}_K(\mathbf{0}, \mathbf{\Delta} + \mathbf{\Sigma} \cdot \gamma\gamma')$. Thus $\eta_S(\mathbf{z}_\lambda)$ is the probability of the null hypothesis given the z-statistic vector for the gene-SNP pair λ . Small values of the lfdr provide evidence for the alternative hypothesis $H_{1,\gamma}$. In order to control the overall false discovery rate (FDR) for the multiple testing problem across all gene-SNP pairs $\lambda \in \Lambda$ we employ an adaptive thresholding procedure for local false discovery rates [9, 13, 14, 17]. For a given set of configurations S , and a given false discovery rate threshold $\alpha \in (0, 1)$, the procedure operates as follows.

- Calculate the lfdr $\eta_S(\mathbf{z}_\lambda)$ for each $\lambda \in \Lambda$.
- Sort the lfdRs from smallest to largest as $\eta_s(\lambda_{(1)}) \leq \cdots \leq \eta_s(\lambda_{(N)})$.
- Let N be the largest integer such that

$$\frac{1}{N} \sum_{i=1}^N \eta_s(\lambda_{(i)}) < \alpha.$$

- Reject hypotheses $H_{0,\lambda_{(i)}}$ for $i = 1, \dots, N$.

It is shown in [9] that the adaptive procedure controls the FDR at level α under very mild conditions. Consequently, we obtain a set of discoveries with FDR below the nominal level α .

Results

In the first part of this section, we conduct a simulation study with 9 tissues. We compare HT-eQTL with the MT-eQTL [9], Meta-Tissue [6] and tissue-by-tissue (TBT) [18–21] methods on different eQTL detection problems. The Meta-Tissue approach leverages the fixed effects and random effects method to address effect size heterogeneity and detect eQTLs across multiple tissues. The TBT approach first evaluates the significance of gene-SNP association in each tissue separately, and then aggregates the information across tissues. We also compare HT-eQTL and MT-eQTL in terms of the model fitting times

and parameter estimation accuracy. Then we apply the two scalable methods, HT-eQTL and Meta-Tissue, to the GTEx v6p data with 20 tissues.

Simulation

In the simulation study, we first generate z-statistics directly from Model (1) with $K = 9$ tissues. We fix the model parameters $\{\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Sigma}\}$ to be the ones estimated from MT-eQTL method on the GTEx pilot data. In particular, the parameters are available from the supplementary material of [9]. For each gene-SNP pair, we first randomly generate a length- K binary configuration vector γ based on the prior probability mass function \mathbf{p} . Given γ , the marginal distribution of the z-statistics is $\mathcal{N}(\boldsymbol{\mu} \cdot \gamma, \boldsymbol{\Delta} + \boldsymbol{\Gamma} \cdot \gamma \gamma')$. Then we simulate a length- K effect size vector from the multivariate Gaussian distribution. We repeat the procedure 10^5 times to obtain the true eQTL configurations and corresponding z-statistic vectors in 10^5 gene-SNP pairs. The true eQTL configurations under the simulation are used to evaluate the efficacy of different methods.

