Inferring novel associations between SNP sets and gene sets in eQTL study using sparse graphical model

Wei Cheng [1], Xiang Zhang [2], Yubao Wu [2], Xiaolin Yin [2], Jing Li [2],
David Heckerman [3], and Wei Wang [4]
[1]Department of Computer Science, University of North Carolina at Chapel Hill, [2]Department of Electrical Engineering and Computer Science, Case Western Reserve University, [3]Microsoft Research
[4]Department of Computer Science, University of California, Los Angeles
[1]weicheng@cs.unc.edu,[2]{xiang.zhang,yubao.wu,xly,jingli}@case.edu,
[3]heckerma@microsoft.com ,[4]weiwang@cs.ucla.edu

ABSTRACT

Genome-wide expression quantitative trait loci (eQTL) studies have emerged as a powerful tool to understand the genetic basis of gene expression and complex traits. The 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 corresponds to biological pathways. In this paper, we propose a sparse (ℓ_1-regularized) graphical model, SET-eQTL, to identify novel associations between sets of SNPs and sets of genes. Such associations are captured by hidden variables connecting SNPs and genes. These hidden variables also naturally model the potential effect of unknown confounding factors. We compare three different methods on a yeast segregant dataset. Extensive experimental results demonstrate that the proposed graphical model SET-eQTL achieves better performance than the other two alternatives.

Categories and Subject Descriptors

H.3.3 [**GWAS**]: eQTL; H.2.8 [**System Biology**]:

General Terms

Algorithms, Experimentation, Theory

Keywords

eQTL, Graphical Model, Gene Set, SNP Set

1. INTRODUCTION

Thanks to the advanced high-throughput technologies for profiling gene expressions and assaying genetic variations, genome-wide study of expression quantitative trait loci (eQTL) has been widely applied to dissect genetic basis of gene expression and molecular mechanisms underlying complex traits [5, 40, 28]. In a typical eQTL study, the association between each expression trait and each single-nucleotide polymorphism (SNP) is assessed separately [8, 51, 45].

Despite the successful applications of this single-locus approach, there are several thorny issues that greatly limit its applicability. First, the large number of SNPs and gene expression traits leads to a huge number of correlated tests [14]. Many SNPs may be genuinely associated with genes but may not reach a stringent genome wide significance threshold after correction for multiple testing. Second, the single-locus approach ignores the joint effect of a set of SNPs on the activities of a set of genes, which may act and interact with each other to achieve a specific cell function. It is widely recognized that genes in the same biological pathway are often co-regulated and may share a common genetic basis [31, 37]. It is a crucial challenge to understand *how multiple, modestly-associated SNPs interact to influence the phenotypes* [23]. Third, confounding factors such as expression heterogeneity may result in spurious associations and mask real signals [29, 42, 13].

Several approaches have been proposed to partially address these challenging issues. To find SNP-SNP interactions, epistasis detection methods have been developed [17, 16, 2, 30]. These methods focus on finding interactions between SNP-pairs. They still suffer the multiple testing problem and are computationally intensive. Recently, machine learning methods, such as Lasso and its variations [45, 22, 24], have been applied to eQTL studies. These methods are effective in addressing the "large p small n" problem (i.e., high dimension and low sample size) and can aggregate associations across multiple SNPs [4]. However, they do not consider the effect of confounding factors which may dramatically affect the results. Statistical models that incorporate confounding factors have been proposed in [26, 43]. These methods are not specifically designed to identify novel associations between SNP sets and gene sets. Pathway analysis methods [7, 11, 46] aim to examine the associations between pre-determined SNP sets (usually from existing knowledgebase, such as GO and KEGG [44, 27]) and the phenotypes. Although this approach is appealing, it is limited to the priori knowledge on the predefined SNP sets/pathways.

To better elucidate the genetic basis of gene expression and understand the underlying biology pathways, it is highly desirable to develop approaches that can automatically infer associations between a group of SNPs and a group of genes. Intuitively, the eQTL data can be modeled using a bipartite graph, where the SNPs are a set of nodes and the genes are another set of nodes. The expression levels of the genes can be treated as a function of SNP combinations represented by the (weighted) edges connecting SNPs and genes. In [19], a method has been proposed to identify cliques in a bipartite graph derived from eQTL data. The cliques are used to model

ACM-BCB '12, October 08-10 2012, Orlando, FL, USA

the hidden correlations between SNP sets and gene sets. However, this method depends on the availability of progeny strain information, which is used as a bridge for modeling the eQTL association graphs. Moreover, it does not consider the confounding factors.

To address the limitations of existing approaches, in this paper, we propose a sparse linear-Gaussian graphical model, SET-eQTL, to infer novel associations between SNP sets and gene sets. The proposed model consists of three layers of nodes as shown in Figure 2. The upper layer nodes correspond to the set of SNPs in the study. The middle layer consists of a set of hidden variables. The hidden variables are used to model both the joint effect of a set of SNPs and the effect of confounding factors. The lower layer nodes correspond to the genes in the study. The nodes in different layers are connected via arcs. Please refer to Section 3 for further details of the proposed model.

To learn the parameters of the proposed model from the eQTL data, which is usually of high dimension and low sample size, we apply an ℓ_1-norm on the parameters [25, 45, 10, 15]. This approach yields a sparse network, where a large number of association weights are zero [33]. In eQTL association networks, most genes are regulated by a subset of SNPs. The matrix that describes the connections between the SNPs and the regulated genes is expected to be sparse. Thus ℓ_1-regularization is a natural choice for this problem.

We apply our model to a yeast data set and show that it has better performance than two alternative methods. We further examine the gene sets connected to hidden variables and find that most these gene sets are strongly correlated with GO categories.

