Graph-regularized dual Lasso for robust eQTL mapping

Wei Cheng\textsuperscript{1}, Xiang Zhang\textsuperscript{2}, Zhishan Guo\textsuperscript{1}, Yu Shi\textsuperscript{3} and Wei Wang\textsuperscript{4,}\textsuperscript{*}

\textsuperscript{1}Department of Computer Science, UNC at Chapel Hill, Chapel Hill, NC 27599, \textsuperscript{2}Department of EECS, Case Western Reserve University, OH 44106, USA \textsuperscript{3}Department of Mathematics, University of Science and Technology of China, Hefei 23002, China and \textsuperscript{4}Department of Computer Science, University of California, Los Angeles, CA 90095, USA

ABSTRACT

Motivation: As a promising tool for dissecting the genetic basis of complex traits, expression quantitative trait loci (eQTL) mapping has attracted increasing research interest. An important issue in eQTL mapping is how to effectively integrate networks representing interactions among genetic markers and genes. Recently, several Lasso-based methods have been proposed to leverage such network information. Despite their success, existing methods have three common limitations: (i) a preprocessing step is usually needed to cluster the networks; (ii) the incompleteness of the networks and the noise in them are not considered; (iii) other available information, such as location of genetic markers and pathway information are not integrated.

Results: To address the limitations of the existing methods, we propose Graph-regularized Dual Lasso (GDL), a robust approach for eQTL mapping. GDL integrates the correlation structures among genetic markers and traits simultaneously. It also takes into account the incompleteness of the networks and is robust to the noise. GDL utilizes graph-based regularizers to model the prior networks and does not require an explicit clustering step. Moreover, it enables further refinement of the partial and noisy networks. We further generalize GDL to incorporate the location of genetic makers and gene-pathway information. We perform extensive experimental evaluations using both simulated and real datasets. Experimental results demonstrate that the proposed methods can effectively integrate various available prior knowledge and significantly outperform the state-of-the-art eQTL mapping methods.

Availability: Software for both C++ version and Matlab version is available at http://www.cs.unc.edu/~weicheng/.

Contact: weiwang@cs.ucla.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

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 dissect genetic basis of complex traits (Bochner, 2003; Michaelson et al., 2009). Several important issues need to be considered in eQTL mapping. First, the number of SNPs is usually much larger than the number of samples (Tibshirani, 1996). Second, the existence of confounding factors, such as expression heterogeneity, may result in spurious associations (Listgarten et al., 2010). Third, SNPs (and genes) usually work together to cause variation in complex traits (Michaelson et al., 2009). The interplay among SNPs and the interplay among genes can be represented as networks and used as prior knowledge (Musani et al., 2007; Pujana et al., 2007). However, such prior knowledge is far from being complete and may contain a lot of noises. Developing effective models to address these issues in eQTL studies has recently attracted increasing research interests (Biganzoli et al., 2006; Kim and Xing, 2012; Lee and Xing, 2012; Lee et al., 2010).

In eQTL studies, two types of networks can be utilized. One is the genetic interaction network (Charles Boone and Andrews, 2007). Modeling genetic interaction (e.g. epistatic effect between SNPs) is essential to understanding the genetic basis of common diseases, since many diseases are complex traits (Lander, 2011). Another type of network is the network among traits, such as the protein–protein interaction (PPI) network or the gene co-expression network. Interacting proteins or genes in a PPI network are likely to be functionally related, i.e. part of a protein complex or in the same biological pathway (von Mering et al., 2002). Effectively utilizing such prior network information can significantly improve the performance of eQTL mapping (Lee and Xing, 2012; Lee et al., 2010).

Figure 1 shows an example of eQTL mapping with prior network knowledge. The interactions among SNPs and genes are represented by matrices $S$ and $G$, respectively. The goal of eQTL mapping is to infer associations between SNPs and genes represented by the coefficient matrix $W$. Suppose that SNP $\mathcal{A}$ is strongly associated with gene $\mathcal{G}$. Using the network prior, the moderate association between SNP $\mathcal{A}$ and gene $\mathcal{G}$ may be identified since $\mathcal{A}$ is related with $\mathcal{G}$ through other node $\mathcal{B}$ and $\mathcal{C}$ have interactions.

To leverage the network prior knowledge, several methods based on Lasso have been proposed (Biganzoli et al., 2006; Kim and Xing, 2012; Lee and Xing, 2012; Lee et al., 2010). In Biganzoli et al. (2006), the group-Lasso penalty is applied to model the genetic interaction network. In (Kim and Xing, 2012) and (Lee et al., 2010), the authors consider groupings of genes and apply a multi-task Lasso penalty. In (Lee and Xing, 2012), the authors further extend the model to consider grouping information of both SNPs and genes. These methods apply a ‘hard’ clustering of SNPs (genes) so that a SNP (gene) cannot belong to multiple groups. However, a SNP may affect multiple genes and a gene may function in multiple pathways. To address this limitation, in (Jenatton et al., 2011), the authors develop a model allowing overlap between different groups.

