Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 2 April 2015
Applications Note

OXFORD

Genome analysis

Applying meta-analysis to Genotype-Tissue Expression data from multiple tissues to identify eQTLs and increase the number of eGenes

Dat Duong [1], Lisa Gai [1], Sagi Snir [2,3], Eun Yong Kang [1], Buhm Han [4,5], Jae Hoon Sul [6,†] and Eleazar Eskin [1,7,†,*]

[1]Department of Computer Science, University of California Los Angeles, 90095, USA
[2]Institute of Evolution, University of Haifa, Haifa, 3498838, Israel
[3]Department of Evolutionary and Environmental Biology, University of Haifa, 3498838, Israel
[4]Department of Convergence Medicine, University of Ulsan College of Medicine, Ulsan, 44610, Republic of Korea
[5]Asan Institute for Life Sciences, Asan Medical Center, Seoul,05505, Republic of Korea
[6]Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, 90095, USA and
[7]Department of Human Genetics, University of California Los Angeles, 90095, USA
[†]These authors contributed equally to this work

[*]To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: There is recent interest in using gene expression data to contextualize findings from traditional genome wide association studies (GWAS). In a specified tissue, expression quantitative trait loci (eQTLs) are genetic variants associated with gene expression, and eGenes are genes whose expression levels are associated to genetic variants. eQTL and eGenes provide great supporting evidence for GWAS hits and provide potential insights into the regulatory pathways involved in disease. If a significant variant or a candidate gene identified by GWAS is also an eQTL or eGene, then there is strong evidence to further study this variant or gene. Multi-tissue gene expression datasets like the Gene Tissue Expression (GTEx) data are used to find eQTLs and eGenes. Unfortunately, these datasets often have small sample size in some tissues. For this reason, there have been many meta-analysis methods designed to combine gene expression data across many tissues to increase power for finding eQTLs and eGenes. These existing methods are not scalable to datasets with many tissues like the GTEx data. Furthermore, these methods ignore a biological insight that the same variant may be associated with the same gene across similar tissues. **Result:** We introduce a meta-analysis model that addresses the problems in these existing methods. We focus on the problem of finding eGenes in data from many tissues, and show that our model is better than other types of meta-analyses.
Availability: Source code is at https://github.com/datduong/RECOV
Contact: eeskin@cs.ucla.edu datdb@cs.ucla.edu

1 Introduction

Expression quantitative trait loci (eQTL) studies find eQTLs, which are genetic variants associated with gene expression, and eGenes, which are genes whose expression levels are associated with at least one genetic variant. eQTL studies are related to traditional genome wide association studies (GWAS) that find variants associated with disease.

Both eQTL studies and GWAS focus on single nucleotide polymorphisms (SNPs). Many SNPs found by GWAS are located in intergenic regions, and often their relationship to the disease phenotype is not obvious. A gene

© The Author XXXX. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

expression is an intermediate phenotype between a causal SNP and a disease (Huang *et al.*, 2014). Thus, eQTL studies may provide biological insights into the mechanism through which the disease occurs. If a significant SNP identified by GWAS is found to be an eQTL, there is a strong evidence to further study the variant. For this reason, top hits in GWAS that are also expression quantitative trait loci (eQTL) are of special interest In fact, recent GWAS have confirmed that many disease-causing variants are eQTLs (Albert, 2016; Liu *et al.*, 2016; Nieuwenhuis *et al.*, 2016). Similarly, genes near GWAS-significant variants that have been identified as eGenes may warrant further study as candidate casual genes. Thus, eQTL studies provide important supporting evidence for GWAS results and potential insights into the regulatory pathways involved in disease.

The underlying approach behind eQTL studies and GWAS is the same. In an eQTL study, one performs association tests between the genotype data and the gene expression (instead of disease statuses) to identify variants which are associated with the gene expression. eQTLs and eGenes may be specific to only one or a group of tissues, as a gene is not always uniformly expressed in every tissue. For example, SNPs associated to schizophrenia have been found to be eQTLs in only the brain tissues, indicating that schizophrenia is affecting how the brain functions (Fromer *et al.*, 2016). For this reason, there have been recent large-scale studies to collect gene expression data in many tissues, such the Genotype-Tissue Expression (GTEx) project (The GTEx Consortium, 2015). This GTEx dataset contains gene expression data in 44 tissues and genotypes of 5 million SNPs for over 300 individuals.

To find eQTLs from the GTEx data and other multi-tissue datasets, one can apply the traditional tissue-by-tissue (TBT) approach in which a separate eQTL study is done for each tissue. However, many tissues do not have enough samples to detect SNPs that are weakly associated to the gene expressions. To address this issue, there have been many efforts in developing different types of meta-analysis, which gather data from many tissues to increase the total sample size and power to detect eQTLs. Two notable methods are Meta-Tissue and eQTLBma. Both have been shown to outperform the traditional TBT method (Flutre *et al.*, 2013; Sul *et al.*, 2013).

Meta-Tissue and eQTLBma have an important limitation that reduces their applicability to large gene expression datasets such as the GTEx data. Both methods are computationally intensive and can be used for datasets containing at most 10 or 20 tissues respectively (Flutre *et al.*, 2013; Sul *et al.*, 2013). Meta-Tissue uses both linear mixed models (LMM) and fixed (or random) effects meta-analysis to combine data from many tissues. Meta-Tissue must estimate the variance components in its LMM setup for every pair of variant and gene expression; thus, its running time is impractical when there are thousands of genes or too many tissues (Sul *et al.*, 2013).

eQTLBma uses a Bayesian approach that considers all possible combinations of tissues in which a SNP is an eQTL. This setup corresponds to 2^T configurations where T is the number of tissues, making the method infeasible when T is 44 like in the GTEx data (Flutre *et al.*, 2013).

As an alternative to Meta-Tissue and eQTLBma, the GTEx consortium used a meta-analysis software called Metasoft, introduced by Han and Eskin (2011). Metasoft is equivalent to Meta-Tissue without the LMM setup (Han and Eskin, 2011, 2012). Metasoft extends the random-effects (RE) meta-analysis model; this extended model is named RE2 (Han and Eskin, 2011).

However, Meta-Tissue, eQTLBma and RE2 assume that a SNP has independent effects on a given gene expression in each tissue. This ignores the fact that the same SNP tends to have similar effects in related tissues (The GTEx Consortium, 2015).

Recently, Acharya *et al.* (2016) introduces a method that amends this shortcoming in Meta-Tissue, eQTLBma and RE2. The model developed by Acharya *et al.* (2016) requires genotype and gene expression data for each individual in each tissue. Their implementation in R using the JARGUAR library needs to load all these data into memory. This can be memory inefficient when there are many genes and tissues.

In this paper, we present a novel meta-analysis method named RECOV. Unlike Meta-Tissue and eQTLBma, RECOV is applicable to large gene expression datasets and can analyze all 44 tissues in the GTEx data. Like JARGUAR, RECOV considers the biological insight that a variant may have similar effects on a gene across tissues. However, unlike JARGUAR, RECOV needs only the summary statistic (i.e. SNP effect and its variance) at each SNP in each tissue and not the complete genotype and gene expression data for each individual. RECOV is based on the RE2 meta-analysis framework and uses a covariance matrix to explicitly model the correlation of a SNP effect on a gene expression in similar tissues.