2. RELATED WORK

Recently, various analytic methods have been developed to address the limitations of the traditional single-locus approach. Epistasis detection methods aim to find the interaction between SNP-pairs [17, 16, 2, 30]. The computational burden of epistasis detection is usually very high due to the large number of interactions that need to be examined [32, 39]. Filtering-based approaches [12, 18, 50], which reduce the search space by selecting a small subset of SNPs for interaction study, may miss important interactions in the SNPs that have been filtered out.

Statistical graphical models and Lasso-based methods [45] have been applied to eQTL study. A tree-guided group lasso has been proposed in [22]. This method directly combines statistical strength across multiple related genes in gene expression data to identify SNPs with pleiotropic effects by leveraging the hierarchical clustering tree over genes. Bayesian methods have also been developed [26, 43]. Confounding factors may greatly affect the results of the eQTL study. To model confounders, a two-step approach can be applied [43, 21]. These methods first learn the confounders that may exhibit broad effects to the gene expression traits. The learned confounders are then used as covariates in the subsequent analysis. Statistical models that incorporate confounders have been proposed [34]. However, none of these methods are specifically designed to find novel associations between SNP sets and gene sets.

Pathway analysis methods have been developed to aggregate the association signals by considering a set of SNPs together [7, 11, 46, 36]. A pathway consists of a set of genes that coordinate to achieve a specific cell function. This approach studies a set of known pathways to find the ones that are highly associated with the phenotype [47]. Although appealing, this approach is limited to the priori knowledge on the predefined gene sets/pathways. On the other hand, the current knowledgebase on the biological pathways is still far from being complete.

In [19], a method is proposed to identify eQTL association cliques

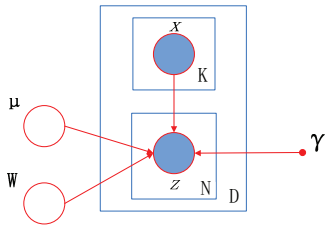


Figure 1: Graphical model for linear regression

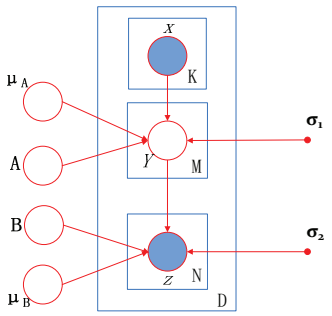


Figure 2: The proposed graphical model with hidden variables

that expose the hidden structure of genotype and expression data. By using the cliques identified, this method can filter out SNP-gene pairs that are unlikely to have significant associations. It models the SNP, progeny and gene expression data as an eQTL association graph, and thus depends on the availability of the progeny strain data as a bridge for modeling the eQTL association graph.

3. METHODS

3.1 The Proposed Graphical Model

Throughout the paper, we assume that, for each sample, the genotype and gene expression are represented by two column vectors. Let $\mathbf{x} = [x_1, x_2, \ldots, x_K]^{\mathrm{T}}$ represent the K SNPs in the study, where $x_i \in \{0, 1, 2\}$ is a random variable corresponding to the i-th SNP. Let $\mathbf{z} = [z_1, z_2, \ldots, z_N]^{\mathrm{T}}$ represent the N genes in the study, where z_j is a continuous random variable corresponding to the j-th gene. The traditional linear regression model for association mapping between $\mathbf{x}$ and $\mathbf{z}$ is

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

where z is a linear function of $\mathbf{x}$ with coefficient matrix $\mathbf{W}$. $\boldsymbol{\mu}$ is a translation factor vector. ϵ is the additive noise of Gaussian distribution with zero-mean and variance $\gamma\mathbf{I}$, where γ is a scalar. That is $\epsilon \sim \mathcal{N}(\mathbf{0}, \gamma\mathbf{I})$. Figure 1 shows the conventional graphical model representation of the linear regression method [4].

To infer associations between SNP sets and gene sets, we propose a graphical model as shown in Figure 2, which is able to capture any potential confounding factors in a natural way. Specifically, we assume that there exist some latent factors regulating the gene expression level, which serves as bridges between the SNPs and the genes. These latent variables are presented as $\mathbf{y} = [y_1, y_2, \ldots, y_M]^{\mathrm{T}}$, where M is the total number of latent variables.

The exact role of these latent factors can be inferred from the network topology of the resulting sparse graphical model learned from the data (by imposing ℓ_1-norm on the likelihood function, which will be discussed later in this section). Figure 3 shows an example of the resulting graphical model. There are two types of hidden

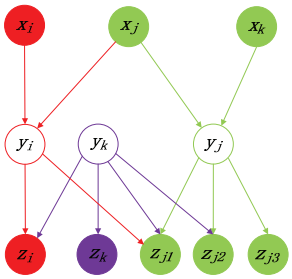


Figure 3: An example of the inferred sparse graphical model

variables. One type consists of hidden variables with zero in-degree (i.e., no connections with the SNPs). These hidden variables correspond to the confounding factors. Another type of hidden variables serve as bridges connecting SNP sets and gene sets. In Figure 3, y_k is a hidden variable modeling confounding effects. y_i and y_j are bridge nodes connecting the SNPs and genes associated with them. Note that this model allows overlaps between different (SNP set, gene set) pairs. It is reasonable because SNPs and genes may play multiple roles in different biology pathways.

