Sparse regression models for unraveling group and individual associations in eQTL mapping

Wei Cheng, Xiang Zhang and Wei Wang

Abstract As a promising tool for dissecting the genetic basis of common diseases, expression quantitative trait loci (eQTL) study has attracted increasing research interest. Traditional eQTL methods focus on testing the associations between individual single-nucleotide polymorphisms (SNPs) and gene expression traits. A major drawback of this approach is that it cannot model the joint effect of a set of SNPs on a set of genes, which may correspond to biological pathways. To alleviate this limitation, in this chapter, we propose *geQTL*, a sparse regression method that can detect both group-wise and individual associations between SNPs and expression traits. *geQTL* can also correct the effects of potential confounders. Our method employs computationally efficient technique, thus it is able to fulfill large scale studies. Moreover, our method can automatically infer the proper number of group-wise associations. We perform extensive experiments on both simulated datasets and yeast datasets to demonstrate the effectiveness and efficiency of the proposed method.

Wei Cheng
NEC Laboratories America, Inc., e-mail: weicheng@nec-labs.com

Xiang Zhang
College Information Sciences and Technology, The Pennsylvania State University, e-mail: xzhang@ist.psu.edu

Wei Wang
Department of Computer Science, University of California, Los Angeles, e-mail: weiwang@cs.ucla.edu

The results show that *geQTL* can effectively detect both individual and group-wise signals and outperforms the state-of-the-arts by a large margin. This book chapter well illustrates that decoupling individual and group-wise associations for association mapping is able to improve eQTL mapping accuracy, and inferring individual and group-wise associations.

Key words: eQTL mapping, group-wise association, computation efficiency

1 Introduction

Expression quantitative trait loci (eQTL) mapping aims at identifying single nucleotide polymorphisms (SNPs) that influence the expression level of genes. It has been widely applied to analyze the genetic basis of gene expression and molecular mechanisms underlying complex traits [2, 18]. In a typical eQTL study, the association between each expression trait and each SNP is assessed separately [8, 29, 25]. This approach does not consider the interactions among SNPs and among genes. However, multiple SNPs may interact with each other and jointly influence the phenotypes [13]. This assumption will inevitably miss complex cases where multiple genetic variants jointly affect the co-expressions of multiple genes. It has been observed in biological experiments that the joint effect of multiple SNPs to a phenotype may be non-additive [13], and genes from the same biological pathway are usually co-regulated [20] by the same genetic basis. The biological process contains both individual effects and joint effects between SNPs and genes [19]. A straightforward approach to detect associations between sets of SNPs and a gene expression level can be done using the standard gene set enrichment analysis[10]. Wu et al. [26] further proposed the variance component models for SNP set testing. Braun et

al. employed aggregation-based approaches to cluster SNPs [3]. In [15], Listgarten et al. further considered the potential confounding factors.

However, there are two limitations for these approaches. First, these methods typically only consider SNPs from pre-defined pathways or gene ontology categories, which are far from being complete. Second, these methods can only detect the mapping of SNP set and a single gene expression level. To better elucidate the genetic basis of gene expression, it is a crucial challenge to understand how multiple modestly-associated SNPs interact to influence the a group of genes [13]. In this chapter, we refer to this kind of eQTL mapping to find associations between group of SNPs and group of gene expression levels as the *group-wise* eQTL mapping. An example is shown in Figure 1. Note that an ideal model should allow overlaps between SNP sets and between gene sets, that is, a SNP or gene may participate in multiple individual and group-wise associations [13]. In literature, *group-wise* eQTL mapping has attracted increasing research interest recently. For example, Xu et al. [5] proposed a two-graph-guided multi-task Lasso approach to infer group-wise eQTL mapping. However, it required the grouping information of both SNPs and genes available as prior knowledge, which may not be practical for many applications. Besides, it is not able to correct the effects of confounding factors.

In this chapter, we propose a novel method, *geQTL*, to automatically detect individual and group-wise associations in eQTL studies. It uses a two-layer feature selection strategy and adopts efficient optimization techniques, which make it suitable for large scale studies. Moreover, geQTL can automatically infer the optimal number of group-wise associations. We perform extensive experiments on both simulated datasets and yeast datasets to demonstrate the effectiveness and efficiency of the proposed method.