In the method section, we describe RECOV in detail and how it can be used to identify eGenes from eQTL studies in more than one tissue. In the result section, we use simulated datasets to show that RECOV has correct false positive rate. We then apply RECOV to real multi-tissue expression data in the GTEx data, and show that our approach detects more eGenes than the previous RE2 and TBT methods.

2 Methods

We begin by introducing the notations that are used in this paper. We use the notation $\mathbf{x} \in \mathbb{R}^n$ to specify a vector $\mathbf{x}$ with dimension n, and $\mathbf{Z} \in \mathbb{R}^{n \times m}$ to specify a matrix $\mathbf{Z}$ with dimension $n \times m$. We use x_i to denote the i^{th} element in $\mathbf{x}$, and likewise, Z_{ij} to specify entry ij in $\mathbf{Z}$. We denote an item k in the set K by $k \in K$, and a set $\{a_1 ... a_K\}$ indexed by k by using $\{a_k\}_{k \in K}$, where the subscript $k \in K$ is omitted whenever the context is clear. The size of the set K is denoted as $|K|$.

2.1 Detecting one eGene via an eQTL study

2.1.1 eQTL study in one tissue

We begin with an eQTL study in one tissue t. An eQTL study finds every eQTL associated with the expression level of a specific gene g. To do this, the study tests each variant v in the set V against the expression of g in a sequential fashion. To set up the problem, suppose we represent the gene expression for m individuals in tissue t as a vector $\mathbf{q} \in \mathbb{R}^m$, and we want to find the effect of variant v on g. Let $\mathbf{s} \in \mathbb{R}^m$ be the standardized genotypes of this v. The eQTL study assumes the following model

$$\mathbf{q} = \beta_{vgt}\mathbf{s} + \boldsymbol{\epsilon}_{vgt} \tag{1}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^m$ is the vector of sampling errors $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma_\epsilon^2 \mathbf{I})$, and $\beta_{vgt} \in \mathbb{R}$ is the true effect size of the variant v on g in tissue t (Darnell *et al.*, 2012; Eskin, 2015; Hormozdiari *et al.*, 2015). The estimate b_{vgt} of the true value β_{vgt} can be computed using the basic least squares equation $b_{vgt} = \arg\min_{\beta_{vgt}} ||\mathbf{q} - \beta_{vgt}\mathbf{s}||_2^2$. This solution is

$$b_{vgt} = (\mathbf{s}^\intercal \mathbf{s})^{-1}\mathbf{s}^\intercal \mathbf{q} \qquad \text{where} \qquad b_{vgt} \sim \mathrm{N}(\beta_{vgt}, (\mathbf{s}^\intercal \mathbf{s})^{-1}\sigma_\epsilon^2) \tag{2}$$

(Abraham and Ledolter, 2006). By using Eq. 2 and writing the null hypothesis $\mathrm{H}_0 : \beta_{vgt} = 0$, one can do a hypothesis test to assert if v has an effect toward g. To do this test, compute the estimate $\hat{\sigma}^2_{\epsilon_{vgt}}$ of $\sigma^2_{\epsilon_{vgt}}$ by

$$\hat{\sigma}^2_{\epsilon_{vgt}} = \frac{1}{m-1}\sum_{i=1}^{m}(q_i - b_{vgt}s_i)^2 \tag{3}$$

and estimate the variance d_{vgt} of b_{vgt} by

$$d_{vgt} = (\mathbf{s}^\intercal \mathbf{s})^{-1}\hat{\sigma}^2_{\epsilon_{vgt}} \tag{4}$$

then compute the p-value $p_{vgt} = \text{p-value}(b_{vgt})$ (Abraham and Ledolter, 2006; Eskin, 2015). If p_{vgt} is less than some significance level , then we

reject $H_0 : \beta_{vgt} = 0$, and conclude that v is an eQTL of g in tissue t. When many variants are tested, we must apply a multiple testing correction; for example, one can apply Bonferroni correction by using the threshold $\alpha/|V|$ where α is the significance level for the whole family of tests. The Bonferroni correction is conservative when there is linkage disequilibrium (LD) in set V. There exist other methods that can handle LD better than the Bonferroni correction (Conneely and Boehnke, 2007; Joo *et al.*, 2014; Hormozdiari *et al.*, 2015; Joo *et al.*, 2016).

2.1.2 Using an eQTL study in one tissue to discover one eGene

Because an eQTL study tests each variant $v \in V$ against gene g in a tissue t, from one single eQTL study, we have a set of p-values $\{p_{vgt}\}_{v \in V}$. The minimum $p_{gt} = \min_{v \in V} \{p_{vgt}\}$ is defined to be the observed eGene statistic at gene g in tissue t (The GTEx Consortium, 2015). Define $\alpha_{p_{gt}} = \text{p-value}(p_{gt})$ to be the eGene p-value (The GTEx Consortium, 2015). The eGene p-value depends on two important factors: the number of variants $|V|$, and the LD of the variants. In practice, $\alpha_{p_{gt}}$ is computed by doing a permutation test (Sul *et al.*, 2015; Duong *et al.*, 2016). In brief, in the k^{th} permutation, one would permute the gene expressions among the individuals, and compute a new $p_{gt}^{(k)} = \min_{v \in V} \{p_{vgt}^{(k)}\}$. $\alpha_{p_{gt}}$ is the ranking of the observed p_{gt} with respect to the density created from many $p_{gt}^{(k)}$. One can then conclude that g is an eGene in tissue t if its eGene p-value $\alpha_{p_{gt}}$ is less than some desired threshold.

2.2 Tissue-by-tissue analysis to find one or many eGenes

When there are genotype-tissue expression data from many tissues, the tissue-by-tissue (TBT) analysis is the standard method to find eGenes (Sul *et al.*, 2013; The GTEx Consortium, 2015). TBT tests whether or not the gene has at least one eQTL in each tissue by examining each tissue individually. Suppose the gene is expressed in T tissues. Then TBT performs T eQTL studies (one test in each tissue). The null hypothesis is that the gene is not an eGene in any tissue. This hypothesis is equivalent to saying that no eQTL is found for this gene in any tissue.

Three layers of multiple testing correction are required since TBT performs one test per gene in each tissue. The first layer of multiple testing correction is applied within a tissue and corrects for LD of the variants tested against the gene. This correction can be done by using the permutation test to compute the eGene p-value for the gene in the tissue (Sul *et al.*, 2013, 2015; The GTEx Consortium, 2015; Duong *et al.*, 2016).

The second layer of multiple testing correction adjusts for the fact that we may test more than one gene within a tissue. For example, the GTEx pilot study tested thousand of genes within one tissue, and then transformed eGene p-values into eGene q-values to control for this multiple testing (Dabney *et al.*, 2010; The GTEx Consortium, 2015). This second layer of multiple testing correction is not needed if only one gene is tested in each tissue.