3.2 Objective Function

From the probability theory, we have that the joint probability of $\mathbf{x}$ and $\mathbf{z}$ is

$$p(\mathbf{x},\mathbf{z}) = \int_{\mathbf{y}} p(\mathbf{x},\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{y}. \tag{2}$$

From the factorization properties of the joint distribution for a directed graphical model, we have

$$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{z}|\mathbf{y})p(\mathbf{x}). \tag{3}$$

Thus, we have

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})} = \int_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})p(\mathbf{z}|\mathbf{y})\mathrm{d}\mathbf{y}. \tag{4}$$

The two conditional probabilities follow normal distributions:

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\boldsymbol{\mu}_{\mathbf{A}}, \sigma_1^2\mathbf{I}_{\mathbf{M}}),$$

and

$$\mathbf{z}|\mathbf{y} \sim \mathcal{N}(\mathbf{z}|\mathbf{B}\mathbf{y}+\boldsymbol{\mu}_{\mathbf{B}}, \sigma_2^2\mathbf{I}_{\mathbf{N}}),$$

where $\mathbf{A} \in \mathbb{R}^{M\times K}$ is the coefficient matrix between $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{B} \in \mathbb{R}^{N\times M}$ is the coefficient matrix between $\mathbf{y}$ and $\mathbf{z}$. $\boldsymbol{\mu}_{\mathbf{A}} \in \mathbb{R}^{M\times 1}$ and $\boldsymbol{\mu}_{\mathbf{B}} \in \mathbb{R}^{N\times 1}$ are the translation factor vectors, of which $\sigma_1^2\mathbf{I}_M$ and $\sigma_2^2\mathbf{I}_N$ are their variances respectively (σ_1 and σ_2 are constant scalars and $\mathbf{I}_M$ and $\mathbf{I}_N$ are identity matrices).

To impose sparsity, we assume that $\mathbf{A}$ and $\mathbf{B}$ follow Laplace distributions:

$$\mathbf{A} \sim \mathbf{Laplace}(\mathbf{0}, 1/\lambda),$$

and

$$\mathbf{B} \sim \mathbf{Laplace}(\mathbf{0}, 1/\gamma).$$

λ and γ are parameters of the ℓ_1-regularization penalty on the objective function. This model is a two-layer linear model and $p(\mathbf{y}|\mathbf{x})$ serves as the conjugate prior of $p(\mathbf{z}|\mathbf{y})$. Thus we have

$$\boldsymbol{\beta}\cdot\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}},\boldsymbol{\Sigma}_{\mathbf{y}}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\boldsymbol{\mu}_{\mathbf{A}}, \sigma_1^2\mathbf{I}_{\mathbf{M}})\cdot\mathcal{N}(\mathbf{z}|\mathbf{B}\mathbf{y}+\boldsymbol{\mu}_{\mathbf{B}}, \sigma_2^2\mathbf{I}_{\mathbf{N}}) \tag{5}$$

where $\boldsymbol{\beta}$ is a scalar, $\boldsymbol{\mu}_{\mathbf{y}}$ and $\boldsymbol{\Sigma}_{\mathbf{y}}$ are the mean and variance of a new normal distribution respectively.

From Equations 4 and 5, we have that

$$p(\mathbf{z}|\mathbf{x}) = \int_{\mathbf{y}} \boldsymbol{\beta}\cdot\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}},\boldsymbol{\Sigma}_{\mathbf{y}})\mathrm{d}\mathbf{y} = \boldsymbol{\beta} \tag{6}$$

Thus, maximizing $p(\mathbf{z}|\mathbf{x})$ is equivalent to maximizing $\boldsymbol{\beta}$. Next, we show the derivation of $\boldsymbol{\beta}$. We first derive the value of $\boldsymbol{\mu}_{\mathbf{y}}$ and $\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}$ by comparing the exponential terms on both sides of Equation 5.

$$\begin{aligned}&\mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\boldsymbol{\mu}_{\mathbf{A}}, \sigma_1^2\mathbf{I}_{\mathbf{M}})\cdot\mathcal{N}(\mathbf{z}|\mathbf{B}\mathbf{y}+\boldsymbol{\mu}_{\mathbf{B}}, \sigma_2^2\mathbf{I}_{\mathbf{N}})\\ &= \frac{1}{(2\pi)^{\frac{M+N}{2}}\sigma_1^M\sigma_2^N}\exp\{-\tfrac{1}{2}[\tfrac{1}{\sigma_1^2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\boldsymbol{\mu}_{\mathbf{A}})^{\mathrm{T}}(\mathbf{y}-\mathbf{A}\mathbf{x}-\boldsymbol{\mu}_{\mathbf{A}})\\ &\quad+\tfrac{1}{\sigma_2^2}(\mathbf{z}-\mathbf{B}\mathbf{y}-\boldsymbol{\mu}_{\mathbf{B}})^{\mathrm{T}}(\mathbf{z}-\mathbf{B}\mathbf{y}-\boldsymbol{\mu}_{\mathbf{B}})]\}\end{aligned} \tag{7}$$

The exponential term in Equation 7 can be expanded as

$$\begin{aligned}\boldsymbol{\Psi} &= -\tfrac{1}{2}[\tfrac{1}{\sigma_1^2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\boldsymbol{\mu}_{\mathbf{A}})^{\mathrm{T}}(\mathbf{y}-\mathbf{A}\mathbf{x})\\ &\quad+\sigma_2^2(\mathbf{z}-\mathbf{B}\mathbf{y}-\boldsymbol{\mu}_{\mathbf{B}})^{\mathrm{T}}(\mathbf{z}-\mathbf{B}\mathbf{y})]\\ &= -\tfrac{1}{2}[\tfrac{1}{\sigma_1^2}(\mathbf{y}^{\mathrm{T}}\mathbf{y}-\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x}-\mathbf{y}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{A}}-\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{y}+\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x}\\ &\quad+\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{A}}-\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}}\mathbf{y}+\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}}\mathbf{A}X+\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{A}})+\tfrac{1}{\sigma_2^2}(\mathbf{z}^{\mathrm{T}}\mathbf{z}-\mathbf{z}^{\mathrm{T}}\mathbf{B}\mathbf{y}\\ &\quad-\mathbf{z}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{B}}-\mathbf{y}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{z}+\mathbf{y}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{y}+\mathbf{y}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{B}}-\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\mathbf{z}+\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\mathbf{B}\mathbf{y}\\ &\quad+\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{B}})]\\ &= -\tfrac{1}{2}[\mathbf{y}^{\mathrm{T}}(\tfrac{1}{\sigma_1^2}\mathbf{I}_M+\tfrac{1}{\sigma_2^2}\mathbf{B}^{\mathrm{T}}\mathbf{B})\mathbf{y}-\tfrac{2}{\sigma_1^2}(\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{y}+\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}}\mathbf{y})\\ &\quad-\tfrac{2}{\sigma_2^2}(\mathbf{z}^{\mathrm{T}}\mathbf{B}\mathbf{y}-\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\mathbf{B}\mathbf{y})+\tfrac{1}{\sigma_1^2}(\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x}+2\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}}\mathbf{A}\mathbf{x}+\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{A}})\\ &\quad+\tfrac{1}{\sigma_2^2}(\mathbf{z}^{\mathrm{T}}\mathbf{z}-2\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\mathbf{z}+\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{B}})]\end{aligned} \tag{8}$$

