Tsinghua Science and Technology

Volume 19 | Issue 6 Article 7

2014

Methods for Population-Based eQTL Analysis in Human Genetics

Lu Tian
Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, NC 28223, USA.

Andrew Quitadamo
Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, NC 28223, USA.

Frederick Lin
Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, NC 28223, USA.

Xinghua Shi
Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, NC 28223, USA.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 07/10 pp624-634
Volume 19, Number 6, December 2014

Methods for Population-Based eQTL Analysis in Human Genetics

Lu Tian, Andrew Quitadamo, Frederick Lin, and Xinghua Shi*

Abstract: Gene expression is a critical process in biological system that is influenced and modulated by many factors including genetic variation. Expression Quantitative Trait Loci (eQTL) analysis provides a powerful way to understand how genetic variants affect gene expression. For genome wide eQTL analysis, the number of genetic variants and that of genes are large and thus the search space is tremendous. Therefore, eQTL analysis brings about computational and statistical challenges. In this paper, we provide a comprehensive review of recent advances in methods for eQTL analysis in population-based studies. We first present traditional pairwise association methods, which are widely used in human genetics. To account for expression heterogeneity, we investigate the methods for correcting confounding factors. Next, we discuss newly developed statistical learning methods including Lasso-based models. In the conclusion, we provide an overview of future method development in analyzing eQTL associations. Although we focus on human genetics in this review, the methods are applicable to many other organisms.

Key words: expression Quantitative Trait Loci (eQTL) analysis; confounding factors; sparse learning models; Lasso

1 Introduction

Gene expression is a critical process in biological system that is influenced and modulated by many factors including genetic variation. Genetic variation reflects the genetic difference among individuals in human population. Such variation can be at different levels, ranging from Single Nucleotide Polymorphisms (SNPs) to structural variation including Copy Number Variants (CNVs). Assessing the effect of genetic variation on gene expression will improve our knowledge in understanding how genetic variation leads to phenotypic variation including its contribution to human health and disease.

Expression Quantitative Trait Loci (eQTL) analysis[1-4] provides a powerful way to understand how genetic variants affect gene expression. Specifically, eQTL analysis treats gene expressions as quantitative molecular phenotypes, and intends to find genetic variants whose genotypes are significantly associated with changes in gene expression. These identified associations can help reveal biochemical processes underlying living systems, discover genetic factors, and pathways that cause disease[5].

In recent years, high-throughput technologies including microarrays and sequencing, have enabled the identification of genetic variation and the quantification of gene expression at whole genome level. Such data has provided a rich resource for eQTL analysis, for both biomedical application and methodology development. For genome wide eQTL analysis, the number of genetic variants and that of genes are large and thus the search space is tremendous. For example, there could be millions of SNPs and over twenty thousand genes in a genome wide eQTL analysis in humans. The search space could easily reach to the scale of 10^9. Consequently, it is yet another "needle-in-a-haysack" problem to seek for a small number significant associations from this vast search space.

• Lu Tian, Andrew Quitadamo, Frederick Lin, and Xinghua Shi are with the Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, NC 28223, USA. E-mail: {ltian, aquitada, flin8, x.shi}@uncc.edu.
* To whom correspondence should be addressed.
Manuscript received: 2014-06-17; accepted: 2014-06-24

Therefore, eQTL analysis brings about computational and statistical challenges.

In practice, there are generally two settings for eQTL analysis in human genetics. The first setting is in family-based studies, where related individuals from a set of family trees are investigated. The second setting is for population studies, where a collection of unrelated individuals is assessed. In this paper, we will provide a comprehensive review of recent advances in methods for eQTL analysis in population-based setting.

In Section 2, we will focus on traditional pairwise association methods, which are widely used in human genetics. We will discuss the methods for correcting confounding factors in heterogeneous gene expression data in Section 3. Next, we will review newly developed statistical learning methods including Lasso in Section 4. In Section 5, we will conclude with a discussion for future method development in analyzing eQTL associations. Although we focus on human genetics in this review, the methods are applicable to many other organisms.

2 Pairwise Association Methods

2.1 Pairwise association procedure

In eQTL analysis, pairwise association methods perform a correlation or regression analysis on the genotype (or probe intensity values) of a genetic variant and the expression of a gene across different individuals. The assumption is that genetic variants are independent of each other, and gene expressions are also independent among the genes. Thus, we can perform association analysis for each pair of variant and gene separately. Multi-test correction will then be applied to correct for the bias introduced by the large number of tests performed separately.

Figure 1 illustrates such pairwise associations between a pair of genetic variant and gene investigated in the same individuals. Suppose we investigate J genetic variants and K genes in n individuals. Specifically, we have a vector $\boldsymbol{x}_j$, $j \in \{1, \cdots, J\}$, for a genetic variant of v_j, which contains its genotype (or probe intensity) in n individuals. Similarly, we use another vector $\boldsymbol{y}_k$, $k \in \{1, \cdots, K\}$, to denote the expression values of a gene g_k with index k in the corresponding n individuals. For each (v_j, g_k) pair, we can perform a correlation (e.g., Pearson correlation or Spearman rank correlation) or linear regression analysis between $\boldsymbol{x}_j$ and $\boldsymbol{y}_k$. This analysis produces a nominal p-value and a correlation noted as r-value.

Afterwards, multi-test correction technique is performed to select significant associations. Common methods to correct for multiple tests include FamilyWise Error Rate (FWER) procedures, random permutations, and False Discovery Rate (FDR) controlled approach. FWER procedures such as the Bonferroni correction tends to be too conservative in practice. In practice, less stringent methods such as random permutation[4, 6-9] or FDR controlled[10-12] approaches are used.

Random permutation based multi-test correction randomly permutes the expression values many times (e.g., 10 000 times). A permutation p-value threshold (e.g., 0.01) is then used to define significant eQTL associations. The most widely used FDR controlled multi-test correction is based on Benjamini-Hochberg procedure[13].