The third layer of multiple testing correction takes into account the fact that one gene is tested T number of times (once per tissue) (Sul *et al.*, 2013). In this paper, we apply Bonferroni correction so that the false-positive threshold for any eGene q-value in each tissue is α/T, where α is 5% for example. In this layer, other multiple testing correction methods can be used instead of the Bonferroni correction. This paper however focuses on the meta-analysis model, and measuring the performances of various multiple testing correction methods is beyond its scope.

2.3 Meta-analysis models for combining eQTL studies across tissues

We motivate the application of meta-analysis for combining eQTL studies across tissues. An eQTL is defined not only with respect to a gene, but also with respect to the tissue in which the gene expression is measured. eQTL studies of the same gene have been analyzed separately at the tissues level (The GTEx Consortium, 2015). We can better detect the effect of a variant on the gene by combining eQTL results across many tissues and modeling the relatedness of the effect sizes of one variant among the tissues.

It is important to emphasize that when using meta-analysis to find many eGenes, one would need only two layers of multiple testing correction. The first layer is applied within a gene to correct for LD because one tests many variants against the gene. The second layer is applied at the gene level because there is usually more than one gene being tested.

We define the notations to be used later. Suppose we have T eQTL studies (one study per tissue) that test the association of a variant v at a gene g. As shown above, denote the effect of this variant in the study (i.e. tissue) t as b_{vgt}, where b_{vgt} is computed using Eq. 2. Denote the variance of b_{vgt} in the study t as d_{vgt} where d_{vgt} is computed using Eq. 4. Let $\mathbf{b}_{vg} \in \mathbb{R}^T$ contain the effects in these T studies, so that $\mathbf{b}_{vg} = \begin{bmatrix} b_{vg1} & b_{vg2} & \cdots & b_{vgT} \end{bmatrix}^\intercal$. Let $\mathbf{D}_{vg} = \mathbf{diag}(d_{vg1} \dots d_{vgT})$.

2.3.1 Random effects (RE) and the RE2 model

The maximum likelihood procedure in RE model assumes that $\mathbf{b}_{vg}$ has the form (Thompson and Sharp, 1997; Han and Eskin, 2011)

$$\mathbf{b}_{vg} = \boldsymbol{\lambda}_{vg} + \boldsymbol{\epsilon}_{vg} \tag{5}$$

The random sampling errors $\boldsymbol{\epsilon}_{vg}$ are estimated from the data and assumed to be $\boldsymbol{\epsilon}_{vg} \sim \text{N}(0, \mathbf{D}_{vg})$. $\boldsymbol{\lambda}_{vg} \in \mathbb{R}^T$ in the RE model is a random variable, that is $\boldsymbol{\lambda}_{vg} \sim \text{N}(\mu_{vg}\mathbf{1}, \tau_{vg}^2\mathbf{I})$ with $\mu_{vg} \in \mathbb{R}$ and $\tau_{vg}^2 \in \mathbb{R}_+$. The effect $\boldsymbol{\lambda}_{vg}$ is thus known as the random effect. μ_{vg} is the common true underlying effect that all the studies inherit. The term τ_{vg}^2 is the heterogeneity among the effects of the variant in T tissues.

Clearly, $\mathbf{b}_{vg}$ comes from the distribution

$$\mathbf{b}_{vg} \sim \text{N}(\mu_{vg}\mathbf{1}, \tau_{vg}^2 I + \mathbf{D}_{vg}) \tag{6}$$

The traditional RE model assumes that if the effect of the variant does not exist in any tissue, then $\mu_{vg} = 0$. However, it has been shown that this traditional null hypothesis does not yield optimal statistical power in detecting eQTLs (Han and Eskin, 2011, 2012). For this reason, the RE2 model assumes a different null hypothesis, that if the effect of the variant does not exist in any tissue, then $\mu_{vg} = 0$ and $\tau_{vg}^2 = 0$. The fact that $\tau_{vg}^2 = 0$ is a result of $\mu_{vg} = 0$ because when the effect does not exist, its variance must not exist (Han and Eskin, 2011, 2012; Kang *et al.*, 2014). We will compare our method against the RE2 model.

The null hypothesis H_0 in RE2 is

$$H_0 : \quad \mu_{vg} = 0 \quad \tau_{vg}^2 = 0 \tag{7}$$

The likelihood ratio test becomes

$$\text{llr}_{vg} = 2 \log \frac{\sup_{\mu_{vg}, \tau_{vg}^2} L(\mathbf{b}_{vg} | \mu_{vg}, \tau_{vg}^2)}{L(\mathbf{b}_{vg} | \mu_{vg} = 0, \tau_{vg}^2 = 0)} \tag{8}$$

The function L denotes the likelihood function of the random variable $\mathbf{b}_{vg}$. The numerator $\sup_{\mu_{vg}, \tau_{vg}^2} L(\mathbf{b}_{vg} | \mu_{vg}, \tau_{vg}^2)$ may be estimated using numerical methods or other heuristic methods. Here, we apply the Nelder-Mead method, which is a heuristic derivative free search method.

In finding the supremum, one implicitly enforces $\tau_{vg}^2 \geq 0$. Due to this restricted parameter space, the asymptotic density of the likelihood ratio is an average of a χ_1^2 and χ_2^2 (Self and Liang, 1987; Han and Eskin, 2011). To find the p-value of this likelihood ratio when T is large, one can use this asymptotic density.

Otherwise, one may compute the likelihood ratio p-value by creating a density of likelihood ratios under the null hypothesis and ranking the observed likelihood ratio with respect to this density. This null density

is made by sampling many instances of $\mathbf{b}_{vg}$ using Eq. 6 with $\mu_{vg} = \tau^2_{vg} = 0$, and computing their corresponding llr_{vg}. If the p-value of llr_{vg} is significant, then v is an eQTL with respect to g in at least one tissue. Because we have 44 tissues in the GTEx data, we will use the asymptotic distribution of the likelihood ratio.

2.3.2 RECOV: Random effects (RE) model with a covariance (COV) term

Here we present an extension to the RE model. We first discuss the covariance term. Eq. 5 of the RE model assumes $\boldsymbol{\lambda}_{vg} \sim \text{N}(\mu_{vg}\mathbf{1}, \tau^2_{vg}\mathbf{I})$ so that the effects of variant v toward gene g are independent across the tissues. However, tissues from the same body part are similar; in fact, many eQTLs are found to be consistent among many tissues (Flutre *et al.*, 2013). From this observation, we must acknowledge that $\boldsymbol{\lambda}_{vg} \sim \text{N}(\mu_{vg}\mathbf{1}, \boldsymbol{\Sigma}_{vg})$ where $\boldsymbol{\Sigma}_{vg}$ is not diagonal. The term $\boldsymbol{\Sigma}_{vg} \in \mathbb{R}^{T\times T}$ models the covariance of effect sizes of v among tissues conditioned on the gene g. The matrix $\boldsymbol{\Sigma}_{vg}$ contains $T \times T$ unknown parameters which are to be estimated. In practice, one has to assume a simpler form for $\boldsymbol{\Sigma}_{vg}$. Here, we assume $\boldsymbol{\Sigma}_{vg} \approx c_{vg}\mathbf{U}_{vg}$. The matrix $\mathbf{U}_{vg}$ is estimated from the data. The term $c_{vg} \geq 0$ is an unknown scaling constant and is to be optimized jointly with the mean of regression coefficient μ_{vg}.