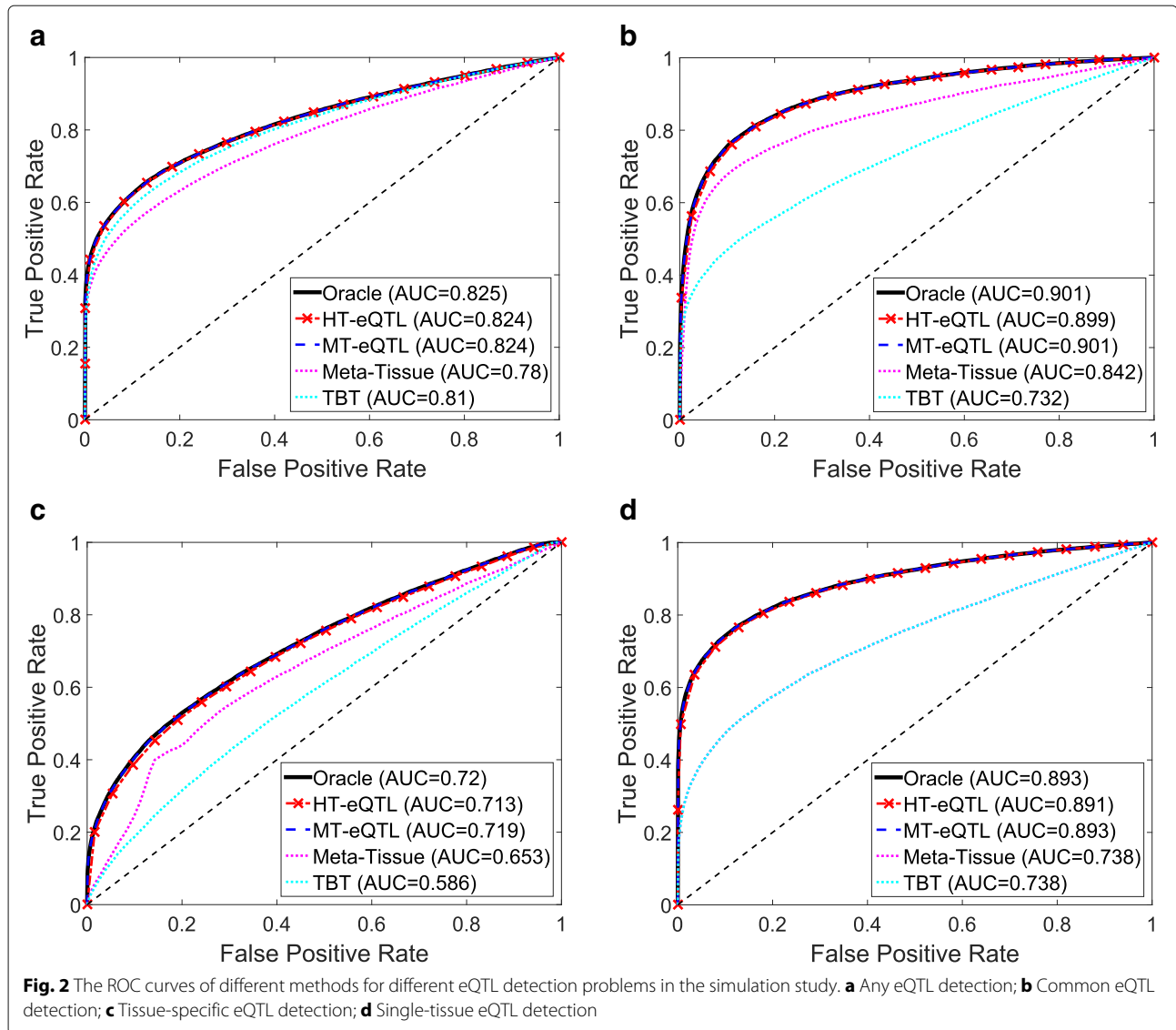
We first compare the computational costs of the MT-eQTL model fitting and the HT-eQTL model fitting (without parallelization). We consider a sequence of nested models with dimensions from 2 to 9. The model fitting times on the simulated data are shown in Fig. 1. We demonstrate that the model fitting time for the MT-eQTL grows exponentially in the number of tissues, while it grows much slower for the HT-eQTL. Namely, the HT-eQTL scales better than the MT-eQTL. This is because the HT-eQTL model fitting only involves the fitting of all the 2-tissue MT-eQTL models and a small overhead induced by assembling the pairwise parameters. When the total number of gene-SNP pairs and the number of tissues are large, the advantage of HT-eQTL is significant. Based on the timing results for MT-eQTL on the 9-tissue GTEx pilot data in [9], we project its fitting time to be more than 30 CPU years on 20 tissues. As we describe later, fitting the HT-eQTL model on the 20-tissue GTEx v6p data only takes less than 3 CPU hours. We remark that the straightforward parallelization of the 2-tissue MT-eQTL model fittings will further reduce the computational cost for HT-eQTL.