2 The Proposed Approach

2.1 Preliminaries

Important notations used in this chapter are listed in Table 1. In this chapter, for each sample, the data of SNPs and genes are denoted by column vectors. Let $\mathbf{x} = [x_1, x_2, \ldots, x_K]^{\mathrm{T}}$ denote the K SNPs. Here, $x_i \in \{0, 1, 2\}$ denotes a random variable corresponding to the i-th SNP (For example, 0, 1, 2 may encode the homozygous major allele, heterozygous allele, and homozygous minor allele, respectively.). Let $\mathbf{z} = [z_1, z_2, \ldots, z_N]^{\mathrm{T}}$ denote the N genes in the study. z_j denotes a continuous random variable corresponding to the j-th gene expression. Let $\mathbf{X} = \{\mathbf{x}_h | 1 \leq h \leq H\} \in \mathbb{R}^{K \times H}$ be the SNP matrix. We use $\mathbf{Z} = \{\mathbf{z}_h | 1 \leq h \leq H\} \in \mathbb{R}^{N \times H}$ to denote the matrix of gene expression levels. H denotes the number of samples in consideration.

The traditional linear regression model for association mapping between $\mathbf{x}$ and $\mathbf{z}$ is

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \varepsilon, \tag{1}$$

where $\mathbf{z}$ is a linear function of $\mathbf{x}$ with coefficient matrix $\mathbf{W}$, $\boldsymbol{\mu}$ is an $N \times 1$ translation factor vector. And ε is the additive noise of Gaussian distribution with zero-mean and variance $\gamma\mathbf{I}$, where γ is a scalar. That is, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \gamma\mathbf{I})$.

In association studies, sparsity is a reasonable assumption because only a small fraction of genetic variants are expected to be associated with a set of gene expression traits. This can be modeled as a feature selection problem. For example, the standard Lasso [25] can be used in association mapping, which applies ℓ_1 penalty on $\mathbf{W}$ for sparsity.

If both $\mathbf{X}$ and $\mathbf{Z}$ are standardized, the objective function of Lasso is formulated as

$$\min_{\mathbf{W}} ||\mathbf{Z} - \mathbf{W}\mathbf{X}||_F^2 + \eta ||\mathbf{W}||_1, \tag{2}$$

where $||\cdot||_F$ denotes the Frobenius norm, $||\cdot||_1$ is the ℓ_1-norm. η is the empirical parameter for the ℓ_1 penalty. $\mathbf{W}$ is the parameter (also called weight) matrix parameterizing the space of linear functions mapping from $\mathbf{X}$ to $\mathbf{Z}$.

Confounding factors, such as unobserved covariates, experimental artifacts and unknown environmental perturbations, may mask real signals and lead to spurious findings. LORS [27] uses a low-rank matrix $\mathbf{L} \in \mathbb{R}^{N\times H}$ to account for the variations caused by hidden factors. The objective function of LORS is

$$\min_{\mathbf{W},\mathbf{L}} ||\mathbf{Z} - \mathbf{W}\mathbf{X} - \mathbf{L}||_F^2 + \eta||\mathbf{W}||_1 + \rho||\mathbf{L}||_*, \tag{3}$$

where $||\cdot||_*$ is the nuclear norm [27]. ρ is the regularization parameter to control the rank of $\mathbf{L}$. $\mathbf{L}$ is a low-rank matrix assuming that there are only a small number of hidden factors influencing the gene expression levels.

When we fix $\{\mathbf{W}$, we can optimize $\{\mathbf{L}\}$ by using singular value decomposition (SVD) according to the following lemma.

Lemma 1. *([16]) Suppose that matrix* $\mathbf{O}$ *has rank r. The solution to the optimization problem*

$$\min_{\mathbf{S}} \frac{1}{2}||\mathbf{O} - \mathbf{S}||_F^2 + \lambda||\mathbf{S}||_* \tag{4}$$

is given by $\widehat{\mathbf{S}} = \mathbf{H}_\lambda(\mathbf{O})$*, where* $\mathbf{H}_\lambda(\mathbf{O}) = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}^{\mathrm{T}}$ *with* $\mathbf{D}_\lambda = \mathrm{diag}[(d_1-\lambda)_+,...,(d_r-\lambda)_+]$*,* $\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$ *is the Singular Value Decomposition (SVD) of* $\mathbf{O}$*,* $\mathbf{D} = \mathrm{diag}[d_1,...,d_r]$*, and* $(d_i - \lambda)_+ = \max((d_i-\lambda),0), (1 \le i \le r)$.

Thus, for fixed $\mathbf{W}$, the formula for updating $\mathbf{L}$ is

$$\mathbf{L} \leftarrow \mathbf{H}_\lambda(\mathbf{Z} - \mathbf{W}\mathbf{X}) \tag{5}$$

Both Lasso and LORS do not consider the existence of group-wise associations. Below, we will introduce the proposed model to infer both group-wise and individual associations for eQTL mapping.