Thus, by comparing the exponential terms on both sides of Equation 5, we get

$$\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} = \frac{1}{\sigma_1^2}\mathbf{I}_M+\frac{1}{\sigma_2^2}\mathbf{B}^{\mathrm{T}}\mathbf{B}, \tag{9}$$

$$\boldsymbol{\mu}_{\mathbf{y}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} = \frac{1}{\sigma_1^2}(\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}+\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}})+\frac{1}{\sigma_2^2}(\mathbf{z}^{\mathrm{T}}\mathbf{B}-\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\mathbf{B}). \tag{10}$$

Further, we have

$$\boldsymbol{\mu}_{\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{y}}[\frac{1}{\sigma_1^2}(\mathbf{A}\mathbf{x}+\boldsymbol{\mu}_{\mathbf{A}})+\frac{1}{\sigma_2^2}(\mathbf{B}^{\mathrm{T}}\mathbf{z}-\mathbf{B}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{B}})]. \tag{11}$$

With $\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}$ and $\boldsymbol{\mu}_{\mathbf{y}}$, we can derive the explicit form of $\boldsymbol{\beta}$ easily by setting $\mathbf{y}=\mathbf{0}$, which leads to the equation below:

$$\begin{aligned}&\boldsymbol{\beta}\cdot\frac{1}{(2\pi)^{\frac{M}{2}}|\boldsymbol{\Sigma}_{\mathbf{y}}|^{\frac{1}{2}}}\exp\{-\tfrac{1}{2}\boldsymbol{\mu}_{\mathbf{y}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\mu}_{\mathbf{y}}\}\\ &= \frac{1}{(2\pi)^{\frac{M+N}{2}}\sigma_1^M\sigma_2^N}\exp\{\boldsymbol{\Psi}_{\mathbf{y}=\mathbf{0}}\},\end{aligned} \tag{12}$$

where $\boldsymbol{\Psi}_{\mathbf{y}=\mathbf{0}}$ is the value of $\boldsymbol{\Psi}$ when $\mathbf{y}=\mathbf{0}$, and thereby

$$\begin{aligned}\boldsymbol{\Psi}_{\mathbf{y}=\mathbf{0}} &= -\tfrac{1}{2}[\tfrac{1}{\sigma_1^2}(\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x}+2\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}}\mathbf{A}\mathbf{x}+\boldsymbol{\mu}_{\mathbf{A}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{A}})\\ &\quad+\tfrac{1}{\sigma_2^2}(\mathbf{z}^{\mathrm{T}}\mathbf{z}-2\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\mathbf{z}+\boldsymbol{\mu}_{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{B}})]\end{aligned} \tag{13}$$

Thus, we get the explicit form of $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = \frac{|\boldsymbol{\Sigma}_{\mathbf{y}}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}\sigma_1^M\sigma_2^N}\exp\{\boldsymbol{\Psi}_{\mathbf{y}=\mathbf{0}}+\tfrac{1}{2}(\boldsymbol{\mu}_{\mathbf{y}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\mu}_{\mathbf{y}})\}. \tag{14}$$

Here, $\boldsymbol{\beta} = p(\mathbf{z}|\mathbf{x},\mathbf{A},\mathbf{B},\boldsymbol{\mu}_{\mathbf{A}},\boldsymbol{\mu}_{\mathbf{B}},\sigma_1,\sigma_2)$ is the likelihood function for one data point $\mathbf{x}$. Let $\mathbf{X}=\{\mathbf{x}_d\}$ and $\mathbf{Z}=\{\mathbf{z}_d\}$ be the sets of D observed data points (genotype and the gene expression profiles for the samples in the study). To maximize $\boldsymbol{\beta}_d$, we can

minimize the negative log-likelihood of $\boldsymbol{\beta}_d$. Thus, our loss function is

$$\begin{aligned}\mathcal{J} &= -\log\prod_{d=1}^{D} p(\mathbf{z}_d|\mathbf{x}_d) \\ &= -\sum_{d=1}^{D}\log p(\mathbf{z}_d|\mathbf{x}_d) \\ &= -\sum_{d=1}^{D}\log \boldsymbol{\beta}_d\end{aligned} \tag{15}$$