Now we compare the parameter estimation from MT-eQTL and HT-eQTL. We particularly focus on the 9-tissue model. The HT-eQTL parameters are obtained by fitting all 2-tissue models and assembling the pairwise parameters as described above. The MT-eQTL parameters are obtained directly by fitting a 9-tissue MT-eQTL model. Regarding the estimation of the correlation matrix $\boldsymbol{\Delta}$, the quartiles of the entry-wise relative errors are (0.86, 2.42–4.36%) and (0.81, 2.00–2.72%) for HT-eQTL and MT-eQTL, respectively. Regarding the estimation of the covariance matrix $\boldsymbol{\Sigma}$, the quartiles of

the entry-wise relative errors are (1.13, 2.41–3.25%) and (0.36, 0.68–1.08%) for HT-eQTL and MT-eQTL, respectively. Namely, although HT-eQTL had larger relative errors than MT-eQTL, both methods estimated the covariance matrices very accurately. For the probability mass vector \mathbf{p} , we calculated the Kullback-Liebler divergence of different estimates from the truth, defined as $D_{KL}(\mathbf{p} \parallel \hat{\mathbf{p}}) = \sum_{i=1}^{2^K} p_i \log(p_i / \hat{p}_i)$. The MT-eQTL estimate has a very small divergence of 0.025 while the HT-eQTL estimate has a slightly larger divergence of 0.141. Overall, the HT-eQTL estimates are slightly less accurate than the MT-eQTL estimates, which is expected because the HT-eQTL method has fewer degrees of freedom than the MT-eQTL method. When there are abundant data relative to the number of parameters, the more complicated MT-eQTL model will result in more accurate estimation. Nevertheless, we emphasize that the HT-eQTL estimates are sufficiently accurate for the eQTL detection purposes (see Fig. 2).

Next, we compare the eQTL detection power of different methods. We particularly focus on the detection of four types of eQTLs: (a) eQTLs in at least one tissue (Any eQTL); (b) eQTLs in all tissues (Common eQTL); (c) eQTLs in at least one tissue but not all tissues (Tissue-Specific eQTL); (d) eQTLs in a single tissue (Single-Tissue eQTL). In addition to the MT-eQTL and HT-eQTL methods, we also consider the Meta-Tissue and TBT approaches. In order to detect Any eQTL, we exploit the random effects model in Meta-Tissue and a minP procedure in TBT, where the minimum p value across tissues is used as the test statistics for each gene-SNP pair. To detect Common eQTL, we use the fixed effects model in Meta-Tissue and a maxP procedure in TBT, where the maximum p values across tissues are used. To detect Tissue-Specific eQTL, we devise a diffP procedure for TBT, where the test statistics for each gene-SNP pair is the difference between the maximum and the minimum p values across tissues. A large value indicates the discrepancy between the two extreme p values is large, and thus provides a strong evidence for the gene-SNP pair to be a tissue-specific eQTL. Similarly, for Meta-Tissue, we exploit the difference of p values from the fixed effects model and the random effects model as the test statistics. Finally, for Single-Tissue eQTL detection, Meta-Tissue reduces to the TBT method. We just use the p values in the primary tissue and ignore those in other tissues. For the MT-eQTL and HT-eQTL methods, we adapt the lfr test statistics in (2) to different testing problems accordingly.