Despite their success, there are three common limitations of these group penalty based approaches. First, a clustering step is usually needed to obtain the grouping information. To address this limitation, (Kim and Xing, 2009; Li and Li, 2008) introduce a network-based fusion penalty on the genes. However, this method does not consider the genetic-interaction network. A two-graph-guided multi-task Lasso approach is developed in (Chen et al., 2012) to make use of gene co-expression network and SNP-correlation network. However, this method does not

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
consider the network prior knowledge. The second limitation of the existing methods is that they do not take into consideration the incompleteness of the networks and the noise in them (von Mering et al., 2002). For example, PPI networks may contain false interactions and miss true interactions (von Mering et al., 2002). Directly using the grouping penalty inferred from the noisy and partial prior networks may introduce new bias and thus impair the performance. Third, in addition to the network information, other prior knowledge, such as location of genetic markers and gene-pathway information are also available. The existing methods cannot incorporate such information.

To address the limitations of the existing methods, we propose a novel approach, Graph-regularized Dual Lasso (GDL), which simultaneously learns the association between SNPs and genes and refines the prior networks. To support ‘soft’ clustering (allowing genes and SNPs to be members of multiple clusters), we adopt the graph regularizer to encode structured penalties from the prior networks. The penalties encourage the connected nodes (SNPs/genes) to have similar coefficients. This enables us to find multiple-correlated genetic markers with pleiotropic effects that affect multiple-correlated genes jointly. To tackle the problem of noisy and incomplete prior networks, we exploit the duality between learning the associations and refining the prior networks to achieve smoother regularization. That is, learning regression coefficients can help to refine the prior networks, and vice versa. For example, in Figure 1, if SNPs $\circ$ and $\oplus$ have strong associations with the same group of genes, they are likely to have interaction, which is not captured in the prior network. An ideal model should allow to update the prior network according to the learned regression coefficients. GDL can also incorporate other available prior knowledge such as the physical location of SNPs and biology pathways to which the genes belong. The resultant optimization problem is convex and can be efficiently solved by using an alternating minimization procedure. We perform extensive empirical evaluation of the proposed method using both simulated and real eQTL datasets. The results demonstrate that GDL is robust to the incomplete and noisy prior knowledge and can significantly improve the accuracy of eQTL mapping compared to the state-of-the-art methods.

2 BACKGROUND: LINEAR REGRESSION WITH GRAPH REGULARIZER

Throughout the article, we assume that, for each sample, the SNPs and genes are represented by column vectors. Important notations are listed in Table 1. Let $x = [x_1, x_2, \ldots, x_K]^T$ represent the $K$ SNPs in the study, where $x_i \in \{0, 1, 2\}$ is a random variable corresponding to the $i$-th SNP (e.g. 0, 1, 2 may encode the homozygous major allele, heterozygous allele and homozygous minor allele, respectively). Let $z = [z_1, z_2, \ldots, z_N]^T$ represent expression levels of 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 $x$ and $z$ is

$$z = Wx + \mu + \epsilon,$$

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

The question now is how to define an appropriate objective function over $W$ that (i) can effectively incorporate the prior network knowledge, and (ii) is robust to the noise and incompleteness in the prior knowledge. Next, we first briefly review Lasso and its variations and then introduce the proposed GDL method.

2.1 Lasso and LORS

Lasso (Tibshirani, 1996) is a method for estimating the regression coefficients $W$ using $\ell_1$ penalty for sparsity. It has been widely used for association mapping problems. Let $X = \{x_d | 1 \leq d \leq D\} \in \mathbb{R}^{K \times D}$ be the SNP matrix and $Z = \{z_d | 1 \leq d \leq D\} \in \mathbb{R}^{N \times D}$ be the gene-expression matrix. Each column of $X$ and $Z$ stands for one sample. The objective function of Lasso is

$$\min_{W} \frac{1}{2} ||Z - WX - \mu I||_F^2 + \eta ||W||_1$$

Table 1. Summary of notations

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>Number of SNPs</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of genes</td>
</tr>
<tr>
<td>$D$</td>
<td>Number of samples</td>
</tr>
<tr>
<td>$X \in \mathbb{R}^{K \times D}$</td>
<td>The SNP matrix</td>
</tr>
<tr>
<td>$Z \in \mathbb{R}^{N \times D}$</td>
<td>The gene matrix data</td>
</tr>
<tr>
<td>$L \in \mathbb{R}^{N \times N}$</td>
<td>A low-rank matrix</td>
</tr>
<tr>
<td>$S_0 \in \mathbb{R}^{K \times K}$</td>
<td>The input affinity matrices of the genetic-interaction network</td>
</tr>
<tr>
<td>$G_S \in \mathbb{R}^{N \times N}$</td>
<td>The input affinity matrices of the network of traits</td>
</tr>
<tr>
<td>$S \in \mathbb{R}^{K \times K}$</td>
<td>The refined affinity matrices of the genetic-interaction network</td>
</tr>
<tr>
<td>$G \in \mathbb{R}^{N \times N}$</td>
<td>The refined affinity matrices of the network of traits</td>
</tr>
<tr>
<td>$W \in \mathbb{R}^{N \times K}$</td>
<td>The coefficient matrix to be inferred</td>
</tr>
<tr>
<td>$R(S)$</td>
<td>The graph regularizer from the genetic-interaction network</td>
</tr>
<tr>
<td>$R(G)$</td>
<td>The graph regularizer from the PPI network</td>
</tr>
<tr>
<td>$D(\cdot, \cdot)$</td>
<td>A non-negative distance measure</td>
</tr>
</tbody>
</table>

Fig. 1. Examples of prior knowledge on genetic-interaction network $S$ and gene-gene interactions represented by PPI network or gene co-expression network $G$. $W$ is the regression coefficients to be learned.
where \( \| \cdot \|_F \) denotes the Frobenius norm, \( \| \cdot \|_1 \) is the \( \ell_1 \)-norm, \( I \) is an \( 1 \times D \) vector of all 1’s, \( \eta \) is the empirical parameter for the \( \ell_1 \) penalty and \( W \) is the parameter (also called weight) matrix parameterizing the space of linear functions mapping from \( X \) to \( Z \).

