BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 8 2013, pages 1026–1034
doi:10.1093/bioinformatics/btt075

Gene expression

Advance Access publication February 17, 2013

Accounting for non-genetic factors by low-rank representation and sparse regression for eQTL mapping

Can Yang[1], Lin Wang[1,2], Shuqin Zhang[1,3] and Hongyu Zhao[1,*]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA, [2]Center for Theoretical Biology, Peking University, Beijing 100871 and [3]Center for Computational Systems Biology, School of Mathematical Sciences, Fudan University, Shanghai 200433, China

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Expression quantitative trait loci (eQTL) studies investigate how gene expression levels are affected by DNA variants. A major challenge in inferring eQTL is that a number of factors, such as unobserved covariates, experimental artifacts and unknown environmental perturbations, may confound the observed expression levels. This may both mask real associations and lead to spurious association findings.

Results: In this article, we introduce a LOw-Rank representation to account for confounding factors and make use of Sparse regression for eQTL mapping (LORS). We integrate the low-rank representation and sparse regression into a unified framework, in which single-nucleotide polymorphisms and gene probes can be jointly analyzed. Given the two model parameters, our formulation is a convex optimization problem. We have developed an efficient algorithm to solve this problem and its convergence is guaranteed. We demonstrate its ability to account for non-genetic effects using simulation, and then apply it to two independent real datasets. Our results indicate that LORS is an effective tool to account for non-genetic effects. First, our detected associations show higher consistency between studies than recently proposed methods. Second, we have identified some new hotspots that can not be identified without accounting for non-genetic effects.

Availability: The software is available at: http://bioinformatics.med.yale.edu/software.aspx.

Contact: hongyu.zhao@yale.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 1, 2012; revised on January 11, 2013; accepted on February 7, 2013

1 INTRODUCTION

Nowadays both gene expression levels and hundreds of thousands of single-nucleotide polymorphisms (SNPs) can be measured by high-throughput technologies. This allows us to systematically explore the relationship between gene expression levels and genotypes: whether a gene is differentially expressed with different genotypes (or alleles) at a specific locus. The loci that are associated with gene expression levels are known as 'expression quantitative trait loci' (eQTL) (Li *et al*., 2012). Recently, a large number of eQTLs have been found in eQTL studies (Cookson *et al*., 2009). These findings provide insights on how gene expression levels are affected by specific genetic variants (Cheung and Spielman, 2009). They may further help to prioritize disease-associated loci and contribute to disease understanding (Nica and Dermitzakis, 2008).

An important issue in eQTL mapping is that a fairly large proportion of the measured gene expression variations may not be caused by genetic variants, but by some other factors, including cellular state (Alter *et al*., 2000), environmental factors (Gibson, 2008) and experimental conditions (Leek *et al*., 2010). A typical example is the batch effect, which may arise when sub-groups of samples were processed by different laboratories, different technicians or on different days. Because these factors are unrelated to genetic variants, we call them non-genetic factors in the rest of the article.

Some of the non-genetic effects can be directly measured. For example, when the batch information is available, the batch effects may be adjusted, e.g an empirical Bayes method named 'Combat' (Li and Rabinovic, 2007). However, in practice, non-genetic factors may not be directly and completely observable and thus remain hidden. For example, Pastinen *et al*. (2006) showed that cell culture conditions have an unnegligible influence on a large number of genes. Gagnon-Bartsch and Speed (2012) reported that a substantial within-batch effect exists in the Microarray Quality Control study (Shi *et al*., 2006). 'Expression heterogeneity' (EH) arises when these hidden factors are not taken into account in statistical analysis. Leek and Storey (2007) showed that EH not only leads to the reduction of statistical power but also spurious association signals in eQTL mapping.

Recently, capturing EH in gene expression studies has drawn the attention of researchers. Many methods have been proposed to infer the hidden factors by some forms of factor analysis, and adjust the inferred factors as if they were observed (Alter *et al*., 2000; Nielsen *et al*., 2002).