We evaluate the performance of different methods using the Receiver Operating Characteristic (ROC) curves for different eQTL detection problems. The results are shown in Fig. 2. In particular, in panel (a), a gene-SNP pair identified by a method is deemed as a true positive if it truly has an eQTL in any tissue; otherwise, it is a false



positive. Similar for the other panels. The Area under a Curve (AUC) is also calculated for each curve. The oracle curves correspond to the lfdR approach based on the true model with the true parameters. In all eQTL detection problems, the MT-eQTL and HT-eQTL methods have comparable performance, very similar to the oracle results. While we expect the MT-eQTL to perform similarly to the oracle procedure, it is surprising that the HT-eQTL, only using information in tissue pairs, also provides comparable (although slightly worse) results to the oracle procedure. Both MT-eQTL and HT-eQTL clearly outperform the Meta-Tissue and TBT approaches in all detection problems.

To sum up, the HT-eQTL method achieves high parameter estimation accuracy and eQTL detection power at a low computational cost. For a large number

of tissues, it provides a preferable alternative to the MT-eQTL method.

GTEX v6p data

The GTEx v6p data constitute the most recent freeze for official GTEx Consortium publications, and can be accessed from the GTEx portal at <http://www.gtexportal.org/home/>. We apply the HT-eQTL method to 20 tissues (selected in consultation with the GTEx Analysis Working Group), including 2 brain tissues, 2 adipose tissues, and a heterogeneous set of 16 other tissues. We consider all 70,724,981 cis gene-SNP pairs where the SNP is within 1Mb of the TSS of the gene.

To obtain model parameters using HT-eQTL, we first fit $\binom{20}{2} = 190$ 2-tissue models, and then assemble all the pairwise parameters following the procedure in the

method section. The probability mass vector \mathbf{p} estimated from the Multi-Probit model is summarized in Fig. 3. We particularly focus on 377 configurations with prior probabilities greater than 10^{-5} . The prior probabilities are added up for configurations in the same Hamming class, providing a general characterization of the multi-tissue eQTL distribution. The parabolic shape estimated from the data is concordant with previous results from the pilot study [11]. The global null configuration (the binary $\mathbf{0}$ vector) has the largest probability of 0.936, and the common eQTL configuration (the binary $\mathbf{1}$ vector) has the second largest probability of 0.0396. Configurations in Hamming class 1 (eQTL in only one tissue) and 19 (eQTL in all but one tissue) have relatively large probabilities. All other configurations have much lower probabilities. We remark that as the number of possible configurations increases exponentially with the number of tissues, the prior probability of each configuration is likely to decrease.

Recall that Σ captures the covariance of effect sizes in different tissues when eQTLs are present. We treat the correlation matrix induced from Σ as the distance metric between tissues, and use the single linkage to conduct hierarchical clustering for the 20 tissues. The dendrogram is shown in Fig. 4. We demonstrate that similar tissues, such as the two adipose tissues and the breast tissue, or the two brain tissues, are grouped together. The whole blood is apparently different from all the other tissues. These findings are concordant with those in the pilot analysis [11].

We also carry out testing of eQTL configurations (at a fixed the FDR level of 5%) for the presence of an eQTL in any tissue, in all tissues, in at least one but not all tissues,

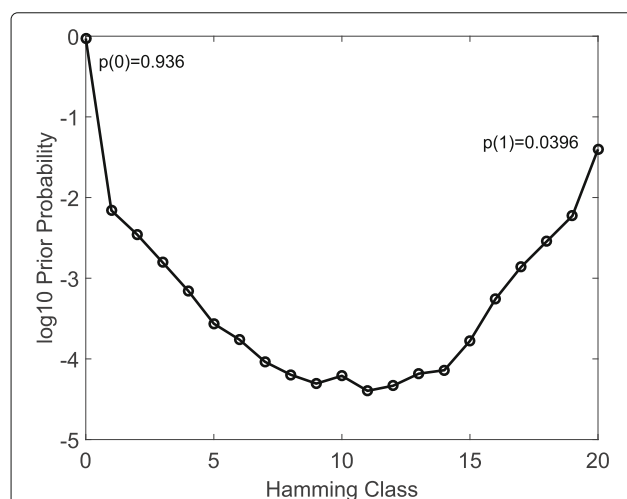


Fig. 3 The summary plot of the probability mass vector estimated from the HT-eQTL method on the GTEx v6p 20-tissue data. The prior probabilities are added up for configurations in the same Hamming class and then log-transformed