Confounding factors, such as unobserved covariates, experimental artifacts and unknown environmental perturbations, may mask real signals and lead to spurious findings. LORS (Yang et al., 2013) uses a low-rank matrix \( L \in \mathbb{R}^{N \times D} \) to account for the variations caused by hidden factors. The objective function of LORS is

\[
\min_{W, \mu, L, S, G} \frac{1}{2} \| Z - WX - \mu I - L \|_F^2 + \eta \| W \|_1 + \lambda \| L \|_* \tag{3}
\]

where \( \| \cdot \|_* \) is the nuclear norm, \( \eta \) is the empirical parameter for the \( \ell_1 \) penalty to control the sparsity of \( W \) and \( \lambda \) is the regularization parameter to control the rank of \( L \). \( L \) is a low-rank matrix assuming that there are only a small number of hidden factors influencing the gene-expression levels.

### 2.2 Graph-regularized Lasso

To incorporate the network prior knowledge, group sparse Lasso (Biganzoli et al., 2006), multi-task Lasso (Obozinski and Taskar, 2006) and SIOL (Lee and Xing, 2012) have been proposed. Group sparse Lasso makes use of grouping information of SNPs; multi-task Lasso makes use of grouping information of genes, while SIOL uses information from both networks. A common drawback of these methods is that the number of groups (SNP and gene clusters) has to be predetermined. To overcome this drawback, we propose to use two graph regularizers to encode the prior network information. Compared with the previous group penalty-based methods, our method does not need to pre-cluster the networks and thus may obtain smoother regularization. Moreover, these methods do not consider confounding factors that may mask real signals and lead to spurious findings. In this article, we further incorporate the idea in LORS (Yang et al., 2013) to tackle the confounding factors simultaneously.

Let \( S_0 \in \mathbb{R}^{K \times K} \) and \( G_0 \in \mathbb{R}^{N \times N} \) be the affinity matrices of the genetic interaction network (e.g. epistatic effect between SNPs) and network of traits (e.g. PPI network or gene co-expression network), and \( D_{S_0} \) and \( D_{G_0} \) be their degree matrices. Given the two networks, we can employ a pairwise comparison between \( w_{ij} \) and \( w_{ij} \) \((1 \leq i < j \leq K)\): if SNPs \( i \) and \( j \) are closely related, \( \| w_{ij} - w_{ij} \|_1 \) is small. The pairwise comparison can be naturally encoded in the weighted fusion penalty

\[
\sum_{ij} \| w_{ij} - w_{ij} \|_1^2 S_{0}(i,j).
\]

This penalty will enforce \( \| w_{ij} - w_{ij} \|_1^2 \sim 0 \) for closely related SNP pairs (with large \( S_{0}(i,j) \) value). Then, the graph regularizer from the genetic-interaction network takes the following form

\[
\mathcal{R}^{(S)} = \frac{1}{2} \sum_{ij} \| w_{ij} - w_{ij} \|_1^2 S_{0}(i,j) = \text{tr}(W(D_{S_0} - S_0)W^T). \tag{4}
\]

Similarly, the graph regularizer for the network of traits is

\[
\mathcal{R}^{(G)} = \text{tr}(W^T(D_{G_0} - G_0)W). \tag{5}
\]

These two regularizers encourage the connected nodes in a graph to have similar coefficients. A heavy penalty occurs if the learned-regression coefficients for neighboring SNPs (genes) are disparate. \((D_{S_0} - S_0)\) and \((D_{G_0} - G_0)\) are known as the combinatorial graph Laplacian, which are positive semi-definite (Chung, 1997). Graph-regularized Lasso (G-Lasso) solves the following optimization problem

\[
\min_{W, \mu, L, S, G} \frac{1}{2} \| Z - WX - \mu I - L \|_F^2 + \eta \| W \|_1 + \lambda \| L \|_* + \alpha \mathcal{R}^{(S)} + \beta \mathcal{R}^{(G)}.
\]

where \( \alpha, \beta > 0 \) are regularization parameters.

### 3 GDL

In eQTL studies, the prior knowledge is usually incomplete and contains noise. It is desirable to refine the prior networks according to the learned regression coefficients. There is a duality between the prior networks and the regression coefficients: learning coefficients can help to refine the prior networks, and vice versa. This leads to mutual reinforcement when learning the two parts simultaneously.