One well-known method that attempts to address these issues is the Surrogate Variable Analysis (SVA; Leek and Storey, 2007). It performs principal component analysis while taking genotypes into consideration and uses permutation to choose the number of principal components. Kang *et al*. (2008) proposed the intersample correlation emended (ICE) eQTL mapping method, in which a linear mixed model was introduced to model the hidden factors. When modeling EH, Kang *et al*. (2008) used the covariance matrix of the gene expression data as the EH covariance matrix in their ICE model. However, this estimate is inconsistent and thus reduces the power of eQTL mapping. Listgarten *et al*. (2010) introduced another linear mixed model, named

[*]To whom correspondence should be addressed.

'LMM-EH', which corrected the inconsistency of the estimated EH covariance matrix. Once the latent covariance matrix has been estimated, LMM-EH can scan every gene-SNP pair. Alternatively, Stegle *et al.* (2010) jointly modeled SNPs, gene probes and hidden confounders into a Bayesian framework. Despite its greatly increased power in eQTL mapping, its heavy computational burden might limit its usage. Fusi *et al.* (2012) proposed another model named 'PANAMA' and borrowed some computational techniques from Gaussian process (Rasmussen and Williams, 2006) and further improved the performance of eQTL mapping. However, during the model optimization, PANAMA may be trapped in a local optimum because the optimization problem is not convex.

In this article, we introduce an alternative formulation to address this issue. We propose a LOw-Rank representation to account for non-genetic factors and make use of Sparse regression for eQTL mapping (LORS). We integrate the low-rank representation and sparse regression into a unified framework, in which SNPs and gene probes can be jointly analyzed. Given the two regularization parameters, the optimization of the model structure is a convex problem. We have developed an efficient algorithm to solve this convex problem and its convergence is guaranteed. We demonstrate its usefulness through its applications to both synthetic data and real data.

2 MODEL

Before introducing our formulation, we summarize the notations used in this article. We consider the following norms of a vector $\mathbf{v} \in \mathbb{R}^n$: the ℓ_1 norm defined as $\|\mathbf{v}\|_1 = \sum_i |v_i|$; the ℓ_2 and the squared ℓ_2 norms defined as $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$ and $\|\mathbf{v}\|_2^2 = \sum_i v_i^2$, respectively. We use the following three norms of a matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$: the Frobenius norm $\|\mathbf{W}\|_F = \sqrt{\sum_{ij} W_{ij}^2}$, the nuclear norm $\|\mathbf{W}\|_* = \sum_{i=1}^r \sigma_i$, where $\sigma_1, \ldots, \sigma_r$ are the singular values of $\mathbf{W}$ and r is the rank of $\mathbf{W}$ and the 'elementwise' ℓ_1 norm $\|\mathbf{W}\|_1 = \sum_{ij} |W_{ij}|$.

Let $\mathbf{Y}$ be an $n \times q$ matrix corresponding to a gene expression dataset, where n is the number of samples and q is the number of genes. Let $\mathbf{X}$ be an $n \times p$ matrix corresponding to a SNP dataset, where p is the number of SNPs. To model the relationship between $\mathbf{Y}$ and $\mathbf{X}$, we propose to decompose $\mathbf{Y}$ as:

$$\mathbf{Y} = \mathbf{1}\mu + \mathbf{XB} + \mathbf{L} + \mathbf{e} \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{p\times q}$ is the coefficient matrix, $\mathbf{1} \in \mathbb{R}^{n\times 1}$ is a vector whose entries are all 1, μ is a $1 \times q$ matrix with $\mu_j, j = 1, \ldots, q$ being the j-th intercept and $\mathbf{e} \in \mathbb{R}^{n\times q}$ is a Gaussian random noise term with zero mean and variance σ^2, i.e. $\mathbf{e}_{ij} \sim \mathcal{N}(0, \sigma^2)$. Here we introduce $\mathbf{L} \in \mathbb{R}^{n\times q}$ in our model to account for the variations caused by a few hidden factors. This model implies that gene expression levels are influenced by genetic factors, non-genetic factors and random noises.

To make the decomposition (1) possible, we make the following assumptions:

- There are only a few hidden factors that may influence gene expression levels. Thus, $\mathbf{L}$ is a low-rank matrix. Here, we also implicitly assume that the hidden factors have global effects rather than local effects.
- The gene expression level may only be affected by a small fraction of SNPs. This implies that the coefficient matrix $\mathbf{B}$ should be sparse.