and in each individual tissue. The number of discoveries are shown in Table 1. As a comparison, we also apply the Meta-Tissue method [6] to the same data set. In particular, we focus on the Any eQTL detection problem, using p values from the random effects model in Meta-Tissue. We apply the Benjamini and Yekutieli approach [22] to control the FDR at the level of 5%. As a result, we obtain over 6.36 million cis pairs from the Meta-Tissue method. About 3.60 million of these pairs are shared with the HT-eQTL method. We further investigate the unique discoveries of each method. As shown in the left panel of Fig. 5, the unique discoveries made by HT-eQTL have very small p values from the Meta-Tissue method, indicating those are likely to be “near” discoveries for the Meta-Tissue method as well. In the right panel of Fig. 5, however, the excessive unique discoveries made by Meta-Tissue have highly enriched large lfr values. The distribution of the lfr values for the unique Meta-Tissue discoveries is striking. It indicates that the majority of the unique Meta-Tissue discoveries are not close to being significant according to the HT-eQTL model. This may be partially due to the inadequacy of the p-value-based FDR control method for highly dependent tests in Meta-Tissue. We further investigated those gene-SNP pairs with large lfr values, and found that many have large effect sizes with opposite signs in different tissues. These gene-SNP pairs cannot be well characterized by Model (1), because the estimated correlations between tissues are large and positive. As a result, they have large lfrs from HT-eQTL. Further research is needed to determine whether those gene-SNP pairs are true eQTLs of interest or not.

Discussion

In this paper, we develop a new method, HT-eQTL, for joint analysis of eQTL in a large number of tissues. The method builds upon the empirical Bayesian framework, MT-eQTL, proposed in [9], and extends it to 20 or more tissues. Like the earlier model, the HT-eQTL model provides a flexible platform for modeling and testing different configurations of eQTLs, while effectively leveraging information across tissues and across gene-SNP pairs. The model fitting procedure only involves the estimation of all 2-tissue models, and the obtained pairwise parameters are then assembled to get the full model parameters. Even in low-dimensional settings, the HT-eQTL method expedites the parameter estimation procedure of the MT-eQTL model with little cost in accuracy. The detection of eQTLs with different configurations is addressed by adaptively thresholding the corresponding local false discovery rates, which efficiently borrow strength across tissues and control the nominal FDR. Finally, the numerical studies demonstrate the efficacy of the proposed method. In the GTEx v6p data analysis, we apply HT-eQTL to 20 tissues. The estimated prior probabilities of