Next, we introduce the GDL. We further relax the constraints from the prior networks (two graph regularizers) introduced in Section 2.2, and integrate the G-Lasso and the dual refinement of graphs into a unified objective function

\[
\min_{W, \mu, L, S, G} \frac{1}{2} \| Z - WX - \mu I - L \|_F^2 + \eta \| W \|_1 + \lambda \| L \|_* + \alpha \text{tr}(W(D_S - S)W^T) + \beta \text{tr}(W^T(D_G - G)W)
\]

\[
+ \gamma \| S - S_0 \|_F^2 + \rho \| G - G_0 \|_F^2.
\]

where \( \gamma, \rho > 0 \) are positive parameters controlling the extent to which the refined networks should be consistent with the original prior networks. \( D_S \) and \( D_G \) are the degree matrices of \( S \) and \( G \). Note that the objective function considers the non-negativity of \( S \) and \( G \). As an extension, the model can be easily extended to incorporate prior knowledge from multiple sources. We only need to revise the last two terms in Equation (7) to

\[
\gamma \sum_{i=1}^{K} \| S - S_0 \|_F^2 + \rho \sum_{i=1}^{K} \| G - G_0 \|_F^2.
\]

where \( f \) and \( e \) are the number of sources for genetic interaction networks and gene trait networks, respectively.

### 3.1 Optimization: an alternating minimization approach

In this section, we present an alternating scheme to optimize the objective function in Equation (7) based on block coordinate techniques. We divide the variables into three sets: \( \{ L \} \), \( \{ S, G \} \) and \( \{ W, \mu \} \). We iteratively update one set of variables while fixing the other two sets. This procedure continues until convergence. Since the objective function is convex, the algorithm will converge to a global optima. The optimization process is as follows. The detailed algorithm is included in the Supplementary Material (Algorithm 1).

1. While fixing \( \{ W, \mu \} \), \( \{ S, G \} \), optimize \( L \) using singular value decomposition (SVD).
Lemma 3.1. (Mazumder et al., 2010) Suppose that matrix \( A \) has rank \( r \). The solution to the optimization problem

\[
\min_{B} \frac{1}{2} \| A - B \| _F^2 + \lambda \| B \| _* ,
\]

is given by \( \hat{B} = H_\lambda (A) \), where \( H_\lambda (A) = UD_\lambda V^T \) with \( D_\lambda = \text{diag}([d_1 - \lambda], \ldots, [d_r - \lambda]) \). \( UDV^T \) is the Singular Value Decomposition (SVD) of \( A \), \( D = \text{diag}([d_1], \ldots, [d_r]) \), and \( (d_i - \lambda)_+ = \max((d_i - \lambda), 0) \), \( 1 \leq i \leq r \).

Thus, for fixed \( \mu, S, \Gamma \), the formula for updating \( L \) is

\[
L \leftarrow H_\lambda (Z - WX - \mu I).
\]

(2) While fixing \( [W, \mu], [L] \), optimize \([S, \Gamma]\) using seminon-negative matrix factorization (semi-NMF) multiplicative updating on \( S \) and \( \Gamma \) iteratively (Ding et al., 2010). For the optimization with non-negative constraints, our updating rule is based on the following two theorems. The proofs of the theorems are given in Section 3.2.

Theorem 3.2. For fixed \( L, \mu, W \) and \( G \), updating \( S \) according to Equation (10) monotonically decreases the value of the objective function in Equation (7) until convergence.

\[
S \leftarrow S \circ \frac{\alpha(W^T W)^+ + 2\gamma S_0}{2\gamma S + \alpha(W^T W) + \alpha \text{diag}(W^T W) J_K}
\]

where \( J_K \) is a \( K \times K \) matrix of all 1’s. \( \circ \) and \( \parallel \parallel \) are element-wise operators. Since \( W^T W \) may take mixed signs, we denote \( W^T W = (W^T W)^+ - (W^T W)^- \). Where \( (W^T W)^+ = \text{diag}((W^T W)_{ij} + (W^T W)_{ji})/2 \) and \( (W^T W)^- = \text{diag}((W^T W)_{ij} - (W^T W)_{ji})/2 \).

Theorem 3.3. For fixed \( L, \mu, W \) and \( S \), updating \( \Gamma \) according to Equation (11) monotonically decreases the value of the objective function in Equation (7) until convergence.

\[
\Gamma \leftarrow \Gamma \circ \frac{\beta(W W^T)^+ + 2\gamma G_0}{2\gamma G + \beta(W W^T) + \beta \text{diag}(W W^T) J_N}
\]

where \( J_N \) is an \( N \times N \) matrix of all 1’s.

The above two theorems are derived from the Karush–Kuhn–Tucker (KKT) complementarity condition (Boyd and Vandenberghe, 2004). We show the updating rule for \( S \) below. The analysis for \( \Gamma \) is similar and omitted. We first formulate the Lagrange function of \( S \) for optimization

\[
L(S) = \text{atr}(W(D_s - S)W^T) + \gamma ||S - S_0||_F^2.
\]

The partial derivative of the Lagrange function with respect to \( S \) is

\[
\nabla_S L = -\alpha W W^T - 2\gamma S_0 + 2\gamma S + \alpha \text{diag}(W W^T) J_K.
\]

Using the KKT complementarity condition for the non-negative constraint on \( S \), we have

\[
\nabla_S L \circ S = 0.
\]

The above formula leads to the updating rule for \( S \) in Equation (10). It has been shown that the multiplicative updating algorithm has first order convergence rate (Ding et al., 2010).