Substituting Equation 14 into Equation 15, the expanded form of the loss function is

$$\begin{aligned}&\mathcal{J}(\mathbf{A},\mathbf{B},\boldsymbol{\mu}_\mathbf{A},\boldsymbol{\mu}_\mathbf{B},\sigma_1,\sigma_2) \\ &= \tfrac{D\cdot N}{2}\ln(2\pi) + D\cdot M\ln(\sigma_1) + D\cdot N\ln(\sigma_2) + \tfrac{D}{2}\ln|\boldsymbol{\Sigma}_\mathbf{y}^{-1}| \\ &+\tfrac{1}{2}\sum_{d=1}^{D}\{\tfrac{1}{\sigma_1^2}(\mathbf{x}_d^\mathrm{T}\mathbf{A}^\mathrm{T}\mathbf{A}\mathbf{x}_d + 2\boldsymbol{\mu}_\mathbf{A}^\mathrm{T}\mathbf{A}\mathbf{x}_d + \boldsymbol{\mu}_\mathbf{A}^\mathrm{T}\boldsymbol{\mu}_\mathbf{A}) \\ &+\tfrac{1}{\sigma_2^2}(\mathbf{z}_d^\mathrm{T}\mathbf{z}_d - 2\boldsymbol{\mu}_\mathbf{B}^\mathrm{T}\mathbf{z}_d + \boldsymbol{\mu}_\mathbf{B}^\mathrm{T}\boldsymbol{\mu}_\mathbf{B}) - [\tfrac{1}{\sigma_1^2}(\mathbf{x}_d^\mathrm{T}\mathbf{A}^\mathrm{T} + \boldsymbol{\mu}_A^\mathrm{T}) \\ &+\tfrac{1}{\sigma_2^2}(\mathbf{z}_d^\mathrm{T}\mathbf{B} - \boldsymbol{\mu}_\mathbf{B}^\mathrm{T}\mathbf{B})]\boldsymbol{\Sigma}_\mathbf{y}[\tfrac{1}{\sigma_1^2}(\mathbf{A}\mathbf{x}_d + \boldsymbol{\mu}_A) + \tfrac{1}{\sigma_2^2}(\mathbf{B}^\mathrm{T}\mathbf{z}_d - \mathbf{B}^\mathrm{T}\boldsymbol{\mu}_\mathbf{B})]\}\end{aligned} \tag{16}$$

Taking into account the prior distributions of **A** and **B**, we have that

$$\begin{aligned}&p(\mathbf{z},\mathbf{A},\mathbf{B}|\mathbf{x},\boldsymbol{\mu}_\mathbf{A},\boldsymbol{\mu}_\mathbf{B},\sigma_1,\sigma_2) \\ &= \boldsymbol{\beta}\cdot\mathbf{Laplace}(\mathbf{A}|\mathbf{0},1/\lambda)\cdot\mathbf{Laplace}(\mathbf{B}|\mathbf{0},1/\gamma)\end{aligned} \tag{17}$$

Thus, we can have the ℓ_1-regularized objective function

$$\min_{\mathbf{A},\mathbf{B},\boldsymbol{\mu}_\mathbf{A},\boldsymbol{\mu}_\mathbf{B},\sigma_1,\sigma_2}\prod_{d=1}^{D}\log p(\mathbf{z}_d,\mathbf{A},\mathbf{B}|\mathbf{x}_d,\boldsymbol{\mu}_\mathbf{A},\boldsymbol{\mu}_\mathbf{B},\sigma_1,\sigma_2),$$

which is identical to

$$\min_{\mathbf{A},\mathbf{B},\boldsymbol{\mu}_\mathbf{A},\boldsymbol{\mu}_\mathbf{B},\sigma_1,\sigma_2}[\mathcal{J} + D\cdot(\lambda||\mathbf{A}||_1 + \gamma||\mathbf{B}||_1)], \tag{18}$$

where $||\cdot||_1$ is the ℓ_1-norm. λ and γ are the *precision* of the prior Laplace distributions of **A** and **B** respectively, serving as the regularization parameters which can be determined by cross or holdout validation.

3.3 Optimization

To optimize the objective function, we use the Orthant-Wise Limited-memory Quasi-Newton (OWL-QN) algorithm described in [1]. The OWL-QN algorithm minimizes functions of the form

$$f(w) = loss(w) + C||w||_1,$$

where *loss*$(\cdot)$ is an arbitrary differentiable loss function, and $||w||_1$ is the ℓ_1-norm of the parameter vector. It is based on the L-BFGS Quasi-Newton algorithm [35], with modifications to deal with the fact that the ℓ_1-norm is not differentiable. The algorithm is proven to converge to a local optimum of the parameter vector. The algorithm is very fast, and capable of scaling efficiently to problems with millions of parameters. Thus it is a good option for our problem where the parameter space is large when dealing with large scale eQTL data.

In addition to the loss function and penalized parameters, the OWL-QN algorithm also requires the gradient of the loss function, which (without detailed derivation) is given in the Appendix.

4. EXPERIMENTAL STUDY

4.1 Data set

We apply our method to a yeast eQTL dataset of 112 yeast segregants generated from a cross of two inbred strains: BY and RM ([6, 38]). The dataset originally includes expression profiles of 6229 gene expression traits and genotype profiles of 2956 SNP markers. After removing those SNP markers with percentage of NAs larger than 0.1 (the imcomplete SNPs are imputed), and merging those markers with the same genotypes, we get 1017 SNP markers. Similarly, we drop genes with missing values and have 4474 expression profiles left.

4.2 Baseline Methods

To compare the performance of different approaches, we developed two baseline methods. One method is based on clustering SNP-gene correlation matrix. The other one is based on clustering the resulting bipartite graph of Lasso.

4.2.1 The CoC-Pearson Method

The Pearson's correlation coefficient matrix captures the statistical correlation between SNPs and genes [41]. Let $\mathbf{X} = \{\mathbf{x}_d\} \in \mathbb{R}^{K\times D}$ be the genotype matrix and $\mathbf{Z} = \{\mathbf{z}_d\} \in \mathbb{R}^{N\times D}$ be the phenotype matrix. The Pearson's correlation coefficient matrix of SNPs and genes is $\mathbf{P} = \{\mathbf{P}_{i,j}\}$, where

$$\mathbf{P}_{i,j} = \frac{cov(\mathbf{X}_{i,\cdot},\mathbf{Z}_{j,\cdot})}{\sigma_{\mathbf{X}_{i,\cdot}}\sigma_{\mathbf{Z}_{j,\cdot}}}, \tag{19}$$

i.e., the covariance of a (SNP,gene) pair divided by the product of their standard deviations.