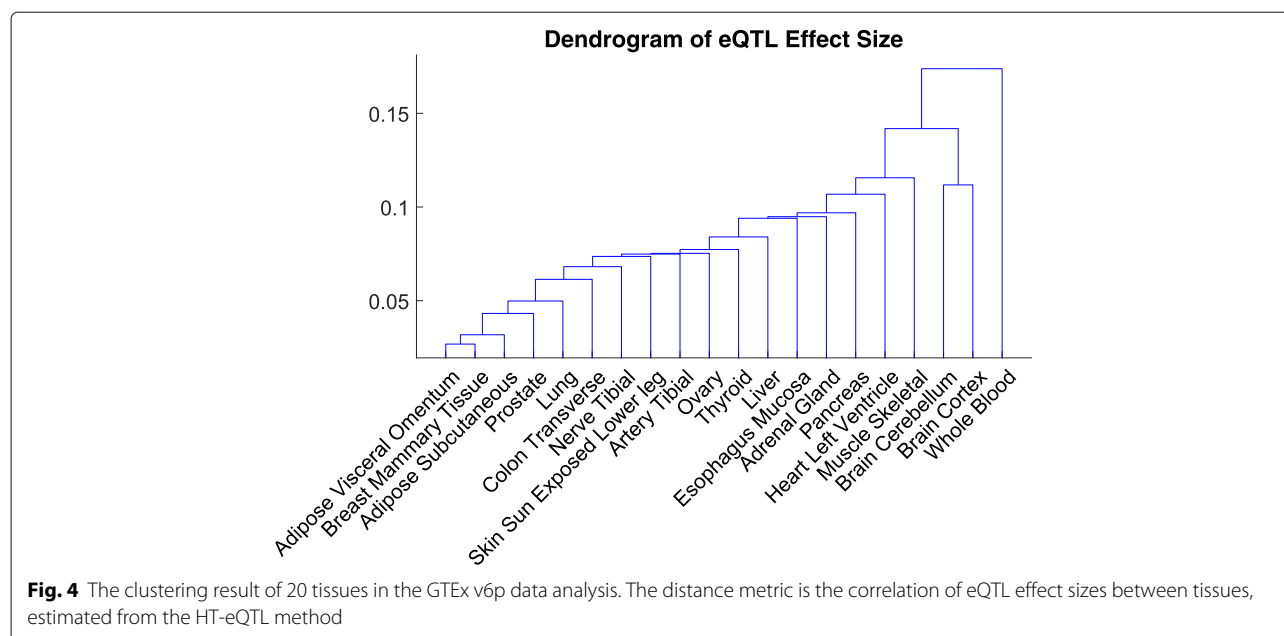


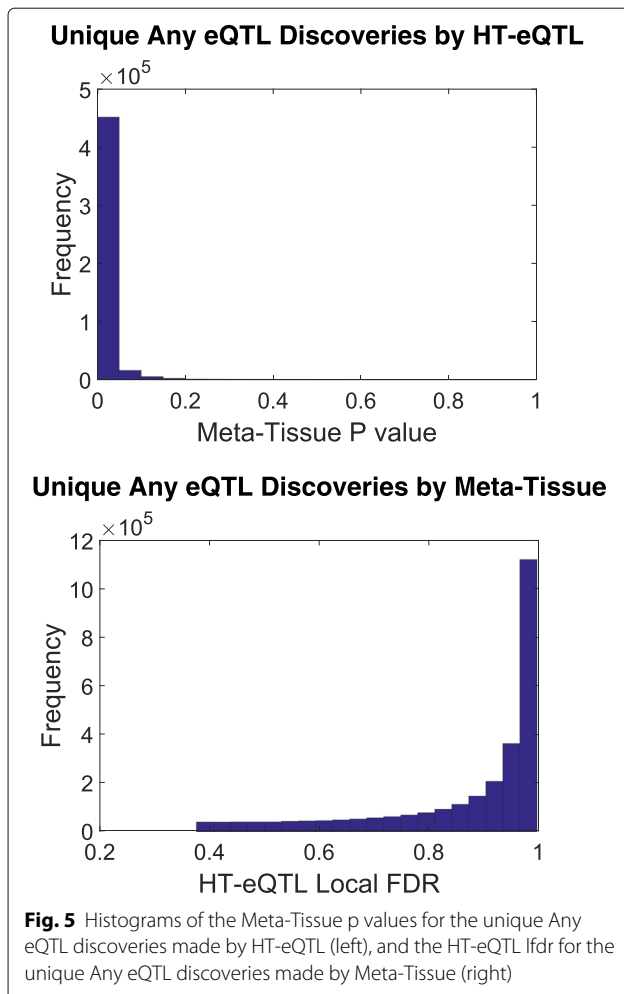
Table 1 The numbers of discoveries and the corresponding percentages of total cis pairs for different eQTL detection problems