The genetic variants investigated can be of different types, ranging from SNPs to CNVs. Although the majority of eQTL studies are focused on SNPs[4, 7-10, 12, 14-16], recent studies have started CNV eQTL analysis[6, 11] and found that CNVs contribute significantly to gene expression variation. Dependent on if intensity values for probes or reads are used or variant genotypes are used, the identified associations can be visualized using scatterplots or box plots. Figure 2 shows two examples of identified eQTL associations.

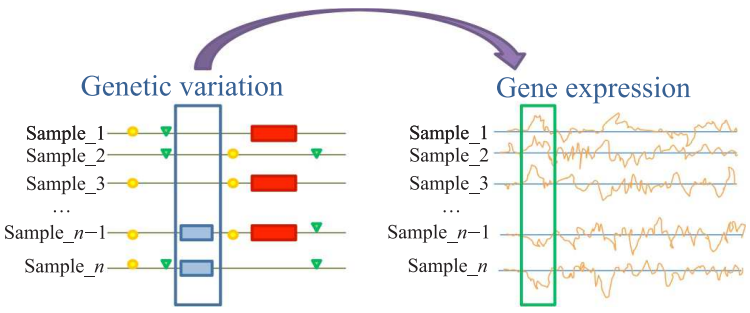


Fig. 1 Pairwise associations between genetic variants and gene expressions in eQTL analysis.

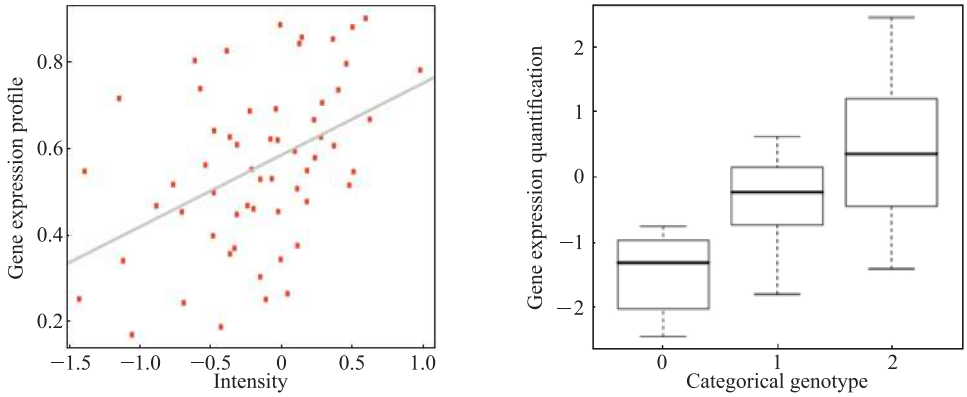


Fig. 2 Examples of significant eQTL associations based on (a) intensity values and (b) variant genotypes.

In panel (a), X axis represents the intensity values for genetic variants, Y axis shows the corresponding gene expression profiles in the same individuals, and the gray line is the linear regression line showing the association. In panel (b), X axis stands for categorical genotypes of genetic variants and Y axis represents the gene expression quantifications in those individuals. Note that the effect that a genetic variant can affect on a gene expression can be negative or positive, which will be reflected by a negative or positive correlation value, respectively.

2.2 *Cis* and *trans* associations

In general, there are typically two categories of eQTL associations that have been investigated, namely *cis* (i.e., local) and *trans* (i.e., remote) associations. Such characterization is mainly based on the distance boundary between genetic variants and a target gene. Cis associations focus on the genetic variants that lie within a window around the midpoint (or transcription start site) of any given gene. Many studies use a distance cutoff of 1 Mb[4,6-10], where variants located within 1 Mb window upstream and downstream of a particular gene are considered for *cis* associations. Other distance cutoffs have also been used such as 100 kb[12], 200 kb[11] or 2.5 Mb[15]. *Trans* associations assess the genetic variants that include genetic variants that map beyond the defined window of the gene on the same chromosome or on a different chromosome from the target gene.

Since *trans* associations require more tests to be performed than *cis* associations do, the multiple testing burden increases for *trans* associations. Thus the power for identifying *trans* associations is lower that that of *cis* associations. Most published studies have focused on *cis* associations only[4,6-9,11,12], although there are some studies include both *cis* and *trans* associations[10,15,16].

2.3 Tools for pairwise eQTL analysis

Over recent years, there have been a wide range of tools made publicly available for eQTL analysis. Although the technical details are different, we see a general trend that methods have been improved to enable large scale eQTL analysis with a large number of variants investigated. Here, we review several tools that are publicly available to the community.

Merlin[17] provides a suite of tools for linkage and association studies, including eQTL analysis. It uses sparse trees, a reduced representation of the full binary tree, to present gene flow in pedigrees. Merlin supports eQTL analysis in population studies as well as in family-based analysis. Merlin does not control for population stratification. If population stratification is a concern for eQTL analysis, population membership can be incorporated as a covariate. Another option is to perform population stratification before running eQTL analysis, or adjust eQTL results based on genomic control methods.

R/qtl[18] is an R package that provides eQTL analysis, with computationally intensive algorithms coded in C. The core technology R/qtl applies is Hidden Markov Model (HMM)[19]. It uses Haley-Knott regression[20] to perform the correlation analysis. snpMatrix[21] is another R package developed for genome-wide association studies and can be used for eQTL analysis.