2.2 geQTL

In geQTL, individual associations between SNPs and genes are modeled by following the Lasso-based strategy. Group-wise associations are inferred using a two-layer feature selection method. Since multiple SNPs may have joint effect on a group of genes, and such effect may be accomplished through complex biological processes, we introduce latent variables to bridge sets of SNPs and sets of genes. Specifically, we assume that there exist latent factors regulating the gene expression level, which serve as bridges between the SNPs and the genes. The latent variables are denoted by $\mathbf{y} = [y_1, y_2, \ldots, y_M]^{\mathrm{T}}$. Here, M ($M \ll min(K,N)$) is the total number of latent variables representing group-wise associations. The relationship between $\mathbf{x}$ and $\mathbf{y}$ can be represented as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}_1, \tag{6}$$

where

$$\boldsymbol{\varepsilon}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_1^2 \mathbf{I}_M).$$

$\mathbf{A} \in \mathbb{R}^{M \times K}$ denotes the matrix of coefficients between $\mathbf{x}$ and $\mathbf{y}$. $\sigma_1^2 \mathbf{I}_M$ denotes the variances of the additive noise. $\mathbf{I}_M$ is an identity matrix. Here we drop the intercept terms because the input data $\mathbf{X}$ and $\mathbf{Z}$ are normalized to zero mean and unit variance as preprocessing.

Similarly, the relationship between $\mathbf{y}$ and $\mathbf{z}$ can be represented as

$$\mathbf{z} = \mathbf{B}\mathbf{y} + \mathbf{C}\mathbf{x} + \boldsymbol{\varepsilon}_2, \tag{7}$$

where

$$\boldsymbol{\varepsilon}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \mathbf{I}_N).$$

$\mathbf{B} \in \mathbb{R}^{N\times M}$ denotes the matrix of coefficients between $\mathbf{y}$ and $\mathbf{z}$, $\mathbf{C} \in \mathbb{R}^{N\times K}$ denotes the matrix of coefficients between $\mathbf{x}$ and $\mathbf{z}$ to encode the individual associations.

Note that Eq. (7) decouples the associations between SNPs and genes into two parts: one for individual associations represented as $\mathbf{Cx}$, and another for group-wise associations represented as $\mathbf{By}$. Next, we infer the group-wise associations by a two-layer feature selection strategy. We first remove the individual associations and denote

$$\tilde{\mathbf{Z}} = \mathbf{Z} - \mathbf{CX}. \tag{8}$$

Thus $\tilde{\mathbf{Z}}$ contains only group-wise effects. Next let

$$\mathbf{Y} = \mathbf{AX}. \tag{9}$$

Thus $\mathbf{Y}$ represents a low-rank transformation of the original SNP matrix. Each row of $\mathbf{Y}$ represents a group of SNPs. From Eq. (7), we have the following multiple-input-multiple-output (MIMO) linear system

$$\tilde{\mathbf{Z}} = \mathbf{BY} + \mathbf{E}, \tag{10}$$

where $\mathbf{E}$ is a Gaussian white-noise term. In Eq. (9) and (10), $\mathbf{A}$ and $\mathbf{B}$ should be sparse since a single gene is often influenced by a small number of SNPs and vice versa [15].

Therefore, the overall objective function is

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{L}} loss(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{L}) \\ &\quad +\rho||\mathbf{L}||_* + \alpha||\mathbf{A}||_1 + \beta||\mathbf{B}||_1 + \gamma||\mathbf{C}||_1, \end{aligned} \tag{11}$$

where $\alpha, \beta, \gamma, \rho$ are the regularization parameters, and the loss function is

$$loss(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{L}) = ||\mathbf{Z}-\mathbf{L}-(\mathbf{BA}+\mathbf{C})\mathbf{X}||_F^2. \quad (12)$$

Here, we choose different penalties for $\mathbf{A},\mathbf{B},\mathbf{C}$ because the sparsities of different matrices are typically of different scales.

2.3 Optimization

The optimization for $\mathbf{L}$ can be achieved by following a similar approach as in [27]. To optimize $\mathbf{A},\mathbf{B},\mathbf{C}$, many tools can be used to optimize the ℓ_1 penalized objective function, e.g., the Orthant-Wise Limited-memory Quasi-Newton (OWL-QN) algorithm [1]. Due to space limitation, we omit the details. In the next, we devise optimization techniques that can dramatically improve the computational efficiency of geQTL.

2.4 Boosting the Computational Efficiency

Given a large number of SNPs and gene expression traits, scalability of the algorithm is a crucial issue. We propose two improved models, geQTL$^+$ and geQTL-ridge, which optimize the search for significant individual associations, which is the main computational bottleneck of the algorithm.