eQTL Configuration	Number ($\times 1E6$)	Percentage (%)
eQTL in <i>Any</i> Tissue	4.088	5.78
eQTL in <i>All</i> Tissues	0.708	1.00
<i>Tissue-Specific</i> eQTL	0.239	0.34
Adipose Subcutaneous	3.640	5.15
Adipose Visceral Omentum	3.536	5.00
Adrenal Gland	3.302	4.67
Artery Tibial	3.671	5.19
Brain Cerebellum	3.329	4.71
Brain Cortex	3.120	4.41
Breast Mammary Tissue	3.507	4.96
Colon Transverse	3.515	4.97
Esophagus Mucosa	3.716	5.25
Heart Left Ventricle	3.433	4.85
Liver	1.727	2.44
Lung	3.576	5.06
Muscle Skeletal	3.581	5.06
Nerve Tibial	3.712	5.25
Ovary	2.999	4.24
Pancreas	3.479	4.92
Prostate	3.021	4.27
Skin Sun Exposed Lower Leg	3.717	5.26
Thyroid	3.758	5.31
Whole Blood	3.147	4.45

The FDR level is fixed at 5% for all testing problems

eQTL configurations show that most eQTLs are common across all tissues or present in a single tissue. The estimated effect sizes provide additional insights into the tissue similarity and clustering. We identify a large number of common and tissue-specific eQTLs. A large proportion of the discoveries are replicated by the Meta-Tissue approach. The additional unique discoveries made by our method are “near” discoveries for the Meta-Tissue method, as illustrated by the highly skewed p-value distributions (see Fig. 5). It indicates that HT-eQTL is able to push the detection boundary in a favorable direction (i.e., more statistical power) while preserving error control.

The HT-eQTL method is a necessary first step in the extension of the multi-tissue eQTL model, and a basis for extensions to 30 or more tissues. There are several future research directions. One the one hand, the proposed method relies on the marginal compatibility of a multivariate Gaussian distribution. In practice, if the joint distribution of the z-statistics deviates from the Gaussian distribution, it may affect the model fitting. One may investigate multivariate transformations to make the z-statistics jointly Gaussian. Another direction is to estimate very high dimensional distributions on the space of configurations. One may explore a hierarchical structure in tissues, where each hierarchy only consists of a moderate number of tissues (or tissue groups). Then the proposed method can be applied to each hierarchy separately and combined afterwards. One could also explore computationally efficient and accurate approximations of the cumulative probabilities of a high-dimensional multivariate Gaussian distribution.



Conclusions

We present a scalable method for multi-tissue eQTL analysis. The method can effectively borrow strength across tissues to improve the power of eQTL detection in a single tissue. It also has superior power to detect eQTL of different configurations. The model parameters capture important biological insights into tissue similarity and specificity. In particular, from the GTEx analysis we observe that most *cis* eQTLs are present in either all tissues or a single tissue. The eQTLs identified by the proposed method provide a valuable resource for subsequent analysis, and may facilitate the discovery of genetic regulatory pathways underlying complex diseases.

Additional file

Additional file 1: Supplementary material of the HT-eQTL model fitting procedure. (PDF 144 kb)

Abbreviations

EM: Expectation-Maximization; eQTL: Expression Quantitative Trait Loci; FDR: False discovery rate; GTEx: Genotype-Tissue Expression project; GWAS: Genome-Wide Association Studies; HT-eQTL: High-Tissue Expression Quantitative Trait Loci analysis; lfrdr: Local false discovery rate; Mb: Mega-base; MT-eQTL: Multi-Tissue Expression Quantitative Trait Loci analysis method; ROC: Receiver Operating Characteristic; SNP: Single nucleotide polymorphism; TBT: Tissue-by-Tissue analysis; TSS: Transcription start site

Acknowledgements

The authors would like to thank members of the GTEx Analysis Working Group for helpful comments and discussions.

Funding

This work was supported by the National Institutes of Health [1R01HG008980-01 to GL, R01MH101819 and R01MH090936 to GL, ABN, FAW, R01HG009125 to ABN]; the National Science Foundation [DMS-1613072 to ABN]; and the National Institute of Environmental Health Sciences [P42ES005948 to FAW, P30ES025128 to DJ].