PLINK[22] provides an integrated environment for genome-wide association studies and family-based studies, with support of eQTL analysis. The main functions of PLINK consist of five aspects, including data management, summary statistic, population

stratification, association analysis, and identity-by-descent estimation. PLINK uses a compact binary file including SNP data as input data for analysis. PLINK provides summary statistics of SNPs, including genotyping rates, allele and genotype frequency, and Hardy-Weinberg equilibrium tests. For family-based studies, single-SNP Mendelian error is also available. In addition, individual heterozygosity rates and sex-check can also be obtained for each individual. PLINK implemented two methods to deal with the population stratification issue. PLINK estimates the average proportion of alleles shared Identical By State (IBS) between any two individuals and then classifies individuals into homogenous clusters. Alternatively, PLINK can use the data-reduction technique of classical multi-dimensional scaling to produce a k-dimensional representation of any substructure. PLINK can perform several statistical tests for case/control studies, and transmission/disequilibrium test for family-based studies. PLINK can calculate inbreeding coefficients as indicator of inherited identical by descent for each individual, using panels of common SNPs.

FastMap[23] is designed for fast eQTL mapping in population-based studies. It achieves fast calculation of test statistic tests (correlations) by efficiently calculating subset sums using the Subset Summation Tree. FastMap was originally developed for the use with inbred mouse strains.

Matrix eQTL[24] is a new addition to the eQTL analysis tools, which provides fast and efficient eQTL mapping. Comparing to previous methods described above (Merlin[17], R/qtl[18], snpMatrix[21], PLINK[22], FastMap[23]), matrix eQTL is much faster and thus can be applied to large datasets. Matrix eQTL utilizes matrix operations that include the following methods: simple linear regressions that do not include covariates and assume uncorrelated homoskedastic errors, ANOVA testing, additive covariates, and heteroskedastic and correlated errors. All the three versions of matrix eQTL (Matlab, Revolution R and R, Goto BLAS) did in fact outperform the others by preprocessing and expressing the most computationally intensive part of the algorithm in terms of large matrix operations. Also, Matrix eQTL computation time remains nearly unchanged when covariates are added to the model. Note matrix eQTL supports covariates for linear regression models. Users can explicitly describe known factors that are not genetic as covariates.

3 Confounding Factors

Expression heterogeneity due to technical, genetic, environmental, or demographic variables is common in gene expression studies. Failing to incorporate these sources of heterogeneity into an analysis can reduce power or induce unwanted dependence across gene, and also introduce spurious signals. Since there are diverse known and unknown confounding factors that contribute to heterogeneity in gene expression, it is critical to take care consideration of these confounding factors for eQTL analysis. Known confounding factors include batch effect, GC content bias, population structure, gender, age, and others. Unknown (hidden) confounding factors can come from unmeasured variables such as temperature, concentration or environmental variation.

To identify, estimate, and incorporate expression heterogeneity, it is commonly used to apply dimension reduction methods, such as Principal Components Analysis (PCA) and factor analysis, to dissect known and unknown confounding factors before performing eQTL analysis. The top PCs, that represent contribution from confounders, are removed or used as covariates for eQTL analysis[10]. Specifically, the PCA procedure is composed of the following iterative steps: (1) Perform eQTL associations on the original expression matrix and obtain N_0 eQTL associations. (2) Apply PCA on the expression matrix and obtain the top PC and residual matrix. (3) Use the residual matrix to perform eQTL analysis. Denote the number of eQTL associations as N_1. In this process, the top PC can be used as covariate or be completely removed from eQTL analysis. (4) If $N_1 > N_0$, repeat the process of (1)-(3) on the residual expression matrix. (5) Continue steps (1)-(4) until reaching an residual expression matrix with top i PCs removed, where $N_i > N_{i-1}$ and $N_i > N_{i+1}$. At this point, we decide that we will use the residual expression matrix with top i PCs removed for eQTL associations, and this selection will give us the optimal power by removing the i confounding factors.

Similar to PCA, another dimension reduction method to account for confounders is called factor analysis. One of such methods is Surrogate Variable Analysis (SVA)[25]. SVA uses the expression data itself to identify groups of genes affected by each unobserved factor and estimates the factor based on the expression of those genes. The SVA algorithm can conceptually be broken down into four basic steps: (1)

Remove the signal due to the measured variable(s) of interest to obtain a residual expression matrix. Apply decomposition to the residual expression matrix to identify signatures of expression heterogeneity in terms of an orthogonal basis of singular vectors. Use a statistical test to determine the singular vectors that represent significantly more variation than expected. (2) Identify subset of genes driving each orthogonal signature of expression heterogeneity through a significance analysis of associations between genes and signatures for expression heterogeneity on the residual expression matrix. (3) For each subset of genes, build a surrogate variable based on the full expression heterogeneity signature of that subset in the original expression data. (4) Include all significant surrogate variables as covariates in subsequent regression analyses, allowing for gene-specific coefficients for each surrogate variable. The author applied SVA to disease class, time course, and genetics of gene expression studies. The results showed that SVA increased the biological accuracy and reproducibility of analyses in genome-wide expression studies.

Traditional methods including PCA and SVA, are able to correct for strong confounding effects, based on the previous procedures that are optimized empirically.

4 Statistical Learning Methods for eQTL Analysis

Other than pairwise association studies in eQTL analysis, new methods have been developed to apply statistical learning methods to identify significant eQTL associations. In this section, we review some of these advanced methods including new statistical methods for correcting confounding factors along with eQTL analysis, Lasso-based models, and graphical models.

4.1 Statistical methods for eQTL analysis with confounding factor correction

Combined with the aforementioned methods in Section 3 for correcting confounding factors in expression data, we can use the tools discussed in Section 2.3 to identify significant eQTL associations. Therefore, eQTL analysis includes two separate components, that is, estimating confounder factors first, and then followed by eQTL analysis. However, new methods have developed[26, 27] to streamline the correction and association processes in the same toolkit.

The study from Ref. [26] provides a method for correcting confounding factors that may complicate eQTL analysis by causing spurious associations including spurious regulatory hotspots. The authors argue that it is the intersample correlation structure inherent in expression data that leads to spurious associations between genetic variants and gene expression that induce spurious regulatory hotspots. The authors develop a statistical method, the Intersample Correlation Emended (ICE) eQTL mapping, that corrects for the spurious associations caused by complex intersample correlation of expression measurements in eQTL mapping.