(3) While fixing \([L], [S, \Gamma]\), optimize \([W, \mu]\) using the coordinate descent algorithm.

Because we use the \( \ell_1 \) penalty on \( W \), we can use the coordinate descent algorithm for the optimization of \( W \), which gives the following updating formula:

\[
W_{ij} = \frac{F(m(i, j), \eta)}{(XX^T)_{ij} + 2\alpha(D_s - S)_{ij} + 2\beta(D_G - G)_{ij}}
\]

where \( F(m(i, j), \eta) = \text{sign}(m(i, j)) \max(|m(i, j)| - \eta, 0) \), and

\[
m(i, j) = (XX^T)_{ij} - \sum_{k=1}^{K} W_{ik} (XX^T)_{kj} - 2\alpha \sum_{k=1}^{K} W_{ik} (D_s - S)_{kj} - 2\beta \sum_{k=1}^{K} (D_G - G)_{ik} W_{kj}.
\]

The solution of updating \( \mu \) can be derived by setting \( \nabla_{\mu} L(\mu) = 0 \), which gives

\[
\mu = \frac{(Z - WX)^T}{D}.
\]

3.2 Convergence analysis

In the following, we investigate the convergence of the algorithm. First, we study the convergence for the second step. We use the auxiliary-function approach (Lee and Seung, 2000) to analyze the convergence of the multiplicative updating formulas. Here we first introduce the definition of auxiliary function.

Definition 3.4. Given a function \( L(h) \) of any parameter \( h \), a function \( Z(h, h) \) is an auxiliary function for \( L(h) \) if the conditions

\[
Z(h, \tilde{h}) \geq L(h) \text{ and } Z(h, h) = L(h),
\]

are satisfied for any given \( h, \tilde{h} \) (Lee and Seung, 2000).

Lemma 3.5. If \( Z \) is an auxiliary function for function \( L(h) \), then \( L(h) \) is non-increasing under the update (Lee and Seung, 2000).

\[
h^{t+1} = \arg\min_{\hat{h}} Z(h, \hat{h}^{t+1}).
\]

Theorem 3.6. Let \( L(S) \) denote the Lagrange function of \( S \) for optimization. The following function
Here $\mathcal{D}(\cdot,\cdot)$ is a non-negative distance measure. Note that the Euclidean distance is used in previous sections. $\mathbf{S}$ and $\mathbf{G}$ are initially given by inputs $\mathbf{S}_0$ and $\mathbf{G}_0$. We refer to this generalized model as the generalized GDL (GGDL). GGDL executes the following two steps iteratively until the termination condition is met: (i) update $\mathbf{W}$ while fixing $\mathbf{S}$ and $\mathbf{G}$ and (ii) update $\mathbf{S}$ and $\mathbf{G}$ according to $\mathbf{W}$, while guarantee that both $\sum_{i,j} D(w_{ij}, w_{ij}) S_{ij}$ and $\sum_{i,j} D(w_{ij}, w_{ij}) G_{ij}$ decrease.

These two steps are based on the aforementioned duality between learning $\mathbf{W}$ and refining $\mathbf{S}$ and $\mathbf{G}$. The detailed algorithm is provided in the Supplementary Material. Next, we illustrate the updating process assuming that $\mathbf{S}$ and $\mathbf{G}$ are unweighted graphs. It can be easily extended to weighted graphs.

Step 1 can be done by using the coordinate decent algorithm. In Step 2, to guarantee that both $\sum_{i,j} D(w_{ij}, w_{ij}) S_{ij}$ and $\sum_{i,j} D(w_{ij}, w_{ij}) G_{ij}$ decrease, we can maintain a fixed number of $1$'s in $\mathbf{S}$ and $\mathbf{G}$. Taking $\mathbf{G}$ as an example, once $G_{ij}$ is selected to change from 0 to 1, another element $G_{k,l}$ with $D(w_{ij}, w_{kl}) < D(w_{ij}, w_{lj})$ should be changed from 1 to 0.

The selection of $(i,j)$ and $(l,j)$ is based on the ranking of $D(w_{ij}, w_{ij})$ $(1 \leq i,j \leq N)$. Specifically, we examine $\kappa$ pairs (the choice of $\kappa$ depends on the user’s belief in the quality of the prior network. For example, it can be $5\%$ of all $(i,j)$ pairs) with the smallest distances. Among them, we pick those having no edges in $\mathbf{G}$. Let $\mathcal{P}_0$ be this set of pairs. Accordingly, we examine $\kappa$ pairs with the largest distances. Among these pairs, we pick up only those having an edge in $\mathbf{G}$. Let $\mathcal{P}_1$ be this set of pairs. The elements of $\mathbf{G}$ corresponding to pairs in $\mathcal{P}_0$ are candidates for updating from 0 to 1, since these pairs of genes are associated with similar SNPs. Similarly, elements of $\mathbf{G}$ corresponding to pairs in $\mathcal{P}_1$ are candidates for updating from 1 to 0.