Availability of data and materials

The GTEx v6p data used in this paper are available from the GTEx portal (although application may be required) at <http://www.gtexportal.org/home/>. The Matlab code for implementing the method, including a numerical example, is publicly available at <https://github.com/reagan0323/MT-eQTL/tree/master/HT-eQTL>.

Authors' contributions

GL, ABN and FAW conceptualized the project and developed the novel methodology and analysis. DJ helped conduct the Meta-Tissue method. GL, ABN and FAW contributed to the analysis and interpretation of results and editing the final manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W 168 Street, New York, USA. ²Center for Human Health and the Environment and Bioinformatics Research Center, North Carolina State University, 850 Main Campus Drive, 27695 Raleigh, USA. ³Department of Statistics and Biological Sciences, North Carolina State University, 2311 Stinson Drive, 27695 Raleigh, USA. ⁴Department of Statistics and Operations Research and Department of Biostatistics, University of North Carolina at Chapel Hill, 318 E Cameron Avenue, 27599 Chapel Hill, USA.

Received: 28 August 2017 Accepted: 28 February 2018

Published online: 09 March 2018

References

1. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet.* 2009;10(3):184–94.
2. Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet.* 2009;10(8):565–77.
3. Peterson CB, Bogomolov M, Benjamini Y, Sabatti C. Treeqtl: hierarchical error control for eqtl findings. *Bioinformatics.* 2016;32(16):2556–8.
4. Bogomolov M, Peterson CB, Benjamini Y, Sabatti C. Testing hypotheses on a tree: new error rates and controlling strategies. *arXiv preprint arXiv:1705.07529.* 2017.

5. Lewin A, Saadi H, Peters JE, Moreno-Moral A, Lee JC, Smith KG, Petretto E, Bottolo L, Richardson S. Mt-hess: an efficient bayesian approach for simultaneous association detection in omics datasets, with application to eqtl mapping in multiple tissues. *Bioinformatics*. 2015;32(4):523–32.
6. Sul JH, Han B, Ye C, Choi T, Eskin E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet*. 2013;9(6):1003491.
7. Han B, Duong D, Sul JH, de Bakker PI, Eskin E, Raychaudhuri S. A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum Mol Genet*. 2016;25(9):1857–66.
8. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet*. 2013;9(5):1003486.
9. Li G, Shabalin AA, Rusyn I, Wright FA, Nobel AB. An empirical bayes approach for multiple tissue eQTL analysis. *Biostatistics*. 2017. kxx048. <https://doi.org/10.1093/biostatistics/kxx048>.
10. Urbut SM, Wang G, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *bioRxiv*. 2017096552.
11. The GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015;348(6235):648–60.
12. Mudholkar GS, Chaubey YP. On the distribution of Fisher's transformation of the correlation coefficient. *Commun Stat Simul Comput*. 1976;5(4):163–72.
13. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc*. 2001;96(456):1151–60.
14. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. 2004;5(2):155–76.
15. Efron B. Size, power and false discovery rates. *Ann Stat*. 2007;35(4):1351–77.
16. Efron B. Microarrays, empirical Bayes and the two-groups model. *Stat Sci*. 2008;23(1):1–22.
17. Sun W, Cai TT. Oracle and adaptive compound decision rules for false discovery rate control. *J Am Stat Assoc*. 2007;102(479):901–12.
18. Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN, Welsh-Bohmer KA, Hulette CM, Denny TN, Goldstein DB. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol*. 2008;6(12):1000001.
19. Dimas AS, Deutsch S, Stranger BE, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*. 2009;325(5945):1246–50.
20. Ding J, Gudjonsson JE, Liang L, et al. Gene expression in skin and lymphoblastoid cells: refined statistical method reveals extensive overlap in cis-eQTL signals. *Am J Hum Genet*. 2010;87(6):779–89.
21. Fu J, Wolfs MG, Deelen P, et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet*. 2012;8(1):1002431.
22. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29:1165–88.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