In this paper, we compute the $\mathbf{U}_{vg}$ at each variant-gene pair as follows. Denote $\mathbf{B}_g = \begin{bmatrix}\mathbf{b}_{1g} & \mathbf{b}_{2g} & \cdots & b_{|V|g}\end{bmatrix}$ so that $\mathbf{B}_g \in \mathbb{R}^{T\times|V|}$. Thus, a column in $\mathbf{B}_g$ contains the effects of a variant in 44 tissues. To avoid reusing the data when testing a single SNP, we remove its effects in the 44 tissues when estimating its covariance term. To do this, we divide all cis-variants of g into 10 separate segments according to their physical locations on the chromosome, and use the 9 segments that do not contain v to compute $\mathbf{U}_{vg}$. In particular, denote $\mathbf{B}_{-vg}$ as the matrix $\mathbf{B}_g$ without the effect sizes of the variants that belong to the same segment as v. $\mathbf{U}_{vg}$ can be estimated as $\mathbf{U}_{vg} = \mathbf{B}_{-vg}\mathbf{B}^{\intercal}_{-vg}$ (after proper scaling is applied to $\mathbf{B}_{-vg}$). This computation is similar to how one would compute a kinship matrix using the genotype matrix (Eskin, 2015). In this scheme, we also hope that the variants in strong LD with v are removed, so that there are fewer vectors in $\mathbf{B}_{-vg}$ that resemble $\mathbf{b}_{vg}$ when computing $\mathbf{U}_{vg}$.

Now, we are ready to introduce this covariance term $\mathbf{U}_{vg}$ to the RE model. We extend the RE model so that when testing a variant v against gene g, we have

$$\mathbf{b}_{vg} = \boldsymbol{\lambda}_{vg} + \epsilon_{vg} \tag{9}$$

where

$$\boldsymbol{\lambda}_{vg} \sim \text{N}(\mu_{vg}\mathbf{1}, c_{vg}\mathbf{U}_{vg}) \quad \epsilon_{vg} \sim \text{N}(0, \mathbf{D}_{vg}) \tag{10}$$

Like before, the matrix $\mathbf{D}_{vg}$ is known because it contains the observed variances of the SNP effects estimated by Eq.4. This form for $\mathbf{D}_{vg}$ is standard in meta-analysis (Thompson and Sharp, 1997). We now have

$$\text{cov}(\mathbf{b}_{vg}) = c_{vg}\mathbf{U}_{vg} + \mathbf{D}_{vg} \tag{11}$$

$$\mathbf{b}_{vg} \sim \text{N}(\mu_{vg}\mathbf{1}, \text{cov}(\mathbf{b}_{vg})) \tag{12}$$

The null hypothesis that v does not effect g in all T tissues is

$$\text{H}_0 : \qquad \mu_{vg} = 0 \qquad c_{vg} = 0 \tag{13}$$

The alternative hypothesis implies that v has an effect in at least one of the T tissues.

Under this setting, the log likelihood ratio to test the hypothesis becomes

$$\text{llr}_{vg} = 2\log\ \frac{\sup_{\mu_{vg},\, c_{vg}} L(\mathbf{b}_{vg}|\mu_{vg},\ c_{vg})}{L(\mathbf{b}_{vg}|\mu_{vg} = 0, c_{vg} = 0)} \tag{14}$$

Like in the RE model, in finding the supremum in the alternative, one enforces $c_{vg} \geq 0$. Due to this restricted parameter space, the asymptotic density of the likelihood ratio is an average of a χ^2_1 and χ^2_2. Alternatively, one can compute the empirical p-value of this likelihood ratio with a permutation test. In any case, if the p-value of the likelihood ratio is significant, then v is an eQTL with respect to g in at least one tissue.

2.4 Using meta-analysis of eQTL in many tissues to identify eGenes

In practice, a set of variants V is tested against g. Here we describe how one can combine the meta-analysis result at each variant $v \in V$ to determine if g is an eGene.

Define $p_{vg} = \text{p-value}(\text{llr}_{vg})$ so that from many variants, we have the set of p-values $\{p_{vg}\}_{v\in V}$. The observed statistic at gene g is $p_g = \min_{v\in V}\{p_{vg}\}$. To determine if p_g is significant, one needs to compute its eGene p-value denoted as α_{p_g} (The GTEx Consortium, 2015). To control for multiple testing when LD exists between the variants, one can compute α_{p_g} using a permutation test (The GTEx Consortium, 2015; Sul *et al.*, 2015; Duong *et al.*, 2016). The permutation test creates a distribution of the observed p_g under the null hypothesis, which can then be used to compute the eGene p-value α_{p_g} of p_g.

This permutation test can be done as follows. Let K be the number of permutations. In the k^{th} permutation, permute the gene expression of g among the individuals in each of the T tissues so that there are T permuted datasets. This permutation reflects the hypothesis that the gene is not an eGene in any tissue. Next, redo the meta-analysis at each variant $v \in V$ so that a new $p^{(k)}_g = \min_{v\in V}\{p^{(k)}_{vg}\}$ is computed. α_{p_g} is the fraction of times the observed p_g is less than $p^{(k)}_g$. The gene g is an eGene in at least one tissue if its eGene p-value α_{p_g} is below some threshold α.

In the pilot GTEx analysis, a set of genes G is being tested at once, so that one has a set $\alpha_G = \{\alpha_{p_g}\}_{g\in G}$. To control for the family wise error rate, one can apply Bonferroni correction to get the threshold $\alpha/|G|$. Any gene $g \in G$ with $\alpha_{p_g} < \alpha/|G|$ is an egene in at least one of the T tissues.

2.4.1 Estimating eGene p-value

The permutation test above must be performed at every pair of variant $v \in V$ and gene $g \in G$ in a tissue t. The entire permutation test requires $K|V||G|T$ permuted datasets, which is time consuming. Here, we introduce an alternate method to estimate the eGene p-value. The permutation test in essence estimates a function f that maps a test statistic to its p-value. There is evidence that the correlation of test statistics at two variants is equal to their LD (Han *et al.*, 2009). This holds true when the test statistics are effect sizes (Han *et al.*, 2009).

In our meta-analysis, to properly estimate the eGene p-value α_{p_g}, we must consider the effect of LD in the set of variants V on the observed statistic at each variant v. At each variant, it does not matter whether we treat its llr_{vg} or its p_{vg} as the observed statistic, because the likelihood ratio and its p-value are two equivalent entities for two reasons. First, the likelihood ratio of each variant v has the same distribution and degree of freedom. Second, the p-value function is 1-to-1 and strictly monotone. Thus, having a null density for the $\max_{v\in V}\{\text{llr}_{vg}\}$ is equivalent to having a null density for the minimum p_g.