In this process, the prior knowledge of gene pathways can be easily incorporated to better refine $\mathbf{G}$. For instance, we can further require that only the gene pairs in $\mathcal{P}_0$ belonging to the same pathway are eligible for updating, and only the gene pairs in $\mathcal{P}_1$ belonging to different pathways are eligible for updating. We denote the set of gene pairs eligible for updating by $\mathcal{P}_0'$ and $\mathcal{P}_1'$, respectively. Then, we choose min $|\mathcal{P}_0'|$, $|\mathcal{P}_1'|$ pairs in set $\mathcal{P}_0'$ with smallest $D(w_{ij}, w_{ij})$ $(1 \leq i,j \leq N)$ and update $G_{ij}$ from 0 to 1. Similarly, we choose min $|\mathcal{P}_0'|$, $|\mathcal{P}_1'|$ pairs in set $\mathcal{P}_1'$ with largest $D(w_{ij}, w_{ij})$ $(1 \leq i,j \leq N)$ and update $G_{ij}$ from 1 to 0.

Obviously, all $D(w_{ij}, w_{ij})$'s are smaller than $D(w_{ij}, w_{ij})$ if $\kappa < \frac{N(N-1)}{2}$. Thus, $\sum_{i,j} D(w_{ij}, w_{ij}) G_{ij}$ is guaranteed to decrease. The updating process ends when no such pairs can be found so that switching their values will result in a decrease of the objective function.

The convergence of GGDL can be observed as follows. The decrease of the objective function value in the first step is straightforward since we minimize it using coordinate decent. In the second step, the change of the objective function value is given by

$$-\alpha D(w_{ij}, w_{ij}) + \alpha D(w_{ij}, w_{ij})$$

which is always negative. Thus, in each iteration, the objective function decreases.
function value decreases. Since the objective function is non-negative, the process eventually converges.

**Theorem 4.1.** GGDl converges to the global optimum if both \( \sum_{i,j} D(w_{ij}, w_{ij}) \) and \( \sum_{i,j} D(w_{ij}, w_{ij}) \) are convex to \( W \).

**Proof:** The last two terms in Equation (22) are linear with respect to \( S \) and \( G \), and convex to \( W \) according to the conditions listed. Thus the objective function is convex over all variables. A convergent result to the global optimum can be guaranteed. \( \square \)

## 5 EXPERIMENTS

In this section, we perform extensive experiments to evaluate the performance of the proposed methods. We use both simulated datasets and real yeast eQTL dataset (Brem et al., 2005). For comparison, we select several state-of-the-art methods, including SIOL (Lee and Xing, 2012), two graph guided multi-task lasso (mtllasso2G) (Chen et al., 2012), sparse group Lasso (Biganzoli et al., 2006), sparse multi-task Lasso (Biganzoli et al., 2006), LORS (Yang et al., 2013) and Lasso (Tibshirani, 1996). For all the methods, the tuning parameters were learned using cross validation.

### 5.1 Simulation study

We first evaluate the performance of the selected methods using simulation study. Note that GGDl requires additional prior knowledge and will be evaluated using real dataset.

We adopt the same setup for the simulation study as that in (Lee and Xing, 2012; Yang et al., 2013) and generate synthetic datasets as follows. 100 SNPs are randomly selected from the yeast eQTL dataset (Brem et al., 2005) (112 samples). Ten gene-expression profiles are generated by \( Z_{ij} = W_{ij}X + \Sigma_{ij} + E_{ij}, (1 \leq j \leq 10) \), where \( E_{ij} \sim \mathcal{N}(0, \sigma^2I) \) denotes Gaussian noise. \( \Sigma_{ij} \) is used to model non-genetic effects, which is drawn from \( \mathcal{N}(0, \tau^2) \), where \( \tau = 0.1 \). \( \Sigma \) is generated by \( MM^T \), where \( M \in \mathbb{R}^{10 \times 5} \) and \( M_{ij} \sim \mathcal{N}(0, 1) \). \( c \) is the number of hidden factors and is set to 10 by default. The association matrix \( W \) is generated as follows. Three sets of randomly selected four SNPs are associated with three gene clusters (1–3), (4–6), (7–10), respectively. In addition, one SNP is associated with two gene clusters (1–3) and (4–6), and one SNP is associated with all genes. The association strength is set to 1 for all selected SNPs. The clustering structures among SNPs and genes serve as the ground truth of the prior network knowledge. Only two of the three SNP (gene) clusters are used in \( W \) to simulate incomplete prior knowledge.

Figure 2 shows the estimated \( W \) matrix by various methods. The \( x \)-axis represents traits (1–10) and \( y \)-axis represents SNPs (1–100). From the figure, we can see that GGDl is more effective than G-Lasso. This is because the dual refinement enables a more robust model. G-Lasso outperforms SIOL and mtllasso2G, indicating that the graph regularizer provides a smoother regularization than the hard clustering based penalty. In addition, SIOL and mtllasso2G do not consider confounding factors. SIOL and mtllasso2G outperform multi-task Lasso and sparse group Lasso since it uses both SNP and gene grouping information, while multi-task Lasso and sparse group Lasso only use one of them. We also observe that all methods utilizing prior grouping knowledge outperform LORS and Lasso which cannot incorporate prior knowledge. LORS outperforms Lasso since it considers the confounding factors.

The ground-truth networks, prior networks and GDL-refined networks are shown in Figure 3. Note that only a portion of the ground-truth networks are used as prior networks. In particular, the information related to gene cluster (7–10) is missing in the prior networks. We observe that the refined matrix \( G \) well captures the missing grouping information of gene cluster (7–10). Similarly, many missing pairwise relationships in \( S \) are recovered in the refined matrix (points in red ellipses).