2.4.1 geQTL$^+$

In a typical eQTL study, we usually have $M \ll min(K,N)$. Thus, the bottleneck of the algorithm is to optimize $\mathbf{C}$. Our strategy is to confine the space of $\mathbf{C}$. The intuition is that we only permit a small fraction of elements in $\mathbf{C}$ to be nonzero. It has been shown that if $\mathbf{Z}$ and $\mathbf{X}$ are standardized with zero mean and unit sum of squares,

then $\mathbf{r} = abs(\mathbf{Z}\mathbf{X}^{\mathrm{T}})$ is equal to the gene-SNP correlations ($\mathbf{r}_{gs} = |cor(z_g, x_s)|$) [22]. Since for many test statistics, e.g., t, F, R^2, and LR, for the simple linear regression problem can be expressed as functions of the sample correlation $\mathbf{r}_{gs}$, e.g., $R^2 = r^2$, and $t = \frac{r\sqrt{n-2}}{1-r^2}$, we can find a threshold according to the required p-value, such that test statistics exceeding the threshold are significant at the required significance level. The test statistics for every gene-SNP pair in $\mathbf{r}$ are compared with the threshold, and only those elements whose $\mathbf{r}$ are greater than the threshold are optimized. We denote $\mathbf{R} \in \mathbb{R}^{N\times K}$ as the indicator matrix indicating which elements in $\mathbf{C}$ can be nonzero (i.e., $\mathbf{r}_{gs} > threshold$).

2.4.2 geQTL-ridge

When N and K are extremely large, optimizing $\mathbf{C}$ may still be time-consuming, since it may take many iterations to converge with the ℓ_1 constraint. Next, we introduce geQTL-ridge, which further improves the time efficiency with slight decrease in accuracy. The key idea is to use ridge regression for individual associations so that we can get a closed form solution for $\mathbf{C}$. The objective function is shown in the following.

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{L}} loss(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{L}) \\ &\quad +\rho||\mathbf{L}||_* + \alpha||\mathbf{A}||_1 + \beta||\mathbf{B}||_1 + \gamma||\mathbf{C}||_2^2, \\ &\quad s.t.\ (\mathbf{C})_{i,j} \text{ is nozero only if } (\mathbf{R})_{i,j} \text{ is } 1. \end{aligned} \tag{13}$$

Theorem 1. *The solution of* $\mathbf{C}$ *in Eq. (13) is*

$$\mathbf{c}_i \leftarrow \mathbf{d}_i\mathbf{X}^{\mathrm{T}}\mathbf{P}_i(\mathbf{P}_i^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{P}_i + \gamma\mathbf{I}_K)^{-1}\mathbf{P}_i^{\mathrm{T}}, \tag{14}$$

where

$$\mathbf{c}_i = (\mathbf{C})_{i,:}, \mathbf{d}_i = (\mathbf{D})_{i,:},$$

$$\mathbf{D} = \mathbf{Z} - \mathbf{L} - \mathbf{BAX},$$

and $\mathbf{P}_i$ *is defined as in formula (19).*

The proof of the Theorem 1 is in the following section.

2.5 Proof of Theorem 1

Proof. Recall that any ridge regression problem

$$\min_{\mathbf{a}} ||\mathbf{b} - \mathbf{aQ}||_2^2 + ||\mathbf{a}\Gamma||_2^2, \tag{15}$$

where $\mathbf{a}$ is a row vector and $\mathbf{Q}$ has linearly independent rows, has the following solution

$$\mathbf{a} = \mathbf{bQ}^{\mathrm{T}}(\mathbf{QQ}^{\mathrm{T}} + \Gamma\Gamma^{\mathrm{T}})^{-1}. \tag{16}$$

Note that

$$loss(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{L}) = ||\mathbf{D} - \mathbf{CX}||_F^2 = \sum_{i=1}^{N} ||\mathbf{d}_i - \mathbf{c}_i\mathbf{X}||_2^2, \tag{17}$$

where $\mathbf{D} = \mathbf{Z} - \mathbf{L} - \mathbf{BAX}$, $\mathbf{c}_i = (\mathbf{C})_{i,:}$ and $\mathbf{d}_i = (\mathbf{D})_{i,:}$.