We empirically find that on average the correlation of the likelihood ratios at any two variants is roughly equal to their LD; that is, on average $\text{cor}(\text{llr}_{ug}, \text{llr}_{vg}) \approx \text{LD}(u, v)$ for any variant $u, v \in V$ (Figure 1). For this reason, any function f that accounts for LD and maps an observed test statistic of a gene into an eGene p-value would be applicable in our case. We can use such a function to convert the observed test statistics p_g at the gene g to eGene p-value α_{p_g} without doing the permutation test. Each gene g has its own LD structure and requires its own function f, because the cis-variants of each gene are non-identical.

We apply MVN-EGENE to estimate the function f for each gene. MVN-EGENE is a software that tests if a gene is an eGene in one tissue.

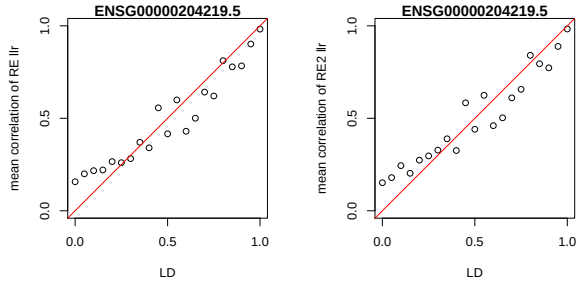


Fig. 1: **Relationship between likelihood ratios of a pair of variants and their LD.** A SNP may be a cis-SNP for multiple genes and therefore tested as an eQTL for multiple genes. Denote cor(llr_u, llr_v) as the correlation between the likelihood ratio statistics of variants u and v in all genes where they are both cis-SNPs. Using cis-SNPs from the gene ENGS00000204219.5, we randomly selected pairs of SNPs that appeared in at least two genes together. We randomly selected pairs of cis-SNPs of each gene that appeared in at least other two genes together from the first 1000 cis-SNPs in the gene. Twenty pairs were selected from LD bins of of width 0.05. The correlation of their likelihod ratios were computed using any other genes where they appear as a cis-SNP. We used both RECOV and RE2 meta-analysis to compute the likelihod ratios at each variant. We then find the mean cor(llr_u , llr_v) in each LD bin, using either RECOV (left) or RE2 (right) likelihood ratios. Plots show correlations after taking absolute value. The identity line is shown in red. Plots for additional genes are shown in the supplemental materials at https://github.com/datduong/RECOV.

MVN-EGENE is designed so that one does not need to do the permutation test when estimating an eGene p-value; it is unable to simultaneously consider more than one tissues like a meta-analysis would.

To compute the function f at a gene, we apply MVN-EGENE at that gene in a tissue (Sul *et al.*, 2015). We assume that the LD does not change much between tissues and it does not matter much which particular tissue is chosen, as long as it has many samples.

In MVN-EGENE, the test statistic for a gene is the most significant effect size taken over all cis-variants. The p-value of this test statistic depends on the LD of the cis-variants. Instead of doing a permutation test to compute this p-value, MVN-EGENE simulates data under the null hypothesis using a multivariate normal distribution. In brief, in one simulation, MVN-EGENE samples the effect sizes of the cis-variants of a gene in a tissue using zero as the mean effect and LD as the covariance matrix. In this simulation, the most significant effect among these effect sizes is taken to be the test statistic at the gene. After many simulations, one can create a null distribution for the observed test statistic. One can easily convert an effect size into a p-value using a normal distribution. By having a null density of the most significant effect size taken over all the variants, one also has the null density of the minimum p-value taken over all the variants. This null density of the minimum p-value in MVN-EGENE properly handles LD at the gene. Here, we use this distribution of minimum p-values as our null density to convert the observed minimum likelihood ratio p-value p_g to its eGene p-value α_{p_g} in both RECOV and RE2.

2.4.2 Estimating genomic control

In the GTEx dataset and other multi-tissue gene expression datasets, the same individual may provide samples for many tissues (Figure 2). Sharing of samples from the same individuals among tissues is known to inflate the false positive rate in a meta-analysis (Han *et al.*, 2016). Before testing whether the RECOV outcome is affected by the fact that tissues share individuals, we test if RECOV inflates the false positive rate when the data is absolutely free of any spurious statistical association. These signals can be due to LD, shared individuals in tissues, batch effects, or correlated expressions of the same gene (or between genes) across tissues. It is important to mention that in the real GTEx data, batch effects have been dealt with by the GTEx consortium by applying PEER factors on the gene expression in each tissue (The GTEx Consortium, 2015). Section 2.4.1 above describes how RECOV and RE2 handle LD in the variants. We now describe how we use a genomic control (GC) factor to remove the effect of shared individuals in the tissues from the meta-analysis results. This GC factor is clearly data dependent as different datasets will require different GC values.

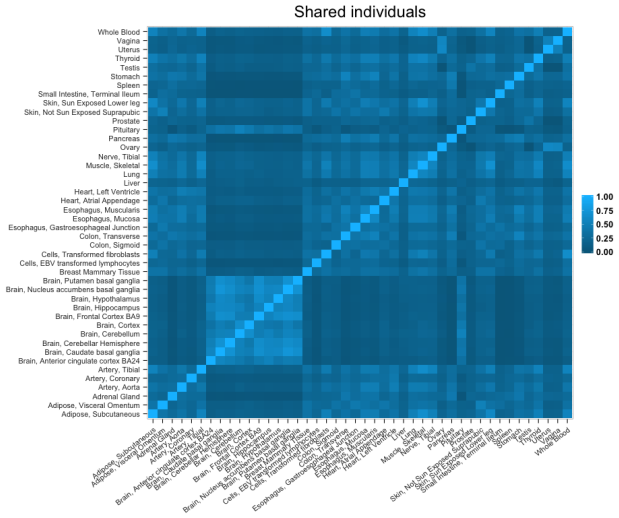


Fig. 2: Shared individuals among the 44 tissues in the GTEx dataset. Degree of overlap of individuals between two tissues is measured using the Jacquard index.

Here, we focus on finding the GC factor for the GTEx data. To do this, we simulate two types of datasets and compare their behaviors. The first type does not contain any spurious statistical signal. The second type contains only signal due to sharing of samples among the tissues, and the number of people shared between pairs of tissues is taken from the GTEx data. Because our goal is to apply RECOV and RE2 to the GTEx data, to avoid data reusing, the SNPs and the gene expressions in both types of datasets are simulated and thus are independent of the values in the GTEx data.

When there is no any spurious statistical association in the data, any alternative hypothesis must be rejected more often than the null hypothesis. We simulate data to demonstrate that RECOV does not inflate false-positive rate in this case. In each simulated dataset, the number of individuals per tissue is taken from the GTEx data but we do not let tissues share individuals. We generate 1,000 SNPs at various minor allele frequency (MAF) without LD, and a random gene expression in each tissue. Between any two tissues, we do not make the expression of the same gene to be correlated. We compute the p-value of the likelihood ratio at each SNP using both RECOV and RE2 model. We repeat this simulation 1,000 times to obtain 1,000,000 p-values each for RECOV and RE2. The histograms of these p-values in both RECOV and RE2 indicate that the null hypothesis is more favored than the alternative hypothesis (Figure 3A,3B).