ICE eQTL mapping performs eQTL analysis using a linear mixed model as shown in Eq. (1). $\boldsymbol{Y}$ is the gene expression matrix and $\boldsymbol{X}$ is the genetic variation matrix. $\boldsymbol{U}$ represents the unknown confounding factors. $\boldsymbol{E}$ is a random variable that represents random noise with a Gaussian white-noise term. $\boldsymbol{Z}$ and $\boldsymbol{C}$ represent the fixed effects of known confounding factors and their coefficients. $\boldsymbol{B}$ is the contribution of genetic variation matrix $\boldsymbol{X}$ to $\boldsymbol{Y}$. F-test is applied to test for the significance of $\boldsymbol{B}$, using the Efficient Mixed-Model Association (EMMA) package for rapid estimation of variance components and maximum likelihood to perform likelihood tests[28].

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{ZC} + \boldsymbol{U} + \boldsymbol{E} \tag{1}$$

Therefore, ICE eQTL directly incorporates the complex correlation structure into the statistical model as a variance component accounting for random effects. It does not need to know in prior which confounding factors to be corrected.

In contrast to SVA[25], ICE can correct for a mixture of strong and moderate confounding effects. It has been shown that ICE eQTL mapping can reduce false positive *trans* associations and increase the number of true *cis* associations, due to its ability to correct for systematic confounding effects inherent in expression datasets.

This new method called LMM-EH-PS[29] aims to correct for population structure in addition to take consideration of expression heterogeneity. LMM-EH-PS is based on a linear mixed model as shown in Eq. (2). Similar to ICE, LMM-EH-PS assumes that the confounder coefficients are drawn from a zero-mean Gaussian distribution as well.

One can simultaneously correct for population structure and expression heterogeneity simultaneously, by simply generating two sets of confounder coefficients independently, using the similarities

between individuals in respectively genetic variation and expression data. Then these two confounder coefficients are added to the regression model to produce a likelihood for one pair of genetic variant and one gene. Since ICE uses the covariance matrix of the gene expression data to model gene heterogeneity, ICE yields deflated p-values[29]. The problem of deflated p-values is solved in LMM-EH-PS by setting the contribution to be tunable in Ref. [29].

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{ZC} + \boldsymbol{E} \tag{2}$$

Another method to correct for confounding factors is based on Probabilistic Estimation of Expression Residuals (PEER) method[30, 31]. PEER uses Bayesian approaches on factor analysis methods to infer hidden determinants from the normalized and preprocessed expression profiles, considering all known covariates. Factors inferred are then used in alternative genetic analyses, as phenotypes in genetic mapping or as covariates for association analysis of other phenotypes. With the ability of learning unmeasured determinants of gene expression variation and cellular features from gene expression, PEER can increase the power of detecting genetic determinants of gene expression profiles. In addition, genetic factors identified can be interpreted as pathway or transcription factor activations by combining with prior biological information, which increases the interpretability of gene expression analyses. Note that PEER requires a larger number of sample than the expected number of factors to be learned. PEER does not support for mixed modeling nor account for population structure, but can incorporate them as covariates. Due to its demonstrated power, PEER has been used extensively[7, 9, 10, 14] for correcting expression heterogeneity in eQTL studies.

In addition to correcting for confounding factors, Ref. [30] provides a software VBQTL, a probabilistic method for dissecting gene expression variation by jointly modeling the underlying global causes of variability (including known and unknown confounder factors) and the genetic effect. VBQTL implements a Bayesian method that captures Eq. (1). It supports model extension where the right hand side of Eq. (1) can include more terms that take consideration of non-linear effects such as epistasis.

The commonly used methods[25, 26, 29-31] accounting for unknown confounding factors we have reviewed are based on an assumption that confounding factors tend to have broad effects, influencing large fractions of gene expression measurements. However, *trans* genetic effects exhibit similar association pattern as confounding factors using existing methods, which may result in over-correction and reduce the detection of true signal of *trans* effects.

In Ref. [27], the authors developed an integrated probabilistic model PANAMA (Probabilistic ANAlysis of genoMic dAta) that considers both the potential hidden confounding factors and genetic regulators. Similar to Refs. [26, 30], PANAMA performs eQTL analysis using a linear mixed model as shown in Eq. (3). PANAMA is different from other methods in that it jointly learns confounding factors while accounting for the effect of genetic variants with a pronounced *trans* regulatory effect. Therefore, PANAMA avoids overlaps between true eQTL associations and the covariance structure induced by the learnt confounders. It has been shown[27] that PANAMA has better performance in identifying the hidden confounding factors, compared with standard linear regression, SVA[25], ICE[26], and PEER[30, 31]. Such improvement leads to increased power for PANAMA to detect both *cis* and *trans* genetic effects.

4.2 Lasso-based models

Sparse modeling, a feature selection method widely used in the machine learning community, has been recently applied to select eQTL associations in eQTL analysis. In this setting, genetic variants are treated as features and gene expressions are viewed as labels. In eQTL analysis, both the feature matrix on genetic variants and the label matrix on gene expressions are usually high-dimensional. That is, the number of features (i.e., genetic variants) and the number of labels (i.e., gene expressions) are significantly larger than the number of samples. Therefore, the problem of eQTL analysis can be formed to a classical feature selection problem and spare learning methods have therefore proposed.

The sparsity of these methods are justified with an assumption that there are only a small number of associations between genetic variants and traits, given an overwhelmingly large number of variant and gene pairs for genome wide eQTL analysis. These methods either separately examine if the correlation of each pair of traits and genetic variants is significant or characterize their association as parameters in a machine learning model.

For simplicity, the eQTL analysis can be formalized

into a linear regression model in Eq. (3). Here we assume that the expression data has been corrected for known and unknown confounding factors using methods proposed in Section 3 or Section 4.1. Note that we can modify the model to include confounding factors by extending the linear model in Eq. (3) to linear mixed model in Eq. (2).