We have

$$\min_{\mathbf{C}} \quad loss(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{L}) = \sum_{i=1}^{N} \min_{\mathbf{c}_i} ||\mathbf{d}_i - \mathbf{c}_i\mathbf{X}||_2^2, \tag{18}$$

Taking into account that $(\mathbf{c}_i)_j$ can be nonzero only if $(\mathbf{R})_{i,j}$ is 1, we introduce $\mathbf{P}_i$, where $\mathbf{P}_i$ has K rows and $l_i = \sum_{j=1}^{K} (\mathbf{R})_{i,j}$ columns. And

$$(\mathbf{P}_i)_{s,t} = \begin{cases} 1, & \textit{if } (\mathbf{R})_{i,s} \textit{ is the } t\textit{-th } 1 \textit{ in } (\mathbf{R})_{i,:}; \\ 0, & \textit{otherwise}. \end{cases} \tag{19}$$

Then $\mathbf{c}_i = \mathbf{c}_i\mathbf{P}_i\mathbf{P}_i^{\mathrm{T}}$, $||\mathbf{d}_i - \mathbf{c}_i\mathbf{X}||_2^2 + \gamma||\mathbf{c}_i||_2^2 = ||\mathbf{d}_i - (\mathbf{c}_i\mathbf{P}_i)(\mathbf{P}_i^{\mathrm{T}}\mathbf{X})||_2^2 + \gamma||\mathbf{c}_i\mathbf{P}_i||_2^2$, and

$$\begin{aligned} &\min_{\mathbf{c}_i} ||\mathbf{d}_i - \mathbf{c}_i\mathbf{X}||_2^2 + \gamma||\mathbf{c}_i||_2^2, \\ &\textit{s.t. } (\mathbf{c}_i)_j \textit{ is nozero only if } (\mathbf{R})_{i,j} \textit{ is } 1, \end{aligned} \tag{20}$$

is solved by

$$\mathbf{c}_i = (\mathbf{c}_i\mathbf{P}_i)\mathbf{P}_i^{\mathrm{T}} = \mathbf{d}_i\mathbf{X}^{\mathrm{T}}\mathbf{P}_i(\mathbf{P}_i^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{P}_i + \gamma\mathbf{I}_K)^{-1}\mathbf{P}_i^{\mathrm{T}}. \tag{21}$$

Therefore,

$$\begin{aligned} \min_{\mathbf{C}} \quad & \textit{loss}(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{L}) + \gamma||\mathbf{C}||_2^2, \\ & \textit{s.t.}(\mathbf{C})_{i,j} \textit{ is nozero only if } (\mathbf{R})_{i,j} \textit{ is } 1, \end{aligned}$$

is solved by $\mathbf{C} = (\mathbf{c}_1^{\mathrm{T}}, ..., \mathbf{c}_N^{\mathrm{T}})^{\mathrm{T}}$, which leads to the update formula given in Eq. (14).

2.6 Determining the Number of Hidden Variables

In Eq. (12), we use $\mathbf{BA} + \mathbf{C}$ to formulate the overall associations between SNPs and expression traits. Two group-wise associations will not share the same group of SNPs (or genes), since otherwise these two group-wise associations can be combined into one. Therefore, every group-wise association should be unique and irreplaceable. Hence, following two conditions should be satisfied

- $\mathbf{A}$ has linearly independent rows. Since $M \ll K$, this condition is equivalent to that $\mathbf{A}$ has full rank;

- **B** has linearly independent columns. Since $M \ll N$, this condition is equivalent to that **B** has full rank.

When these two conditions are met, we have

$$M = \text{rank}(\mathbf{A}) = \text{rank}(\mathbf{B}) = \text{rank}(\mathbf{BA}). \tag{22}$$

The last equality holds because both **A** and **B** have full rank.

We have the following observation. The singular value decomposition (SVD) of **BA** has the form

$$\mathbf{BA} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}},$$

where **U** and **V** are unitary (orthogonal in our case) matrices, and Σ is a rectangular diagonal matrix with non-negative real numbers on the diagonal, which corresponds to singular values of **BA**. Since **U** and **V** are unitary and hence have full rank, we have

$$\begin{aligned} \text{rank}(\mathbf{BA}) =& \text{rank}(\mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}) = \text{rank}(\Sigma) \\ =& \text{the number of nonzero singular values of } \mathbf{BA}. \end{aligned} \tag{23}$$

We compute **BA** by minimizing Eq. (12), which gives

$$\mathbf{BA} = (\mathbf{Z} - \mathbf{L} - \mathbf{CX})\mathbf{X}^{\mathrm{T}}(\mathbf{XX}^{\mathrm{T}})^{-1}. \tag{24}$$

Combine (22), (23), and (24), we find

$$\begin{aligned} M = & \text{ the number of nonzero singular values of} \\ & (\mathbf{Z} - \mathbf{L} - \mathbf{CX})\mathbf{X}^{\mathrm{T}}(\mathbf{XX}^{\mathrm{T}})^{-1}. \end{aligned} \tag{25}$$

Due to the existence of noise, we should allow small singular values to be considered as zero. Therefore, we can draw a plot with singular values of $(\mathbf{Z} - \mathbf{L} -$

$\mathbf{CX})\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}$ in descending order and set M to be k, if the first k singular values are large and significantly greater than the $(k+1)$-th singular value.