To measure the effect strictly caused by shared individuals, we simulate datasets as above, but we allow tissues to share individuals. The number of people shared between pairs of tissues is taken from the GTEx data. In each simulation, we compute the likelihood ratio p-values at 1,000 SNPs, and

repeat the simulation 1,000 times to obtain 1,000,000 p-values. We observe that these p-values shift toward 0 when the tissues share samples that are from the same individuals (Figure 3C,3D). In this case, we estimate the RECOV and RE2 GC factor to be 1.2947 and 1.1045 respectively. These GC factors are used to removed the effect caused by shared individuals in tissues that may inflate the false-positive rate. To compute a GC factor, one converts the median of the observed p-values into a chi-square statistic, then finds a multiplying factor to scale this new statistic to a chi-square random variable that has p-value at 0.50 (Devlin and Roeder, 1999).

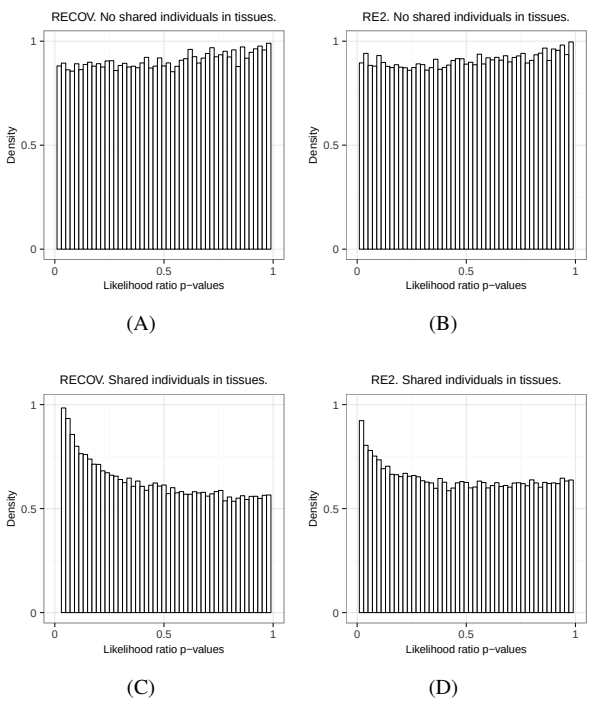


Fig. 3: (A) RECOV and (B) RE2 applied to datasets where the tissues do not share individuals. (C) RECOV and (D) RE2 applied to datasets where the tissues share individuals.

3 Results

3.1 RECOV controls false-positive rate

When using any meta-analysis method to find an eGene, one needs to apply the method at every cis-variant of the specified gene in order to determine if that gene has at least one eQTL. Thus, the global false positive rate (FPR) of RECOV and RE2 depends on the FPR at a single cis-variant. For this reason, we measure the false positive rate (FPR) of RECOV and RE2 at a single variant.

To attain the FPR at one variant, we simulate 1,000 datasets for this single variant under the null hypothesis where this variant is not associated with the gene expression in any of 44 tissues. The MAF of the SNP is randomly chosen and kept the same in all 1,000 datasets. Then in each dataset, the genotype for this SNP and the gene expression are simulated independently of the values in the GTEx data.

To make the simulated data more realistic, we first let tissues have the same number of individuals, and the same amount of samples shared between each pair of tissues, as in the GTEx data. Second, we set expression levels of the same gene from the same individual be correlated with the average correlation of 0.5 across tissues, using the sampling method described in (Sul et al., 2013). This correlation of expression can occur when the tissues of an individual have been exposed to the same environmental factors.

In each of the 1,000 datasets, we estimate the effect size and variance of this single variant on the gene expression in each tissue. RECOV and RE2 take these effect sizes and variances and produce a meta-analysis p-value for this variant. The GC factor estimated in section 2.4.2 is used to transform this p-value in each simulation. This removes only the effect of shared individuals, which is not explicitly modeled in RECOV and RE2. The FPR of this single variant is the fraction of times its transformed p-values are significant.

We repeat this experiment for 1,000 independent variants, so that we have 1,000 measures of FPR for RECOV and RE2. We use the significance level of 0.05 ($\alpha = 0.05$). We find that RECOV attains correct FPR for the majority of variants tested. In RECOV, the median FPR among the 1,000 variants is 0.05 and the 75% and 95% quantiles are 0.06 and 0.09. In RE2, the median FPR is 0.05 and the 75% and 95% quantiles are 0.07 and 0.10. These results demonstrate that RECOV and RE2 control the false positive rates in a realistic setting.

3.2 RECOV discovers more eGenes in GTEx data

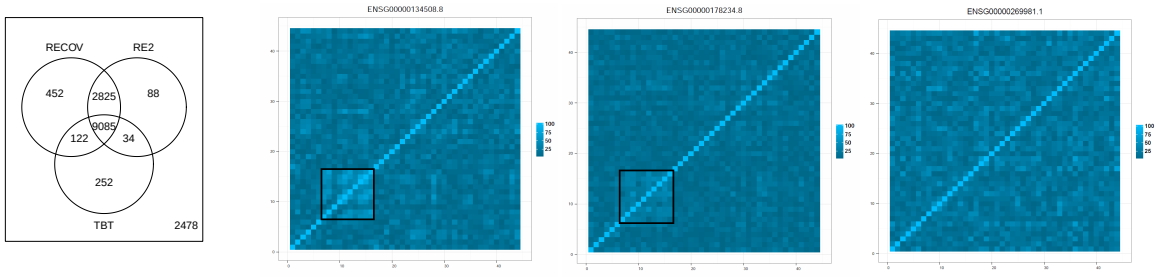
We apply RECOV, RE2, and TBT to the real multi-tissue eQTL dataset from the GTEx consortium. We use GTEx Pilot Dataset V6 released on January 12, 2015. The GTEx consortium has performed RNA-seq on 44 tissues from hundreds of individuals, and we select 15,336 genes that have expression data in all 44 tissues. The consortium has already applied PEER factors to every gene expression in each tissue to remove any batch effects (The GTEx Consortium, 2015). For genotype data, we use GTEx data with imputation that has 5 million genetic variants, all SNPs. Like in the original GTEx pilot study, for each gene, we use its cis-SNPs, which are defined to be located within 1Mb from its transcription start site (The GTEx Consortium, 2015). Not all variants are genotyped in every tissue, because the 44 tissues contain samples from different individuals. We use only cis-variants that are genotyped in all 44 tissues. The median number of cis-variants tested per gene is 3,744.

For each of the 15,336 genes, we apply RECOV, RE2, and TBT to every cis-SNP. For each cis-SNP of a gene, our test statistic is the log likelihood ratio (for RECOV and RE2) or SNP-effect (for TBT). These test statistics are converted into p-values by using a chi-square distribution (for RECOV and RE2) or normal distribution (for TBT). These p-values are then transformed using the GC factors to remove the effect of shared individuals in the tissues. Finally, the most significant p-value among all cis-variants is converted into an eGene p-value by method in 2.4.1 (for RECOV and RE2) or by EGENE-MVN (for TBT).

After computing the eGene p-values for 15,336 genes, we use Bonferroni correction to control for multiple testing correction at 5% level to identify significant eGene p-values; thus each gene has a significance threshold of 0.05/15,336.