Using 50 simulated datasets with different Gaussian noise \((\sigma^2 = 1 \text{ and } \sigma^2 = 5)\), we compare the proposed methods with alternative state-of-the-art approaches. For each setting, we use 30 samples for test and 82 samples for training. We report the averaged result from 50 realizations. Figure 4 shows the ROC curves of TPR-FPR for performance comparison, together with the areas under the precision-recall curve (AUCs) (Chen et al., 2012). The association strengths between SNPs and genes are set to be 0.1, 1 and 3, respectively. It is clear that GDL outperforms all alternative methods by effectively using and refining the prior network knowledge. We also computed test errors. On average, GDL achieved the best test error rate of 0.9122, and the order of the other methods in terms of the test errors is: G-Lasso (0.9276), SIOL (0.9485), mtllasso2G (0.9521), Multi-task Lasso (0.9723), Sparse group Lasso (0.9814), LORS (1.0429) and Lasso (1.2153).
To evaluate the effectiveness of dual refinement, we compare GDL and G-Lasso since the only difference between these two methods is whether the prior networks are refined during the optimization process. We add noises to the prior networks by randomly shuffling the elements in them. Furthermore, we use the signal-to-noise ratio defined as $\text{SNR} = \frac{\sqrt{WN}}{\sigma} \sqrt{\frac{N}{E}}$ (Yang et al., 2013) to measure the noise ratio in the eQTL datasets. Here, we fix $C = 10$, $r = 0.1$, and use different $\sigma$’s to control SNR.

Figure 5 shows the results for different SNRs. For a fixed SNR, we vary the percentage of noises in the prior networks and compare the performance of selected methods. From the results, we can see that G-Lasso is more sensitive to noises in the prior networks than GDL is. Moreover, when the SNR is low, the advantage of GDL is more prominent. These results indicate using dual refinement can dramatically improve the accuracy of the identified associations.

5.2 Yeast eQTL study

We apply the proposed methods to a yeast (Saccharomyces cerevisiae) eQTL dataset of 112 yeast segregants generated from a cross of two inbred strains (Brem et al., 2005). The dataset originally includes expression profiles of 6229 gene-expression traits and genotype profiles of 2956 SNPs. After removing SNPs with $>10\%$ missing values and merging consecutive SNPs high linkage disequilibrium, we get 1017 SNPs with unique genotypes (Huang et al., 2009). After removing the ones with missing values, 4474 expression profiles are selected. The genetic interaction network is generated as in (Lee and Xing, 2012). We use the PPI network downloaded from BioGRID (http://thebiogrid.org/) to represent the prior network among genes. It takes $\sim$1 day for GGDL, and $\sim$10 h for GDL to run into completion.

5.2.1 cis- and trans-enrichment analysis

We follow the standard cis-enrichment analysis (Listgarten et al., 2010) to compare the performance of two competing models. The intuition behind cis-enrichment analysis is that more cis-acting SNPs are expected than trans-acting SNPs. A two-step procedure is used in the cis-enrichment analysis (Listgarten et al., 2010): (i) for each model, we apply a one-tailed Mann–Whitney test on each SNP to test the null hypothesis that the model ranks its cis hypotheses no better than its trans hypotheses, (ii) for each pair of models compared, we perform a two-tailed paired Wilcoxon sign-rank test on the $P$-values obtained from the previous step. The null hypothesis is that the median difference of the $P$-values in the Mann–Whitney test for each SNP is zero. The trans-enrichment is implemented using similar strategy (Brem et al., 2003), in which genes regulated by transcription factors (obtained from http://www.yeastract.com/download.php) are used as trans-acting signals.

In addition to the methods evaluated in the simulation study, GGDL is also evaluated here (with $\alpha = 100000$, $\eta = 5, \lambda = 8, \alpha = 15, \beta = 1$) for GDL, $\eta = 5, \lambda = 8, \alpha = 15, \beta = 1, \eta = 15, \rho = 1$.) The Euclidean
distance is used as the distance metric. We rank pairs of SNPs and genes according to the learned $W$. $S$ is refined if the locations of the two SNPs are <500 bp. $G$ is refined if the two genes are in the same pathway. The pathway information is downloaded from Saccharomyces Genome Database [SGD (http://www.yeastgenome.org/)].

The results of pairwise comparison of selected models are shown in Table 2. In this table, 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 the proposed GGDL and GDL have significantly better enrichment scores than the other models. By incorporating genomic location and pathway information, GGDL performs better than GDL with $P$-value < 0.0001. The effectiveness of the dual refinement on prior graphs is demonstrated by GDL's better performance over G-Lasso. Note that the performance ranking of these models is consistent with that in the simulation study.

The top-1000 significant associations given by GGDL, GDL and G-Lasso are shown in Figure 6. We can see that GGDL and GDL have stronger cis-regulatory signals than G-Lasso does. In total, these methods each detected ~600 associations according to non-zero $W$ values. We estimate FDR using 50 permutations as proposed in (Yang et al., 2013). With FDR $\leq$ 0.01, GGDL obtains ~4500 significant associations. The plots of all identified significant associations for different methods are given in the Supplementary Material.