Based on the discussion above, in order to find optimal M, we can first use Lasso to infer the initial value of $\mathbf{C}$. Then, using Eq. 25, we can infer the optimal M at this stage. After that, we can optimize new $\mathbf{C}$, and calculate new optimal M. We can repeat this procedure until M became stable or reach maximal number of iterations.

3 Experimental Study

In this section, we perform extensive experimental study using both simulated and real eQTL datasets to evaluate the performance of our methods. For comparison, we select several state-of-the-art eQTL methods, including two-graph guided multi-task lasso (MTLasso2G) [5], FaST-LMM [15], SET-eQTL [6], LORS [27], Matrix eQTL [22] and Lasso [25]. Note that we did not compare with our previous work, GDL, in [7] because it needs to incorporate many prior knowledge, that is not relevant to this work. For all the methods, the tuning parameters are learned using cross validation. The discussion of setting proper number of group-wise associations M is included in the supplementary material. The shrinkage of the coefficients is also presented in the supplementary material.

3.1 Simulated Data

We use a similar setup for simulation study to that in [27]. First, 100 SNPs are randomly selected from the yeast eQTL dataset [21]. This gives birth to the matrix $\mathbf{X}$. 100 gene expression profiles are generated by $\mathbf{Z}_{j*} = \beta_{j*}\mathbf{X} + \Xi_{j*} + \mathbf{E}_{j*}$ $(1 \leq j \leq N)$, where $\mathbf{E}_{j*} \sim \mathcal{N}(0, \phi I)$ $(\phi = 0.1)$ is used to simulate the Gaussian noise. To

simulate the effects of confounding factors, we use Ξ_{j*}, drawn from $\mathcal{N}(\mathbf{0}, \tau\Lambda)$. In this chapter, we set $\tau = 0.1$. Λ is given by $\mathbf{F}\mathbf{F}^{\mathrm{T}}$. Here, $\mathbf{F} \in \mathbb{R}^{H\times J}$ and $\mathbf{F}_{ij} \sim \mathcal{N}(0,1)$. J denotes hidden factor number. In this chapter, we set J to 10.

In the left most of Figure 2, we illustrate β. Here, we set the association strength to 1. Totally, there exist four group-wise associations with different scales. The diagonal line represents the individual signals in *cis*-regulation.

In Figure 2, we report the associations inferred by geQTL. Recall that group-wise associations can be inferred from matrix $\mathbf{A}$ and $\mathbf{B}$, and individual associations can be inferred from matrix $\mathbf{C}$. It is obvious that geQTL can detect both group-wise and individual signals.

We use $SNR = \sqrt{\frac{Var(\beta\mathbf{X})}{Var(\Xi+\mathbf{E})}}$ to denote the signal-to-noise ratio [27] in the eQTL datasets. Here, we fix $J = 10, \tau = 0.1$. The *SNR*'s are controlled by using different ϕ's. Using 50 simulated datasets with different *SNR*'s, we compare the proposed methods with the selected methods. Because FaST-LMM requires the input of genomic locations information (e.g., chromosome, base pair, etc), we will compare it on the real data set. The results are averaged over 50 different simulated datasets. $\mathbf{BA} + \mathbf{C}$ is used to represent the association matrix in our method. Figure 3 shows the ROC curve of TPR-FPR (true positive rate - false positive rate) for performance comparison. Typically, we care more about the TPR when the FPR is small because it is important to evaluate the performance of model when controlling the maximum tolerated FPR. Thus, in Figure 3, the ROC of interest for eQTL are generally shown in the range [0,0.1]. The corresponding areas under the TPR-FPR curve are shown in Figure 4.

It can be seen that geQTL and geQTL$^+$ outperform all alternative methods by a large margin since they considers both individual and group-wise associations. We also observe that geQTL-ridge is not as good as geQTL and geQTL$^+$. This is because geQTL-ridge does not provide a sparse solution for individual association-

s. MTLasso2G is comparable to LORS. LORS can correct the effects of the confounders, however, it is not able to detect group-wise mappings. We also observe that by decoupling individual and group-wise associations, the proposed models (geQTL, geQTL$^+$, and geQTL-ridge) are more robust to noise than other methods.