To detect associated SNP sets and gene sets, one straightforward method is to co-cluster the correlation matrix. We adopt the implementation of co-clustering algorithm using information theory proposed in [20]. The algorithm monotonically increases the preserved mutual information by intertwining both the row and column clusterings at all stages and find the optimal co-clustering maximizing the mutual information between the clustered random variables subject to constraints on the number of row and column clusters. The implementation is publicly available from `http://www.cs.utexas.edu/users/dml/Software/cocluster.html`. We refer to this method as CoC-Pearson.

4.2.2 The BGC-Lasso Method

Another baseline method is based on clustering the bipartite graph generated by applying Lasso [45]. We refer to this approach as BGC-Lasso, which consists of the following two steps.

- Step 1: Learn SNP-gene association bipartite graph with Lasso;
- Step 2: Perform bipartite graph clustering.

Lasso is popular method used in detecting SNP-gene associations in eQTL studies. With ℓ_1-penalty, it is suitable for detecting the sparse association bipartite graph in step 1. The objective function of Lasso is

$$\min_{\mathbf{W}}\frac{1}{2}||\mathbf{W}\mathbf{X} - \mathbf{Z}||_2 + \eta||\mathbf{W}||_1 \tag{20}$$

where $||\cdot||_2$ is the ℓ_2-norm. η is the empirical parameter for the ℓ_1 penalty, and **W** is the parameter (also called weight) matrix parameterizing the space of linear functions mapping from **X** to **Z**. The gradient of the least-squares loss function is

$$\begin{aligned}\nabla_\mathbf{W}\mathcal{L}(\mathbf{W}) &= \nabla_\mathbf{W}\tfrac{1}{2}\cdot||\mathbf{W}\mathbf{X} - \mathbf{Z}||_2 \\ &= \tfrac{1}{2}\cdot\nabla_\mathbf{W}\mathrm{tr}[(\mathbf{W}\mathbf{X} - \mathbf{Z})^\mathrm{T}(\mathbf{W}\mathbf{X} - \mathbf{Z})] \\ &= \mathbf{W}\mathbf{X}\mathbf{X}^\mathrm{T} - \mathbf{Z}\mathbf{X}^\mathrm{T}\end{aligned} \tag{21}$$

In step 2, we apply bipartite graph clustering algorithm to identify association cliques. The intuition of the clustering algorithm is shown in Figure 4. For a given number of clusters, we minimize the normalized cut (Ncut) [49], and thus discover a set of dense subgraphs. Each discovered dense subgraph corresponds to a pair of associated SNP set and gene set. We use the algorithm in [9] to extract subgraphs.

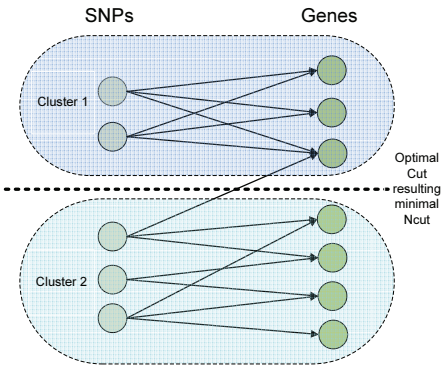


Figure 4: Bipartite graph clustering

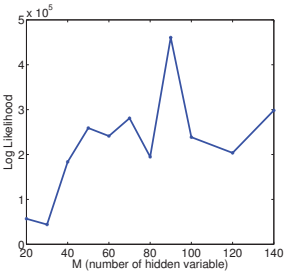


Figure 5: Tuning number of hidden variables

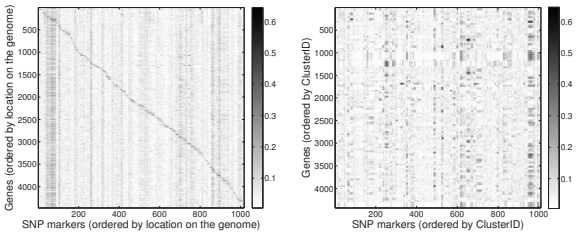
4.3 Parameter Tuning

In the proposed SET-eQTL method, we apply cross validation to tune the three parameters, M, λ and γ. We randomly divide the dataset into two groups of equal size, one as the training set, and the other as the testing set. We first fix the values of λ and γ to tune M. After finding the optimal M, we tune the other two parameters. Figure 5 shows the log likelihood of the data with respect to different M. From the figure, we observe that the optimal value of M is 90. The optimal setting of other parameters are determined similarly with λ=40 and γ=100. A similar approach is apply to tune the parameter η in the BGC-Lasso method. For a fair comparison, we set the number of clusters to be 90 for both BGC-Lasso and CoC-Pearson.

4.4 Gene Ontology Enrichment Analysis

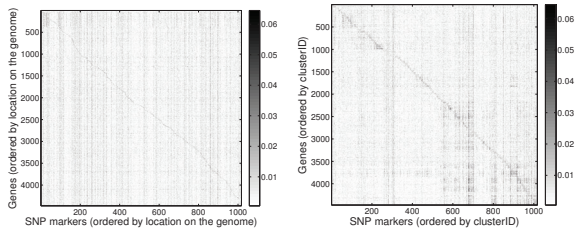
Hidden variables may model the joint effect of SNPs and hidden confounders that have influence on a group of genes. To better understand the learned model, we look for correlations between a set of genes associated with a hidden variable and GO categories (Biological Process Ontology) [44]. In particular, for each gene set H, we identify the GO category whose set of genes is most correlated with H. We measure correlation by a p-value determined by the Fisher's exact test. Since multiple gene sets H need to be examined, the raw p-values need to be calibrated because of the multiple testing problem [48]. To compute calibrated p-values for each H, we perform a randomization test, wherein we apply the same test to 1000 randomly created gene sets that have the same number of genes as H.