Figure 4A shows the Venn diagram of the numbers of eGenes found by TBT, RE2, and RECOV. The majority of tested genes are found to be candidate eGenes. This is expected because there are many tissues tested. It is likely that a gene contains at least one eQTLs in some tissue, which significantly increases the total number of eGenes. Both RE2 and RECOV find more candidate eGenes than TBT. This result agrees with previous findings where applying meta-analysis to multi-tissue datasets yields better outcome than the simple TBT approach (Flutre *et al.*, 2013; Sul *et al.*, 2013).

RECOV detects the highest number of eGenes among the three methods. Out of the 15,336 genes tested, RECOV finds that 81.40% of those genes are eGenes while TBT and RE2 find 61.90% and 78.45% of genes



(A) (B) (C) (D)

Fig. 4: (A) Venn diagram of the numbers of eGenes found by TBT, RE2, and RECOV. (B) The correlation of SNP-effects for the gene ENSG00000134508.8 in 44 tissues (tissue names are omitted). The correlation is computed by using the matrix $\mathbf{B}_g$ in section 2.3.2 where the formula is $\mathbf{B}_g\mathbf{B}_g^\intercal$ (after proper scaling and removal of nearby SNPs). Black box indicates the brain tissues. ENSG00000134508.8 is found to be an eGene by only the RECOV method. The correlation of SNP-effects for gene (C) ENSG00000269981.1 and (D) ENSG00000178234.8 in 44 tissues (tissue names are omitted). ENSG00000269981.1 and ENSG00000178234.8 are found to be eGenes by only the RE2 method.

are eGenes, respectively. This shows that our approach detects 3% more eGenes than RE2 and about 20% more eGenes than TBT.

Next, we consider a case study in each method to understand the circumstances where one method outperforms the other two. We begin with the simple TBT method. In figure 4A, there are 252 genes detected only in the TBT method. Previous publications have reported TBT to be the most powerful option to detect genes with eQTLs that are found in only one tissue (Sul *et al.*, 2013; The GTEx Consortium, 2015). In the TBT method, one analyzes each tissue independently, and is able to determine the number of tissues in which a gene is an eGene. In our result, out of these 252 genes, 225 are eGenes in only 1 tissue, 25 are in 2 tissues, and only 2 are in 3 tissues. This finding agrees with Figure 2 in Sul *et al.* (2013).

Of the 452 genes discovered by only RECOV, the average RECOV eGene p-value is $8.52\text{E}^{-9}(\pm 1.51\text{E}^{-8})$; whereas the average RE2 eGene p-value is $4.18\text{E}^{-3}(\pm 2.85\text{E}^{-2})$. To understand why RECOV discovers genes which are not found by TBT and RE2, consider the protein-coding gene CABLES1 (Gencode id ENSG00000134508.8) which is only detected by RECOV. From the GTEx portal, CABLES1 is expressed mostly in brain tissues, yet it does not have any brain-specific eQTLs. RECOV is a meta-analysis method that pools samples across tissues to increase signals of eQTLs. Thus, it is better than TBT when the sample size per tissue is not large enough so that eQTL signals may be undetected. Unlike RE2, the meta-analysis of RECOV considers correlation of the cis-variants across the tissues; thus RECOV would be better than RE2 if CABLES1 has a consistent correlation pattern. This is indeed the case (Figure 4B). CABLES1's RECOV and RE2 eGene p-value are 4.94E^{-13} and 5E^{-5}, respectively.

Of the 88 genes discovered by only RE2, the average RE2 eGene p-value is $1.15\text{E}^{-8}(\pm 1.44\text{E}^{-8})$; whereas the average RECOV eGene p-value is $1.85\text{E}^{-4}(\pm 2.32\text{E}^{-4})$. We suspect that these 88 genes are genes with eQTLs in multiple tissues. However, due to low sample size, these eQTLs signals may be undetected or do not produce an eGene q-value less than the significance threshold in TBT analysis. As a case study, consider the protein-coding gene GALNT11 (Gencode id ENSG00000178234.8) which is detected by only RE2. Like CABLES1, GALNT11 is expressed mostly in the brain tissues (The GTEx Consortium, 2015). Unlike CABLES1, GALNT11 has eQTL signals in the frontal cortex brain tissue, but these signals produce an eGene q-value of 0.0189 which is higher than the TBT significance threshold. In this case, a meta-analysis approach is more suitable because it combines data from many tissues to improve the eGene p-value. GALNT11's cis-variants have correlated effect sizes across the brain tissues, but this pattern does not stand out from the rest of the tissues when compared to that of CABLES1 (Figure 4C). For this reason, GALNT11's RECOV p-value is higher than its RE2 p-value (3.50E^{-4} vs 7.08E^{-8}). RE2 may also have better performance than RECOV in cases where the cis-variants do not have an obvious correlation pattern across the 44 tissues. As an example, consider the pseudogene RP11-34P13.16 (Gencode id ENSG00000269981.1) which is not tissue-specific (The GTEx Consortium, 2015). The effect sizes of its cis-variants appear to be randomly correlated (Figure 4D), and its RECOV and RE2 p-value are 1.50E^{-4} and 1.37E^{-8}, respectively. Altogether, these attributes may have caused the differences between RECOV and RE2.

4 Discussion

In this paper, we introduce a new random effects meta-analysis method named RECOV that is motivated by the insight that the same SNP may have similar effect on the same gene expression in related tissues. We explicitly model this phenomena by adding a covariance matrix to the existing RE2 model introduced by Han and Eskin (2011). RECOV controls the FPR at a variant level and finds more potential eGenes than RE2 and TBT method when applied to the GTEx data. Using RECOV, we obtain a 3% gain in number of eGenes found compared to RE2, using no additional data.

RECOV scales well to large numbers of tissues compared to previous methods for meta-analysis in gene expression data. For example, MetaTissue and eQTLBma can only handle up to 10 and 20 tissues respectively (Flutre *et al.*, 2013; Sul *et al.*, 2013). RECOV also requires only the summary statistics of the SNP effect on the gene expression in each tissue. These summary statistics are often readily available in gene expression data. Thus, unlike the model by Acharya *et al.* (2016), RECOV requires minimal data preparation.

RECOV, and the RE2 it extends, require optimizing for the two parameters in the log likelihood ratio. These parameters are the mean effect size and the scaling factor for the covariance matrix, and can be estimated by using efficient heuristic methods. We note that the TBT method avoids this optimization. This is a speed-performance trade-off, as it has been shown here and in other papers that meta-analysis approach is better than TBT when applied to multi-tissue data (Flutre *et al.*, 2013; Sul *et al.*, 2013; Acharya *et al.*, 2016). Unlike TBT, RECOV does not provide information about the specific subset of tissues in which the gene is an eGene. This problem is inherent to meta-analysis methods, which only test whether a gene is an eGene in at least one tissue.

Next, we address our use of GC factor for RECOV. The GC factor is traditionally used to correct for inflation due to population structure in classic GWAS but in this paper, we use it to correct for inflation from any unmodeled source. We found that this inflation is due to tissues containing samples from the same individuals. This problem of sample sharing is not the same as the problem of population structure in GWAS (Han and Eskin, 2011, 2012). The value of the GC factor depends on the choice of the covariance matrix $\mathbf{U}_{vg}$ in RECOV. As shown in this paper, when $\mathbf{U}_{vg} = \mathbf{I}$ for RE2 the GC factor is 1.1045, whereas when $\mathbf{U}_{vg} = \mathbf{B}_{-vg}\mathbf{B}_{-vg}^{\intercal}$ for RECOV the GC factor is 1.2947.