Based on these assumptions, we propose to solve the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{B},\mu,\mathbf{L}} \ \|\mathbf{Y} - \mathbf{XB} - \mathbf{1}\mu - \mathbf{L}\|_F^2 \\ &\text{s.t.} \qquad \operatorname{rank}(\mathbf{L}) \le r_0, \quad \|\mathbf{B}\|_1 \le t_0 \end{aligned} \tag{2}$$

where $\|\mathbf{B}\|_1$ is the elementwise ℓ_1 norm defined before, r_0 and t_0 are some fixed constants. To make the minimization problem tractable, we relax the rank operator on $\mathbf{L}$ with the nuclear norm, which has been proven to be an effective convex surrogate of the rank operator (Recht *et al.*, 2010). Now we rewrite (2) in a Lagrange form

$$\min_{\mathbf{B},\mu,\mathbf{L}} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{XB} - \mathbf{1}\mu - \mathbf{L}\|_F^2 + \rho\|\mathbf{B}\|_1 + \lambda\|\mathbf{L}\|_* \tag{3}$$

where $\|\mathbf{L}\|_*$ is the nuclear norm of $\mathbf{L}$, ρ and λ are regularization parameters that control the sparsity of $\mathbf{B}$ and the rank of $\mathbf{L}$, respectively. Now it is a convex optimization problem and can be solved efficiently.

Missing data are commonly encountered when analyzing gene expression data. Here we extend our basic model (3) in the following to handle missing data naturally.

Suppose we only observed a subset of entries in $\mathbf{Y}$, indexed by Ω. The unobserved entries are indexed by $\Omega^\perp$. Mathematically, we can define an orthogonal projection operator $\mathcal{P}$ that projects a matrix $\mathbf{W}$ onto the linear space of matrices supported by Ω:

$$\mathcal{P}_\Omega(\mathbf{W})(i,j) = \begin{cases} 0, & \text{if } (i,j) \in \Omega \\ W_{ij}, & \text{if } (i,j) \notin \Omega \end{cases} \tag{4}$$

and $\mathcal{P}_{\Omega^\perp}(\mathbf{W})$ is its complementary projection, i.e. $\mathcal{P}_\Omega(\mathbf{W}) + \mathcal{P}_{\Omega^\perp}(\mathbf{W}) = \mathbf{W}$.

Because we want to find a sparse coefficient matrix $\mathbf{B}$ and a low-rank matrix $\mathbf{L}$ based on the observed data, we propose to solve the following optimization problem:

$$\min_{\mathbf{B},\mu,\mathbf{L}} \quad \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y} - \mathbf{XB} - \mathbf{1}\mu - \mathbf{L})\|_F^2 + \rho\|\mathbf{B}\|_1 + \lambda\|\mathbf{L}\|_* \tag{5}$$

where the first term $\|\mathcal{P}_\Omega(\mathbf{Y} - \mathbf{XB} - \mathbf{1}\mu - \mathbf{L})\|_F^2$ is the sum of squared errors on the observed entries indexed by Ω.

3 ALGORITHM

To solve the optimization problem (3) efficiently, we need the following lemma [the proof can be found in (Mazumder *et al.*, 2010)]:

Lemma 1. *Suppose matrix $\mathbf{W}_{m\times n}$ has rank r. The solution to the optimization problem*

$$\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{W} - \mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}\|_* \tag{6}$$

is given by $\widehat{\mathbf{Z}} = \mathbf{S}_\lambda(\mathbf{W})$ where

$$\mathbf{S}_\lambda(\mathbf{W}) = \mathbf{UD}_\lambda\mathbf{V}^T \ \textit{with}\ \mathbf{D}_\lambda = \mathit{diag}[(d_1 - \lambda)_+, \ldots, (d_r - \lambda)_+] \tag{7}$$

$\mathbf{UDV}^T$ *is the Singular Value Decomposition (SVD) of* $\mathbf{W}$, $\mathbf{D} = \mathit{diag}[d_1, \ldots, d_r]$, *and* $t_+ = \mathit{max}(t, 0)$.