3.2 Yeast eQTL Data

We also validated *geQTL* using the bench mark dataset–yeast (Saccharomyces cerevisiae) eQTL dataset. The dataset contains 112 yeast segregants generated from a cross of two inbred strains [21]. Originally, It contains 6229 gene epxressions and 2956 SNPs. SNPs with $> 10\%$ missing values in the remaining SNPs are imputed using the function fill.geno in R/qtl [4]. The neighboring SNPs with the same genotype profiles are combined, resulting in 1027 genotype profiles. Remove gene expression traits with missing values, we get 4474 expression profiles.

3.2.1 cis- and trans- analysis

We follow the standard *cis*-enrichment analysis that is used in [14, 17] for evaluation. Moreover, we use the *trans*-enrichment with a similar strategy [28]. Genes regulated by transcription factors (obtained from http://www.yeastract.com/download.php) are treated as *trans*-acting signals.

In Table 2, we report the pairwise comparison using cis- and trans- enrichment analysis. We do not list geQTL separately from geQTL+, since geQTL+ is a faster version of geQTL. In this table, the methods are sorted (from top to bottom in the left column and from left to right in the top row) in decreasing order of performance. A p-value shows how significant a method on the left column outperforms a method in the top row in terms of *cis* and *trans* enrichments. We observe that

geQTL$^+$ has significantly better *cis*-enrichment scores than the other models. For *trans*-enrichment, geQTL$^+$ is the best, and MTLasso2G comes in second, outperforming FaST-LMM, SET-eQTL, LORS, Matrix eQTL and Lasso. LORS outperforms Matrix eQTL and Lasso for both *cis*- and *trans*-enrichment. This is because LORS considers confounding factors while Matrix eQTL and Lasso does not. In total, these methods each detected about 6000 associations according to non-zero $\mathbf{W}$ values. We estimate FDR using 50 permutations as proposed in [27]. With FDR ≤ 0.01, geQTL$^+$ obtains about 4500 significant associations. The plots of all identified significant associations for different methods are given in Figure 5. Obviously, we can see that $\mathbf{C} + \mathbf{B} \times \mathbf{A}$ and $\mathbf{C}$ of geQTL$^+$ report weaker *trans*-regulatory bands while stronger *cis*-regulatory signals than other competitors.

3.2.2 Gene ontology enrichment analysis on detected group-wise associations for yeast

We further evaluate the quality of detected groups of genes by measuring the correlations between the detected groups of genes and the GO (Gene Ontology) categories [24]. Specifically, the GO enrichment test is calculated by DAVID [11]. In this chapter, gene sets reported by our algorithm with calibrated p-values less than 0.01 are considered as significantly enriched.

Since SET-eQTL is the only previous approach capable of detecting group-wise association mapping, we compare the groups of genes detected by geQTL and those by SET-eQTL. For SET-eQTL, 90 out of 150 gene sets are significantly enriched. By contrast, 28 out of 30 gene sets reported by geQTL are significantly enriched. This well illustrates that the effectiveness of geQTL to infer group-wise associations. The number of SNPs in each group reported by geQTL and their genomic locations are shown in Figure 6. We can clearly observe that SNPs in the same group are often

physically close to each other. This is reasonable because SNPs nearby usually jointly affect the expression level of a set of genes that achieves a specific cell function [19].

3.2.3 Reproducibility of eQTLs between studies

To further evaluate the identified associations, we investigate the consistency of calls between two independent studies [23]. We examine the reproducibility based on the following two criteria [9, 27, 12]:

- Reproducibility of detected SNP-gene associations: Let L_1 and L_2 be the sets of SNP-gene associations detected in the two yeast datasets, respectively. We can rank the associations according to the weights (or q-values for FaST-LMM). Let L_1^T and L_2^T be the top T most significant associations from the two datasets. The reproducibility is defined as $\frac{|L_1^T \bigcap L_2^T|}{T}$.
- Reproducibility of *trans* regulatory hotspots: For each SNP, we count the number of associated genes from the detected SNP-gene associations. We use this number as the *regulatory degree* of each SNP. For FaST-LMM, SNP-Gene pairs with a q-value < 0.001 are considered significant associations. We also tried different cutoffs for FaST-LMM (from 0.01 to 0.001), the results are similar. SNPs with large regulatory degrees are often referred to as hotspots. We sort SNPs in descending order of their regulatory degrees. We denote the sorted SNPs lists as S_1 and S_2 for the two yeast datasets. Let S_1^T and S_2^T be the top T SNPs in the sorted SNP lists. The trans calling consistency of reported hotspots is denoted by $\frac{|S_1^T \bigcap S_2^T|}{T}$.