RECOV is a general framework for meta-analysis that can be used with any covariance matrix. The choice of matrix used in this paper (described in section 2.3.2) reflects our assumptions about the behavior of the same SNP in different tissues, namely it has correlated effects on the same gene expression across tissues. Although here we present results from one choice of $\mathbf{U}_{vg}$, there are many ways to select the covariance matrix and other types of $\mathbf{U}_{vg}$ may give better results. For example, if we instead assume the same SNP has correlated effects on the expressions of different genes across the tissues, one strategy would be to make $\mathbf{U}_{vg}$ by combining information from neighboring genes of g, using knowledge of from a gene-gene interaction network. The choice of covariance matrix also likely depends on the dataset to which RECOV is applied and may take a very different form in other applications, such as meta-analysis of traditional GWAS. The problem of selecting the most suitable covariance matrix for RECOV is a rich topic for future work.

Funding

D.D., L.G., E.K. and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589 and 1331176, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782 and R01-ES022282. B.H. is supported by the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea (2015-0222) and the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (HI14C1731). E.E. is supported in part by the NIH BD2K award, U54EB020403. J.E. is supported by National Institute of Health grants R01ES024995 and U01HG007912, NSF CAREER Award #1254200, and an Alfred P. Sloan Fellowship (J.E.). The authors declare that they have no competing interests.

References

Abraham, B. and Ledolter, J. (2006). Introduction to regression modeling.

Acharya, C. R., McCarthy, J. M., Owzar, K., and Allen, A. S. (2016). Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. *BMC bioinformatics*, **17**(1), 257.

Albert, F. W. (2016). Brains, genes and power. *Nature Neuroscience*, **19**(11), 1428–1430.

Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, **81**(6), 1158–1168.

Corbeil, R. R. and Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, **18**(1), 31.

Dabney, A., Storey, J. D., and Warnes, G. (2010). qvalue: Q-value estimation for false discovery rate control. *R package version*, **1**(0).

Darnell, G., Duong, D., Han, B., and Eskin, E. (2012). Incorporating prior information into association studies. *Bioinformatics*, **28**(12), i147–i153.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55**(4), 997–1004.

Duong, D., Zou, J., Hormozdiari, F., Sul, J. H., Ernst, J., Han, B., and Eskin, E. (2016). Using genomic annotations increases statistical power to detect eGenes. *Bioinformatics*, **32**(12), i156–i163.

Eskin, E. (2015). Discovering genes involved in disease and the mystery of missing heritability. *Communications of the ACM*, **58**(10), 80–87.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, **9**(5), e1003486.

Fraley, C. and Burns, P. J. (1995). Large-scale estimation of variance and covariance components. *SIAM J. Sci. Comput.*, **16**(1), 192–209.

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., Ruderfer, D. M., Oh, E. C., Topol, A., Shah, H. R., Klei, L. L., Kramer, R., Pinto, D., Gumus, Z. H., Cicek, A. E., Dang, K. K., Browne, A., Lu, C., Xie, L., Readhead, B., Stahl, E. A., Xiao, J., Parvizi, M., Hamamsy, T., Fullard, J. F., Wang, Y.-C., Mahajan, M. C., Derry, J. M. J., Dudley, J. T., Hemby, S. E., Logsdon, B. A., Talbot, K., Raj, T., Bennett, D. A., De Jager, P. L., Zhu, J., Zhang, B., Sullivan, P. F., Chess, A., Purcell, S. M., Shinobu, L. A., Mangravite, L. M., Toyoshiba, H., Gur, R. E., Hahn, C.-G., Lewis, D. A., Haroutunian, V., Peters, M. A., Lipska, B. K., Buxbaum, J. D., Schadt, E. E., Hirai, K., Roeder, K., Brennand, K. J., Katsanis, N., Domenici, E., Devlin, B., and Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, **19**(11), 1442–1453.

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**(4), 1440.

Han, B. and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet*, **88**(5), 586–98.

Han, B. and Eskin, E. (2012). Interpreting meta-analyses of genome-wide association studies. *PLoS Genet*, **8**(3), e1002555.

Han, B., Kang, H. M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*, **5**(4), e1000456.

Han, B., Duong, D., Sul, J. H., de Bakker, P. I. W., Eskin, E., and Raychaudhuri, S. (2016). A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum. Mol. Genet.*, **25**(9), 1857–1866.

Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. *Bioinformatics*, **31**(12), i206–i213.

Huang, Y.-T., VanderWeele, T. J., and Lin, X. (2014). Joint analysis of snp and gene expression data in genetic association studies of complex diseases. *The annals of applied statistics*, **8**(1), 352.

Joo, J. J., Sul, J., Han, B., Ye, C., and Eskin, E. (2014). Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. *Genome Biol*, **15**(4), r61.

Joo, J. W. J., Hormozdiari, F., Han, B., and Eskin, E. (2016). Multiple testing correction in linear mixed models. *Genome Biol*, **17**(1).

Kang, E. Y., Han, B., Furlotte, N., Joo, J. W., Shih, D., Davis, C. R., Lusis, J. A., and Eskin, E. (2014). Meta-analysis identifies gene-by-environment interactions as demonstrated in a study of 4,965 mice. *PLoS Genet*, **10**(1), e1004022.

Liu, G., Hu, Y., Jin, S., Zhang, F., Jiang, Q., and Hao, J. (2016). Cis-eQTLs regulate reducedLST1gene andNCR3gene expression and contribute to increased autoimmune disease risk: Table 1. *Proceedings of the National Academy of Sciences*, **113**(42), E6321–E6322.

Nieuwenhuis, M. A., Siedlinski, M., van den Berge, M., Granell, R., Li, X., Niens, M., van der Vlies, P., Altmüller, J., Nürnberg, P., Kerkhof, M., van Schayck, O. C., Riemersma, R. A., van der Molen, T., de Monchy, J. G., Bossé, Y., Sandford, A., Bruijnzeel-Koomen, C. A., van Wijk, R. G., ten Hacken, N. H., Timens, W., Boezen, H. M., Henderson, J., Kabesch, M., Vonk, J. M., Postma, D. S., and Koppelman, G. H. (2016). Combining genomewide association study and lung eQTL analysis provides evidence for novel genes associated with asthma. *Allergy*.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**(398), 605–610.

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**(6), e1003491.

Sul, J. H., Raj, T., de Jong, S., de Bakker, P. I., Raychaudhuri, S., Ophoff, R. A., Stranger, B. E., Eskin, E., and Han, B. (2015). Accurate and fast multiple-testing correction in eQTL studies. *The American Journal of Human Genetics*, **96**(6), 857–868.

The GTEx Consortium (2015). The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**(6235), 648–660.

Thompson, S. G. and Sharp, S. J. (1997). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statist. Med.*, **18**(3), S82.