In Table 1, each row represents the gene set associated with a hidden variable. For those 90 discovered gene sets, due to space limitation, we only list the 20 gene sets with the smallest calibrated p-values and the 10 gene sets with largest calibrated p-values. The calibrated p-values for the gene sets associated with hidden variables are listed in the third column in the table. The forth column shows the false discovery rate (FDR) [3] of the gene sets. We observed that with an FDR significance threshold 0.05, 74 out of 90 gene sets are significant. These gene sets may represent novel biological pathways. The remaining hidden variables may represent hidden confounders.



(a) Before Clustering (b) After Clustering

Figure 6: Pearson's correlation coefficient matrix



(a) Before Clustering (b) After Clustering

Figure 7: Absolute value of the W Matrix from Lasso

4.5 Comparison with Baseline Methods

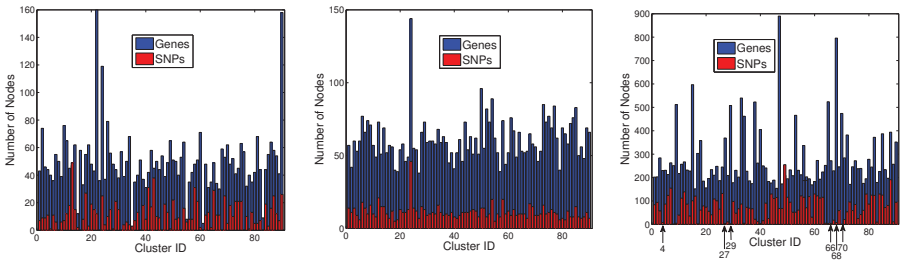
In this section, we analyze the performance of the proposed SET-eQTL method, and the two baseline methods CoC-Pearson and BGC-Lasso.

Figures 6(a) and (b) show the Pearson's correlation coefficient matrices before and after clustering. Figure 7(a) and (b) show the W matrices learned by Lasso before and after clustering. In Figure 6(a) and Figure 7(a), the SNPs and genes are ordered by their locations on the genome. As can be seen, there exists a diagonal line in both of the figures. This is reasonable and indicates strong cis-regulation effect of SNPs to the nearby genes. After clustering, as shown in Figure 6(b), bright blocks representing clusters of SNPs and genes are detected. In Figure 7(b), clustering highlights the diagonal line for the BGC-Lasso method. This indicates that BGC-Lasso favors to cluster genes and SNPs which are closed to each other together, and thus preserves the cis-regulation effect.

Leveraging the results from GO analysis, we are able to better compare the performance of the three methods. Figure 8 and Figure 9 show the number of genes and SNPs within each cluster and the corresponding calibrated p-value (Fisher's exact test) of each discovered gene set. For SET-eQTL, the hidden variable IDs are used as the cluster IDs. It can be seen that the two baseline methods identify less significant gene sets. From Figure 8(c) and Figure 9(c), for the SET-eQTL method, we also observe that the gene sets with large calibrated p-values tend to have very small SNP set associated with them. Those clusters are labeled in both two figures.

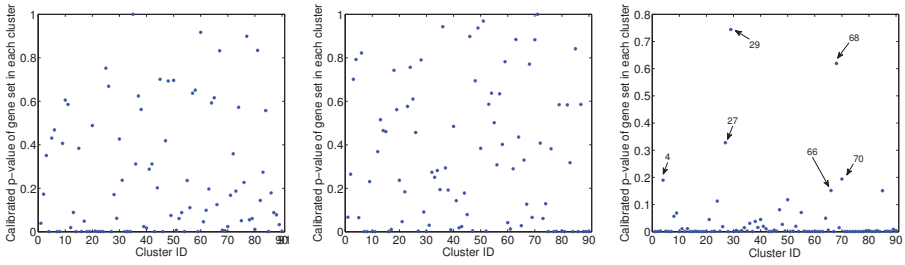
Gene Set Size	Raw p-value	Calibrated p-value	FDR	GO Categories
272	7.89×10^{-11}	0.000999001	0.0019	cellular amino acid biosynthetic process
246	9.73656×10^{-11}	0.000999001	0.0019	cellular amino acid biosynthetic process
193	1.38557×10^{-10}	0.000999001	0.0019	cellular amino acid biosynthetic process
303	1.31797×10^{-09}	0.000999001	0.0019	oxidation-reduction process
175	1.67657×10^{-09}	0.000999001	0.0019	sterol biosynthetic process
245	2.33971×10^{-09}	0.000999001	0.0019	oxidation-reduction process
394	3.89874×10^{-09}	0.000999001	0.0019	cellular amino acid biosynthetic process
358	5.03219×10^{-09}	0.000999001	0.0019	oxidation-reduction process
202	9.33119×10^{-09}	0.000999001	0.0019	cellular amino acid biosynthetic process
203	1.00467×10^{-08}	0.000999001	0.0019	cellular amino acid biosynthetic process
238	1.58219×10^{-08}	0.000999001	0.0019	oxidation-reduction process
217	3.27484×10^{-08}	0.000999001	0.0019	cellular aldehyde metabolic process
233	3.42894×10^{-08}	0.000999001	0.0019	transmembrane transport
185	3.98969×10^{-08}	0.000999001	0.0019	oxidation-reduction process
174	5.49288×10^{-08}	0.000999001	0.0019	cellular amino acid biosynthetic process
239	6.39407×10^{-08}	0.000999001	0.0019	arginine biosynthetic process
156	9.79353×10^{-08}	0.000999001	0.0019	transmembrane transport
284	1.03826×10^{-07}	0.000999001	0.0019	oxidation-reduction process
195	1.22546×10^{-07}	0.000999001	0.0019	oxidation-reduction process
273	1.29062×10^{-07}	0.000999001	0.0019	oxidation-reduction process
...	...	...	...	...
212	0.000405478	0.070929071	0.0798	cellular amino acid biosynthetic process
890	0.000418442	0.080919081	0.0899	cellular aldehyde metabolic process
248	0.00067832	0.112887113	0.1239	oxidation-reduction process
272	0.000839893	0.151848152	0.1248	histidine biosynthetic process
387	0.000887019	0.150849151	0.1248	ion transport
195	0.000897876	0.117882118	0.1248	cellular response to nitrogen starvation
474	0.001130071	0.193806194	0.1964	cytokinesis, completion of separation
230	0.001410782	0.18981019	0.1964	oxidation-reduction process
369	0.002035531	0.327672328	0.3351	oxidation-reduction process
796	0.005892367	0.619380619	0.6263	regulation of transcription by chromatin organization
508	0.007873655	0.744255744	0.7443	RNA processing