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E} \tag{3}$$

The method of sparse modeling is to find the optimal matrix $\boldsymbol{B}$ with J rows and K columns, by minimizing the square loss function plus a regularization term in Eq. (4). Here, $\boldsymbol{X}$ is an $n \times J$ matrix, where each row contains the measurements of the J genetic variants in that individual, and each column contains n observations for one genetic variant. $\boldsymbol{Y}$ is an $n \times K$ matrix, where each row contains the quantitative profiles of K genes in that individual, and each column contains n observations for one gene. $\boldsymbol{B}_{J\times K} = \{\boldsymbol{b}^1, \cdots, \boldsymbol{b}^K\}$ is the association coefficient matrix denoting the connection strengths between traits and genetic variants and $\boldsymbol{E}$ is a Gaussian white-noise term with constant variance σ^2. The coefficient matrix $\boldsymbol{B}$ represents the association value between every genetic variant and each gene. Any non-zero value in $\boldsymbol{B}$ indicates that the genetic variant is associated with the corresponding gene, with a weight determined by the value.

$$L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{XB}\|_2 + \lambda \|\boldsymbol{B}\|_q \tag{4}$$

Lasso (least absolute shrinkage and selection operator)[32] method is a special case with q=1. With l_1 norm, a Lasso model favors a sparse solution with a small number of non-zero terms[33]. In a Lasso model for eQTL analysis, the parameters of the majority of the genetic variants are shrunk to zeros, and those variants corresponding non-zero terms are selected as the identified eQTL associations.

However, both pair-wise correlation methods in Section 2.3 and a Lasso model neglect the relationship among genetic variants or that among the expression levels of multiple genes. These methods assume that genetic variants are independent and gene expressions are not correlated. This assumption will inevitably miss many complex yet observed cases where multiple genetic variants jointly affect the co-expressions of multiple genes.

To account for relatedness of different genes or traits, multi-task Lasso models are proposed[34-36] to impose sparsity over all of the variant and gene pairs. For high-dimensional problems where K and J are large, Eq. (3) is well-posed only if certain regularization is introduced. A standard multi-task Lasso model can be shown as Eq. (5). Here, $\boldsymbol{b}^k$ is the k-column of $\boldsymbol{B}$ representing the association coefficients of all genetic variants to the k-th trait and $\boldsymbol{b}_j$ is the j-th row of $\boldsymbol{B}$ meaning the association strengths of the j-th genetic variant to all traits.

$$\text{minimize}_B \sum_{k=1}^{K} \|\boldsymbol{y}^k - \boldsymbol{Xb}^k\|_2^2 + \lambda \sum_{j=1}^{J} \delta_j \|\boldsymbol{b}_j\|_2 \tag{5}$$

Several extensions of the multi-task Lasso model have been proposed to take consideration of the network structure underlying the relatedness of genes. For example, based on the idea that co-expressed genes share a larger common set of genetic variants comparing those independent genes, a tree guided Lasso model[35] is developed. In this model, the tree structure can be obtained using a hierarchical clustering tree on labels or provided by users. Note that the tree guided Lasso is an extension of a group-Lasso model. Additionally, an adaptive multi-task Lasso model[36] considers the correlation among gene expressions, while incorporating the priors on SNPs such as regulatory features for these SNPs.

Another Lasso-based model, a graph-guided multi-task Lasso model[34,37] can also be used to estimate genetic variants that perturb a subset of highly correlated genes. Let $G_1 = (V_1, E_1)$ be a weighted graph that represents the relatedness of the genes, where V_1 is the set of vertices and E_1 is the set of edges. Genes are vertices and edges represent the relatedness between two genes in G_1. The idea here is that if two genes are correlated and an genetic variant is associated with one of these two genes, then the probability that this variant is associated with the other gene is higher.

We can formulate this idea by adding a regularization term to Eq. (5). As seen in Eq. (6), a regularization term is added to reflect the contribution from correlated genes, with a structure dictated by a graph G_1. Here, $w(e_{m,l})$ is a weight assigned to the edge $e_{m,l}$ in graph G_1 and $r_{m,l}$ is the correlation between $\boldsymbol{y}^m$ and $\boldsymbol{y}^l$. Then a graph-guided multi-task Lasso can be defined as the solution of Eq. (6). Such a graph guided multi-task model in Eq. (6) can learn the associations between one particular genetic variant and a group of correlated genes. The associations among genetic variants and genes will be reflected by $\boldsymbol{B}$ matrix.

The regularization term in Eq. (6) is closely related to

that in fused Lasso[38]. Thus, a graph-guided multi-task Lasso model[34, 37] can be viewed as a generalization of the fused Lasso model in that fusion is dictated by the topology of input graphs, rather than physical proximity.

$$\text{minimize}_{\boldsymbol{B}} \quad \sum_{k=1}^{K} \|\boldsymbol{y}^k - \boldsymbol{X}\boldsymbol{b}^k\|_2^2 + \lambda\|\boldsymbol{B}\|_1 + \gamma \sum_{e_{m,l}\in E_1} w(e_{m,l}) \sum_{j=1}^{J} |b_{jm} - \text{sign}(r_{m,l})b_{jl}| \quad (6)$$

The graph-guided multi-task Lasso model[34, 36, 37] can be further extended to incorporate the correlation among genetic variants. The rationale is that if one genetic variant is associated with the expression of a specific gene, then another genetic variant, that is highly related with this genetic variant, would be more likely to be associated with the expression of this particular gene. This rationale can be formalized as adding another regularization term to Eq. (6), which leads to a two-graph guided multi-task Lasso model as proposed in Ref. [39].