We adopt an alternating strategy to solve problem (3). For fixed $\mathbf{B}$ and μ, the optimization problem becomes

$$\min_{\mathbf{L}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{B}-\mathbf{1}\mu-\mathbf{L}\|_F^2+\lambda\|\mathbf{L}\|_* \tag{8}$$

By Lemma 1, we have a closed-form solution for $\mathbf{L}$:

$$\mathbf{L}=\mathbf{S}_\lambda(\mathbf{Y}-\mathbf{X}\mathbf{B}-\mathbf{1}\mu) \tag{9}$$

For fixed $\mathbf{L}$, the optimization problem becomes

$$\min_{\mathbf{B},\mu} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{B}-\mathbf{1}\mu-\mathbf{L}\|_F^2+\rho\|\mathbf{B}\|_1 \tag{10}$$

It can be further decomposed into q independent Lasso problems (Tibshirani, 1996):

$$\min_{\mathbf{B}_j,\mu_j} \frac{1}{2}\|\mathbf{Y}_j-\mathbf{L}_j-\mathbf{X}\mathbf{B}_j-\mu_j\|_2^2+\rho\|\mathbf{B}_j\|_1, j=1,\ldots,q \tag{11}$$

where $\mathbf{Y}_j, \mathbf{L}_j$ and $\mathbf{B}_j$ are the j-th column of $\mathbf{Y}, \mathbf{L}$ and $\mathbf{B}$, respectively. The Lasso problem can be solved efficiently by the coordinate descent algorithm (Friedman *et al.*, 2007, 2010). Now we have Algorithm 1:

Algorithm 1 A fast algorithm to solve problem (3)

```
• Input: $\mathbf{Y}\in\mathbb{R}^{n\times q}, \mathbf{X}\in\mathbb{R}^{n\times p}, \rho, \lambda$. Initialize $\mathbf{B}\leftarrow\mathbf{0}$, $\mu\leftarrow\mathbf{0}$.
• Iterate until convergence:
    • $\mathbf{L}$-step: $\mathbf{L}=\mathbf{S}_\lambda(\mathbf{Y}-\mathbf{X}\mathbf{B}-\mathbf{1}\mu)$.
    • $(\mathbf{B},\mu)$-step: Solve $q$ independent Lasso problems (11) by the coordinate descent algorithm.
• Output: $\mathbf{B}$, $\mathbf{L}$, $\mu$.
```

So far we have developed the algorithm for solving problem (3). To derive an algorithm to solve optimization problem (5), we need the following lemma [its proof was given by (Mazumder *et al.*, 2010)]:

LEMMA 2. *Soft-impute algorithm*

$$\begin{aligned}\min_{\mathbf{Z}} &\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{W}-\mathbf{Z})\|_F^2+\lambda\|\mathbf{Z}\|_* \\ &=\min_{\mathbf{Z}}\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{W})-[\mathbf{Z}-\mathcal{P}_{\Omega^\perp}(\mathbf{Z})]\|_F^2+\lambda\|\mathbf{Z}\|_* \\ &=\min_{\mathbf{Z}}\frac{1}{2}\|[\mathcal{P}_\Omega(\mathbf{W})+\mathcal{P}_{\Omega^\perp}(\mathbf{Z})]-Z\|_F^2+\lambda\|\mathbf{Z}\|_*\end{aligned} \tag{12}$$

By Lemma 1, the optimal value $\mathbf{Z}^*$ *of the optimization problem (12) can be obtained via updating* $\mathbf{Z}$ *using*

$$\mathbf{Z}\leftarrow\mathbf{S}_\lambda(\mathcal{P}_\Omega(\mathbf{W})+\mathcal{P}_\Omega^\perp(\mathbf{Z})) \tag{13}$$

with an arbitrary initialization.