Table 1: GO enrichment analysis of the gene sets associated with hidden variables



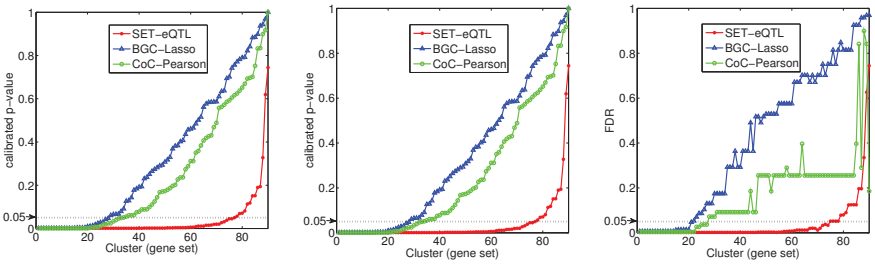
(a) CoC-Pearson (b) BGC-Lasso (c) SET-eQTL

Figure 8: Number of nodes within each cluster



(a) CoC-Pearson (b) BGC-Lasso (c) SET-eQTL

Figure 9: Calibrated p-values of gene sets associated with clusters



(a) Raw p-value (b) Calibrated p-value (c) FDR

Figure 10: Comparison of three methods with GO enrichment analysis

Method	Average raw p-value	Average calibrated p-value	Average FDR	Number of gene sets with calibrated p-values < 0.05
CoC-Pearson	0.002894184	0.248684649	0.170045556	33
BGC-Lasso	0.004638634	0.33035853	0.415928889	28
SET-eQTL	**0.000287769**	**0.037484737**	**0.038482222**	**77**

Table 2: GO enrichment analysis of the gene sets identified by CoC-Pearson, BGC-Lasso and SET-eQTL

This is a strong indicator that these hidden variables may correspond to confounding factors.

To further compare the three methods quantitatively, we apply GO enrichment analysis on the gene sets learned by the three methods. Table 2 shows the average raw p-value, average calibrated p-value, average FDR and the number of significant gene sets (with significance level 0.05 after correction for multiple testing). The original statistics of GO enrichment analysis on the gene sets learned by the three methods are shown in Figure 10. The clusters are arranged by the ascending order with respect to their calibrated p-values. As can been seen from Table 2 and Figure 10, the raw and calibrated p-values, and FDRs of SET-eQTL are all much less than those of BGC-Lasso and CoC-Pearson.

5. CONCLUSION

A crucial challenge in eQTL study is to understand how multiple SNPs interact with each other to jointly affect the expression level of genes. In this paper, we propose a sparse graphical model to identify novel associations between SNP sets and gene sets. The proposed model can also take potential confounding factors into account. ℓ_1-regularization is applied to learn the sparse structure of the graphical model. Using a yeast eQTL data set, we have shown that the proposed method has superior performance over the other two clustering-based methods.The inferred gene sets are strongly correlated with Gene Ontology categories.

6. REFERENCES

[1] G. Andrew and J. Gao. Scalable training of l1-regularized log-linear models. *International Conference on Machine Learning*, 2007.

[2] D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[3] 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 (Methodological), 57(1):289–300, 1995.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] B. R. Bochner. New technologies to assess genotype?henotype relationships. *Nature Reviews Genetics*, 4:309–314, 2003.

[6] R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA*, page 1572Ű1577, 2005.

[7] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing gwas results: A review of statistical methods and recommendations for their application. *American journal of human genetics*, 86(1):6–22, Jan 2010.

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

[9] I. S. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell*, 29:2007, 2007.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[11] C. C. Elbers, K. R. v. Eijk, L. Franke, F. Mulder, Y. T. v. d. Schouw, C. Wijmenga, and N. C. Onland-Moret. Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genetic epidemiology*, 33(5):419–31, Jul 2009.

[12] D. M. Evans, J. Marchini, A. P. Morris, and L. R. Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genetics*, 2: e157, 2006.

[13] Y. Gilad, S. A. Rifkin, and J. K. Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, 24:408–415, Aug 2008.

[14] P. Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York: Springer, 2005.

[15] Y. Guan and J. G. Dy. sparse probabilistic principal component analysis. *International Conference on Aritificial Intelligence and Statistics*, 2009.

[16] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6:95–108, 2005.

[17] J. Hoh and J. Ott. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, 4:701–709, 2003.

[18] J. Hoh, A. Wille, R. Zee, S. Cheng, R. Reynolds, K. Lindpaintner, and J. Ott. Selecting snps in two-stage analysis of disease association data: a model-free approach. *Annals of Human Genetics*, 64:413–417, 2000.

[19] Y. Huang, S. Wuchty, M. T. Ferdig, and T. M. Przytycka. Graph theoretical approach to study eqtl: a case study of plasmodium falciparum. *ISMB*, pages i15–i20, 2009.

[20] S. M. I. S. Dhillon and D. S. Modha. Information-theoretic co-clustering. *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD)*, pages 89–98, 2003.

[21] J. D. S. Jeffrey T. Leek. Capturing heterogeneity in gene expression