The two-graph guided multi-task Lasso model can be then learned by minimizing the objective function in Eq. (7). The optimization problem in Eq. (7) can be efficiently solved by a coordinate-descent algorithm as in Ref. [34]. Here, $G_2 = (V_2, E_2)$ is a graph that represents the correlation among genetic variants. V_2 is the set of vertices representing genetic variants, and E_2 is the set of edges representing the correlation among these variants. $w(e_{f,g})$ is a weight assigned to the edge $e_{m,l}$ in graph G_2 and $r_{f,g}$ is the correlation between $\boldsymbol{y}^f$ and $\boldsymbol{y}^g$.

$$\text{minimize}_{B} \quad \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + \lambda\|\boldsymbol{B}\|_1 + \gamma_1 \sum_{e_{m,l}\in E_1} w(e_{m,l}) \sum_{j=1}^{J} |b_{jm} - \text{sign}(r_{m,l})b_{jl}| + \gamma_2 \sum_{e_{f,g}\in E_2} w(e_{f,g}) \sum_{k=1}^{K} |b_{fk} - \text{sign}(r_{f,g})b_{gk}| \quad (7)$$

The two-graph guided multi-task Lasso model captures the scenario that multiple genetic variants, by forming a subnetwork, can affect the expression of multiple correlated genes. Such a model has a nice sparse property that it allows flexible structured sparsity both on genetic variants and genes. The two-graph guided multi-task Lasso model can be seen as a generalization of several multi-task feature selection methods proposed before[34-37].

Another model, named jointly structured input-output Lasso model[40], has been developed to capture the association between a set of genetic variants and a set of correlated genes. The model implements an l_1/l_2 regularized multi-task regression[41]. The structured input-output Lasso model can be achieved by solving the following optimization problem.

$$\text{minimize}_{\boldsymbol{B}} \quad \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + \lambda_1\|\boldsymbol{B}\|_1 + \gamma_1 \sum_{k=1}^{K} \sum_{g\in V_1} \|b_k^g\|_2 + \gamma_2 \sum_{j=1}^{J} \sum_{h\in V_2} \|b_k^h\|_2 \quad (8)$$

The last two terms in Eq. (8) reflect the contribution from the groups, denoted as V_1, of genetic variants, and the contribution from the groups of genes as defined as V_2, respectively.

All of the methods described so far in Section 4.2 have been focused on applying Lasso to linear models. More recently, a new method has been proposed that combines mixed linear model, as shown in Eq. (2), and Lasso for eQTL analysis[42]. This method, termed as LMM-Lasso, can thus correct for population structure and capture associations in a Lasso model in eQTL analysis. LMM-Lasso scans all SNPs at the same time while accounting for their interdependencies to assess their associations with gene expression.

5 Conclusions

In this paper, we review recent method development in estimating the contribution of genetic variation to gene expression, under the overarching framework of eQTL analysis. With rapid accumulation of genetic variation and gene expression data in human genetics, it is demanding to design efficient and fast methods to perform such data intensive analysis.

Although pairwise association methods still dominate the current practice of eQTL analysis, we witness an emerging trend to develop more statistical models that capture not only the confounding factors, but also complex relationship in human genetics data. Traditional methods, including PCA and SVA, have been used routinely to correct for strong confounding effects. Newly developed methods like PEER have demonstrated their increased power in taking account of confounders.

In addition to pairwise association analysis, machine learning methods, including sparse learning models, have been recently applied to eQTL analysis. In particular, Lasso models have been widely studied

to identify a small set of eQTL associations due to its sparsity. In this paper, we discuss a suite of Lasso-based methods to account for the correlation of genetic variants and/or the relatedness of different genes. Extensions of Lasso methods have also been applied to look at complex relationships among eQTLs including redundancy[43]. Besides the sparse methods in the Lasso family, graphical models[44-46] have been proposed for eQTL analysis.

Previous integrative analysis of eQTL methods has shown the gain of power to use ensemble of machine learning methods for eQTL analysis[47-49]. With the further development in applying statistical and machine learning approaches to the problem of identifying eQTL associations, we expect the ensemble strategy will gain more attention in future studies.

The methods we present in this paper has been designed to estimate eQTL associations in one specific tissue. Gene expression is differentially regulated across tissues. Hence, eQTLs are tissue specific. A recent study[14] has collected genomic and expression information in different tissues and performed eQTL analysis for these tissues independently. New methods have been developed for cross-tissue eQTL analysis that incorporates data from multiple tissues to improve power for eQTL analysis[50,51]. With more tissue-specific data available, we anticipate more methods will be available to assess eQTL associations shared and specific in different tissues. Such analysis will provide more information to understand genetic variation contributes to differential gene expression changes across tissues, and leads to phenotypic variation including disease susceptibility and manifestation.

Acknowledgements

The authors would like to thank anonymous reviewers for their valuable comments and suggestions. This work was supported in part by a Faculty Research Grant from the University of North Carolina at Charlotte.

References

[1] M. V. Rockman and L. Kruglyak, Genetics of global gene expression, *Nat. Rev. Genet*., vol. 7, pp. 862-872, 2006.

[2] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, Mapping complex disease traits with global gene expression, *Nat. Rev. Genet.*, vol. 10, no. 3, pp. 184-194, 2009.

[3] V. G. Cheung and R. S. Spielman, Genetics of human gene expression: Mapping DNA variants that influence gene expression, *Nat. Rev. Genet.*, vol. 10, no. 9, pp. 595-604, 2009.

[4] B. E. Stranger, M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch, R. Lyle, S. Hunt, B. Kahl, S. E. Antonarakis, S. Tavar, P. Deloukas, and E. T. Dermitzakis, Genome-wide associations of gene expression variation in humans, *PLoS Genet*., vol. 1, no. 6, p. e78, 2005.

[5] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, et al., Relative impact of nucleotide and copy number variation on gene expression phenotypes, *Science*, vol. 315, no. 5813, pp. 848-853, 2007.