### 5.2.2 Refinement of the prior networks

To investigate to what extent GGDL is able to refine the prior networks and study the effect of different parameter settings on $\kappa$, we intentionally change 75% elements in the original prior PPI network and genetic-interaction network to random noises. We feed the new networks to GGDL and evaluate the refined networks. The results are shown in Figure 7. We can see that for both PPI and genetic-interaction networks, many elements are recovered by GGDL. This demonstrates the effectiveness of GGDL. Moreover, when the number of SNP (gene) pairs ($\kappa$) examined for updating reaches 100,000, both PPI and genetic-iteration networks are well refined.
5.2.3 Hotspots analysis

In this section, we study whether GGDL can help detect more biologically relevant associations than the alternatives. Specifically, we examine the hotspots which affect >10 gene traits (Lee and Xing, 2012). The top-15 hotspots detected by GGDL are listed in Table 3. The top-15 hotspots detected by other methods are included in the Supplementary Material. From Table 3, we observe that for all hotspots, the associated genes are enriched with at least one GO category. Note that GGDL and GDL detect one hotspot (12), which cannot be detected by G-Lasso. They also detect one hotspot (6), which cannot be detected by SIOL. The number of hotspots that are significant enriched is listed in Table 4. From the table, we can see that GGDL slightly outperforms GDL since it incorporates the location of SNPs and gene-pathway information.

6 DISCUSSION

As a promising tool for dissecting the genetic basis of common diseases, eQTL study has attracted increasing research interest.

Table 3. Summary of the top-15 hotspots detected by GGDL

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Loci</th>
<th>GO</th>
<th>Hits</th>
<th>GDL (all)</th>
<th>GDL (hits)</th>
<th>G-Lasso (all)</th>
<th>G-Lasso (hits)</th>
<th>SIOL (all)</th>
<th>SIOL (hits)</th>
<th>LORS (all)</th>
<th>LORS (hits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>XII:1056097</td>
<td>(1)**</td>
<td>7</td>
<td>31</td>
<td>32</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>31</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>III:81832..92391</td>
<td>(2)**</td>
<td>5</td>
<td>29</td>
<td>28</td>
<td>5</td>
<td>58</td>
<td>5</td>
<td>22</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>XII:1056103</td>
<td>(1)**</td>
<td>7</td>
<td>29</td>
<td>28</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
<td>III:79091</td>
<td>(2)**</td>
<td>6</td>
<td>29</td>
<td>28</td>
<td>6</td>
<td>28</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>III:175799..177850</td>
<td>(3)*</td>
<td>3</td>
<td>26</td>
<td>23</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>18</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>XII:105925..105930 (1)**</td>
<td>7</td>
<td>27</td>
<td>27</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>III:105042</td>
<td>(2)**</td>
<td>6</td>
<td>23</td>
<td>25</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>19</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>III:201166..201167</td>
<td>(3)**</td>
<td>3</td>
<td>23</td>
<td>22</td>
<td>3</td>
<td>13</td>
<td>2</td>
<td>23</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>XII:1054278..1054302</td>
<td>(1)**</td>
<td>7</td>
<td>26</td>
<td>24</td>
<td>7</td>
<td>24</td>
<td>5</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>III:100213</td>
<td>(2)**</td>
<td>5</td>
<td>23</td>
<td>23</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>III:209932</td>
<td>(3)*</td>
<td>3</td>
<td>21</td>
<td>19</td>
<td>3</td>
<td>16</td>
<td>4</td>
<td>15</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>XII:659357..662627</td>
<td>(4)*</td>
<td>4</td>
<td>19</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>37</td>
<td>9</td>
<td>36</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>19</td>
<td>III:210748..210748</td>
<td>(5)*</td>
<td>4</td>
<td>24</td>
<td>18</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>11</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>19</td>
<td>VIII:111679..111680</td>
<td>(6)*</td>
<td>3</td>
<td>20</td>
<td>3</td>
<td>19</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>19</td>
<td>VIII:111682..111690</td>
<td>(7)*</td>
<td>5</td>
<td>21</td>
<td>20</td>
<td>5</td>
<td>57</td>
<td>6</td>
<td>22</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Total hits</td>
<td>75</td>
<td>74</td>
<td>70</td>
<td>53</td>
<td>77</td>
<td>10</td>
<td>59</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

The traditional eQTL methods focus on testing the associations between individual 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.

Recent advancement in high-throughput biology has made a variety of biological interaction networks available. Effectively integrating such prior knowledge is essential for accurate and robust eQTL mapping. However, the prior networks are often noisy and incomplete. In this article, we propose novel graph-regularized-regression models to take into account the prior networks of SNPs and genes simultaneously. Exploiting the duality between the learned coefficients and incomplete prior networks enables more robust model. We also generalize our model to integrate other types of information, such as SNP locations and gene pathways. The experimental results on both simulated...
and real eQTL datasets demonstrate that our models outperform alternative methods. In particular, the proposed dual refinement regularization can significantly improve the performance of eQTL mapping.

**Funding:** National Institutes of Health (grants R01HG006703 and P50 GM076468-08); NSF IIS-1313606; NSF IIS-1162374 and IIS-1218036.

**Conflict of Interest:** none declared.

**REFERENCES**