In Figure 7 (a), we show the consistency of the top 4500 associations between different studies. In Figure 7 (b), we list the reproducibility of *trans* regulatory hotspots

reported by different approaches. Overall, geQTL$^+$ yielded results with greater consistency all other methods. This well illustrates the superiority of geQTL$^+$.

4 Conclusion

In literature, much efforts have been done on eQTL mapping. Traditional eQTL mapping approaches can not detect the group-wise associations between sets of SNPs and sets of genes. To achieve that, we propose a fast approach, *geQTL*, to detect both individual and group-wise associations for eQTL mapping. *geQTL* can also correct the effects of potential confounders. We also introduce efficient algorithms to scale up the computation so that the algorithms are able to tackle large scale data sets. Additionally, the proposed model provides an effective strategy to automatically infer the proper number of group-wise associations. We perform extensive experiments on both simulated datasets and yeast datasets to demonstrate the effectiveness and efficiency of the proposed method. Inferring individual and group-wise associations also helps us better explain the genetic basis of gene expression. Due to scalability issue, our model simply assume random errors between different genes are independent and have the same variance. That is the reason why current model only identified a small number of group-wise associations. Our future work will further incorporate the relationships between genes by integrating gene co-expression network or protein-protein-interaction network.

References

[1] Andrew, G. and Gao, J. (2007). Scalable training of l1-regularized log-linear models. *ICML'07*.

[2] Bochner, B. R. (2003). New technologies to assess genotype henotype relationships. *Nature Reviews Genetics*, **4**, 309–314.

[3] Braun, R. and Buetow, K. (2011). Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data. *PLoS Genet.*, **7**(6), e1002101.

[4] Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**(7), 889–890.

[5] Chen, X., Shi, X., Xu, X., Wang, Z., Mills, R., Lee, C., and Xu, J. (2012). A two-graph guided multi-task lasso approach for eqtl mapping. In *AISTATS'12*, pages 208–217.

[6] Cheng, W., Zhang, X., Wu, Y., Yin, X., Li, J., Heckerman, D., and Wang, W. (2012). Inferring novel associations between snp sets and gene sets in eqtl study using sparse graphical model. In *ACM-BCB'12*, pages 466–473.

[7] Cheng, W., Zhang, X., Guo, Z., Shi, Y., and Wang, W. (2014). Graph-regularized dual lasso for robust eQTL mapping. *Bioinformatics*, pages 139–148.

[8] Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M., and Burdick, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, pages 1365–1369.

[9] Fusi, N., Stegle, O., and Lawrence, N. D. (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol.*, **8**(1), e1002330.

[10] Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, **24**(23), 2784–2785.

[11] Huang, d. a. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**(1), 44–57.

[12] Joo, J. W., Sul, J. H., 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.

[13] Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, **470**(7333), 187–197.

[14] Listgarten, J., Kadie, C., Schadt, E. E., and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **107**(38), 16465–16470.

[15] Listgarten, J., Lippert, C., Kang, E. Y., Xiang, J., Kadie, C. M., and Heckerman, D. (2013). A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics*, **29**(12), 1526–1533.

[16] Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *JMLR*, **11**, 2287–2322.

[17] McClurg, P., Janes, J., Wu, C., Delano, D. L., Walker, J. R., Batalov, S., Takahashi, J. S., Shimomura, K., Kohsaka, A., Bass, J., Wiltshire, T., and Su, A. I. (2007). Genomewide association analysis in diverse inbred mice: power and population structure. *Genetics*, **176**(1), 675–683.

[18] Michaelson, J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, **48(3)**, 265–276.

[19] Musani, S. K., Shriner, D., Liu, N., Feng, R., Coffey, C. S., Yi, N., Tiwari, H. K., and Allison, D. B. (2007). Detection of gene x gene interactions in genome-wide association studies of human population data. *Human Heredity*, pages 67–84.

[20] Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., and Muneesh Tewari, e. a. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics*, pages 1338–1349.

[21] Rachel B. Brem, John D. Storey, J. W. and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, pages 701–03.

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

[23] Smith, E. N. and Kruglyak, L. (2008). Gene-environment interaction in yeast gene expression. *PLoS Biol*, page e83.

[24] The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25(1)**, 25–29.

[25] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, **58(1)**, 267–288.

[26] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**(1), 82–93.

[27] Yang, C., Wang, L., Zhang, S., and Zhao, H. (2013). Accounting for non-genetic factors by low-rank representation and sparse regression for eQTL mapping. *Bioinformatics*, pages 1026–1034.

[28] Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.*, **35**(1), 57–64.

[29] Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E., and Schadt, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, pages 854–61.