[6] A. Schlattl, S. Anders, S. M. Waszak, W. Huber, and J. O. Korbel, Relating CNVs to transcriptome data at fine resolution: Assessment of the effect of variant size, type, and overlap with functional regions, *Genome Res*., vol. 21, no. 12, pp. 2004-2013, 2011.

[7] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, *Series B*, vol. 57, no. 1, p. 289300, 1995.

[8] GTEx Consortium, The Genotype-Tissue Expression (GTEx) project, *Nat. Genet*., vol. 45, no. 6, pp. 580-585, 2013.

[9] B. E. Stranger, S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, et al., Patterns of cis regulatory variation in diverse human populations, *PLoS Genet*., vol. 8, no. 4, p. e1002639, 2012. doi: 10.1371/journal.pgen.1002639.

[10] Q. Li, J. H. Seo, B. Stranger, A. McKenna, I. Pe'er, T. Laframboise, M. Brown, S. Tyekucheva, and M. L. Freedman, Integrative eQTL-based analyses reveal the biology of breast cancer risk loci, *Cell*, vol. 152, no. 3, pp. 633-641, 2013.

[11] S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis, Transcriptome genetics using second generation sequencing in a Caucasian population, *Science*, vol. 464, no. 7289, pp. 773-777, 2010.

[12] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J. B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard, Understanding mechanisms underlying human gene expression variation with RNA sequencing, *Science*, vol. 464, no. 7289, pp. 768-772, 2010.

[13] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. 't Hoen, J. Monlong, M. A. Rivas, M. González-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, et al., Transcriptome and genome sequencing uncovers functional variation in humans, *Nature*, vol. 501, no. 7468, pp. 506-522, 2013.

[14] L. Liang, N. Morar, A. L. Dixon, G. M. Lathrop, G. R. Abecasis, M. F. Moffatt, and W. O. Cookson, A cross-platform analysis of 14 177 expression quantitative trait loci derived from lymphoblastoid cell lines, *Genome Res*., vol. 23, no. 4, pp. 716-726, 2013. doi: 10.1101/gr.142521.112.

[15] A. Kreimer and I. Pe'er, Variants in exons and in transcription factors affect gene expression in trans, *Genome Biol*., vol. 14, no. 7, p. R71, 2013.

[16] V. G. Cheung, R. R. Nayak, I. X. Wang, S. Elwyn, S. M. Cousins, M. Morley, and R. S. Spielman, Polymorphic cis- and trans-regulation of human gene expression, *PLoS Biol.*, vol. 8, no. 9, p. e1000480, 2010.

[17] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon, Merlin–rapid analysis of dense genetic maps using sparse gene flow trees, *Nat. Genet.*, vol. 30, no. 1, pp. 97-101, 2002.

[18] K. W. Broman, H. Wu, S. Sen, and G. A. Churchill, R/qtl: QTL mapping in experimental crosses, *Bioinformatics*, vol. 19, p. 889, 2003.

[19] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164-171, 1970.

[20] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, et al., PLINK: A tool set for whole-genome association and population-based linkage analyses, *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559-575, 2007.

[21] D. Leung, An R package for analysis of whole-genome association studies, *Hum. Hered.*, vol. 64, pp. 45-51, 2007.

[22] C. Haley and S. Knott, A simple regression method for mapping quantitative trait loci in line crosses using flanking markers, *Heredity*, vol. 69, pp. 315-324, 1992.

[23] D. M. Gatti, A. A. Shabalin, T. C. Lam, F. A. Wright, I. Rusyn, and A. B. Nobel, FastMap: Fast eQTL mapping in homozygous populations, *Bioinformatics*, vol. 25, no. 4, pp. 482-489, 2009.

[24] A. A. Shabalin, Matrix eQTL: Ultra fast eQTL analysis via large matrix operations, *Bioinformatics*, vol. 28, no. 10, pp. 1353-1358, 2012.

[25] H. M. Kang, C. Ye, and E. Eskin, Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots, *Genetics*, vol. 180, no. 4, pp. 1909-1925, 2008. doi: 10.1534/genetics.108.094201.

[26] H. M. Kang, N. A. Zaitlen, A. Kirby, C. M. Wade, D. Heckerman, M. Daly, and E. Eskin, Efficient control for population structure in model organism association mapping, *Genetics*, vol. 178, pp. 1709-1723, 2008.

[27] J. T. Leek and J. D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genet.*, vol. 3, no. 9, pp. 1724-1735, 2007.

[28] J. Listgarten, C. Kadie, E. Schadt, and D. Heckerman, Correction for hidden confounders in the genetic analysis of gene expression, *Proc. Natl. Acad. Sci. USA*, vol. 107, p. 16465, 2010.

[29] O. Stegle, L. Parts, R. Durbin, and J. Winn, A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies, *PLoS Comput. Biol.*, vol. 6, no. 5, p. e1000770, 2010.

[30] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses, *Nat. Protoc.*, vol. 7, no. 3, pp. 500-507, 2012. doi: 10.1038/nprot.2011.457.

[31] N. Fusi, O. Stegle, and N. D. Lawrence, Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies, *PLoS Comput. Biol.*, vol. 8, no. 1, p. e1002330, 2012. doi: 10.1371/journal.pcbi.1002330.

[32] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.*, vol. 58, no. 1, pp. 267-288, 1996.

[33] E. J. Candes, M. B. Wakin, and S. P. Boyd, Enhancing sparsity by reweighted l1 minimization, *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877-905, 2008.

[34] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society*, pp. 91-108, 2005.

[35] S. Kim and E. P. Xing, Statistical estimation of correlated genome associations to a quantitative trait network, *PLoS Genetics*, vol. 5, no. 8, 2009.

[36] S. Kim and E. P. Xing, Tree-guided group lasso for multi-task regression with structured sparsity, in *The 27^{th} International Conference on Machine Learning (ICML)*, 2010.

[37] S. Lee, J. Zhu, and E. P. Xing, Adaptive multi-task lasso: With application to eQTL detection, in *The 24th Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.

[38] X. Chen, S. Kim, Q. Lin, J. G. Carbonell, and E. P. Xing, Graph-structured multi-task regression and an efficient optimization method for general fused lasso, arXiv: 1005. 3579, 2010.

[39] X. Chen, X. Shi, X. Xu, Z. Wang, R. E. Mills, C. Lee, and J. Xu, A two-graph guided multi-task lasso approach for eQTL mapping, in *Proceedings of the 15th International Conference of Artificial Intelligence and Statistics (AISTATS 2012)*, La Palma, Canary Islands, 2012, pp. 208-217.

[40] S. Lee and E. P. Xing, Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs, *Bioinformatics*, vol. 28, no. 12, pp. i137-i146, 2012.

[41] G. Obozinski, B. Taskar, and M. Jordan, Joint covariate selection for grouped classification, Technical Report, Department of Statistics, University of California, Berkeley, USA, 2006.

[42] B. Rakitsch, C. Lippert, O. Stegle, K. Borgwardt, A Lasso multi-marker mixed model for association mapping with population structure correction, *Bioinformatics*, vol. 29, no. 2, pp. 206-214, 2013.

[43] Z. Wang, J. Xu, and X. Shi, Finding alternative eQTLs by exploring sparse model space, *Journal of Computational Biology*, vol. 21, no. 5, pp. 385-393, 2014.

[44] W. Cheng, X. Zhang, W. Wang, Y. Wu, X. Yin, J. Li, and D. Heckerman, Inferring novel associations between SNP sets and gene sets in eQTL study using sparse graphical model, in *Proceedings of the ACM International Conference on Bioinformatics and Computational Biology (ACMBCB)*, 2012, pp. 466-472.

[45] X. Zhang, W. Cheng, J. Listgarten, C. Kadie, S. Huang, W. Wang, and H. Heckerman, Learning transcriptional

regulatory relationships using sparse graphical models, *PLoS One*, 2012. doi:10.1371/journal.pone.0035762.

[46] L. Zhang and S. Kim, Learning gene networks under SNP perturbations using eQTL datasets, *PLoS Computational Biology*, 2014. doi: 10.137/journal.pcbi.1003420.

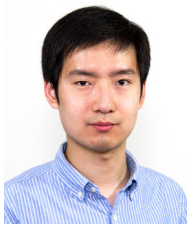
[47] J. J. Michaelson, S. Loguercio, and A. Beyer, Detection and interpretation of expression quantitative trait loci (eQTL), *Methods*, vol. 48, no. 3, pp. 265-276, 2009.

[48] S. Loguercio, R. W. Overall, J. J. Michaelson, T. Wiltshire, M. T. Pletcher, B. H. Miller, J. R. Walker, G. Kempermann, A. I. Su, and A. Beyer, Integrative analysis of low- and high-resolution eQTL, *PLoS One*, vol. 5, no. 11, p. e13920, 2010. doi: 10.1371/journal.pone.0013920.

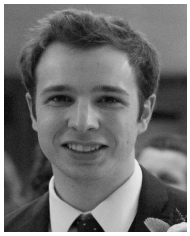
[49] J. J. Michaelson, R. Alberts, K. Schughart, and A. Beyer, Data driven assessment of eQTL mapping methods, *BMC Genomics*, vol. 11, p. 502, 2010. doi: 10.1186/1471-2164-11-502.

[50] T. Flutre, X. Wen, J. Pritchard, and M. Stephens, A statistical framework for joint eQTL analysis in multiple tissues, *PLoS Genet*., vol. 9, no. 5, p. e1003486, 2013.

[51] J. H. Sul, B. Han, C. Ye, T. Choi, and E. Eskin, Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches, *PLoS Genet*., vol. 9, no. 6, p. e1003491, 2013. doi: 10.1371/journal.pgen.1003491.



Lu Tian is currently a Professional Science Master student in the Department of Bioinformatics at the University of North Carolina at Charlotte. He is interested in analyzing large-scale genomic, expression and interaction data sets to identify markers associated with multiple phenotypes. He got his BS degree of biological science from Nankai University, China in 2006. Before joining UNCC, he performed genome-wide association study of complex human disease and clinical traits at the Center of Cancer Genomics, Wake Forest University Health Sciences.



Andrew Quitadamo graduated in 2013 from the University of New Hampshire with a BS degree in biochemistry, molecular and cellular biology. He is currently working on a PhD degree in bioinformatics at the University of North Carolina at Charlotte. His research interests include epigenetics, networks, and human disease.

Fredrick Lin is a recent graduate from the Professional Science Master's in Bioinformatics program at the University of North Carolina at Charlotte. He is interested in studying phenotypic differences affected through biological processes through epigenetics and functional genomics.

Xinghua Shi is an assistant professor in the Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte. Before joining UNC Charlotte, she was a postdoctoral research fellow at Brigham and Womens Hospital and Harvard Medical School, a NIH T32 medical genetics training fellow at Harvard Medical School, a visiting research fellow in the Medical and Population Genetics program at Broad Institute, and an associate in the Quantitative Genetics Program at Harvard School of Public Health. She received her PhD and MS degrees in computer science from the University of Chicago in 2008 and 2003, and MEng and BEng degrees in computer science from Beijing Institute of Technology, China in 2001 and 1998, respectively. She is a recipient of the Wells Fargo Foundation Fund for Faculty Excellence from Charlotte Research Institute in 2013, and a recipient of the Faculty Research Grant from the University of North Carolina at Charlotte during 2014-2015. Her research interest is in computational systems biology, particularly, the design and development of tools and algorithms to solve large-scale computational problems in biology and biomedical research. She is currently focused on integrating genetic and epigenetic datasets to study how genetic architecture affects biological processes and complex phenotypes at the systems level. She is also interested in complex network analysis, genetic privacy, and big data analytics in biomedical research.