We also adopt the alternating strategy to solve (5). For fixed $\mathbf{B}$ and μ, optimization problem (5) becomes

$$\min_{L}\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y}-\mathbf{X}\mathbf{B}-\mathbf{1}\mu-\mathbf{L})\|_F^2+\lambda\|\mathbf{L}\|_* \tag{14}$$

By Lemma 2 we have

$$\mathbf{L}\leftarrow\mathbf{S}_\lambda(\mathcal{P}_\Omega(\mathbf{Y}-\mathbf{X}\mathbf{B})+P_\Omega^\perp(\mathbf{L})) \tag{15}$$

For fixed $\mathbf{L}$, optimization problem (5) becomes

$$(\mathbf{B},\mu)=\arg\min_{(\mathbf{B},\mu)}\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y}-\mathbf{L}-\mathbf{X}\mathbf{B}-\mathbf{1}\mu)\|_F^2+\rho\|\mathbf{B}\|_1 \tag{16}$$

Again, this problem can be decomposed into q independent Lasso problems as follows:

$$\min_{(\mathbf{B}_j,\mu_j)}\frac{1}{2}\|\mathcal{P}_{\Omega_j}(\mathbf{Y}_j-\mathbf{L}_j-\mathbf{X}\mathbf{B}_j-\mu_j)\|_2^2+\rho\|\mathbf{B}_j\|_1, j=1,\ldots,q \tag{17}$$

Now we have Algorithm 2:

Algorithm 2 A fast algorithm to solve problem (5)

```
• Input: $\mathbf{Y}\in\mathbb{R}^{n\times q}, \mathbf{X}\in\mathbb{R}^{n\times p}, \rho, \lambda$. Initialize $\mathbf{B}\leftarrow\mathbf{0}$, $\mu\leftarrow\mathbf{0}$.
• Iterate until convergence:
    • $\mathbf{L}$-step: iteratively update $\mathbf{L}$ using (15).
    • $(\mathbf{B},\mu)$-step: Solve $q$ independent Lasso problems (17) using the coordinate descent algorithm.
• Output: $\mathbf{B}$, $\mathbf{L}$, $\mu$.
```

The convergence analysis of our algorithms and the CPU timings are provided in the Supplementary Document.

4 PARAMETER TUNING

We have two parameters that need to be tuned in our models. Here we propose a cross-validation-like strategy to select these two parameters. The idea is as follows: Let Ω be the index of the observed entries of $\mathbf{Y}$. We randomly divide Ω into training entries Ω_1 and testing entries Ω_2: $\Omega_1\bigcup\Omega_2=\Omega$ and $\Omega_1\bigcap\Omega_2=\emptyset$. The sizes of Ω_1 and Ω_2 are roughly the same. We may solve problem (5) on a grid of (ρ,λ) values on the training data:

$$\min_{\mathbf{B},\mu,\mathbf{L}}\quad\frac{1}{2}\|\mathcal{P}_{\Omega_1}(\mathbf{Y}-\mathbf{X}\mathbf{B}-\mathbf{1}\mu-\mathbf{L})\|_F^2+\rho\|\mathbf{B}\|_1+\lambda\|\mathbf{L}\|_* \tag{18}$$

Then we evaluate the prediction error (19) on the testing data

$$Err(\rho,\lambda)=\frac{1}{2}\|\mathcal{P}_{\Omega_2}(\mathbf{Y}-\mathbf{X}\mathbf{B}(\rho,\lambda)-\mathbf{1}\mu(\rho,\lambda)-\mathbf{L}(\rho,\lambda))\|_F^2 \tag{19}$$

where we write $\mathbf{B}$, μ and $\mathbf{L}$ as $\mathbf{B}(\rho,\lambda)$, $\mu(\rho,\lambda)$ and $\mathbf{L}(\rho,\lambda)$ to emphasize that $\mathbf{B}$, μ and $\mathbf{L}$ depend on the parameters ρ and λ. We can then choose the parameter setting (ρ^*,λ^*), which minimizes the prediction error (19).

However, searching for two parameters on a grid of values may be too computationally expensive when dealing with large datasets. Instead, we search a good λ value with fixing $\rho=\infty$ and then perform a one dimensional search on a sequence of ρ values. In our implementation, we first set the maximum rank of $\mathbf{L}$, denoted as $\text{rank}_{\max}(\mathbf{L})$, equal to $\min(n,q)/2$. Then, we start from a large $\lambda_{\max}$, which equals to the second largest singular value of matrix $\mathcal{P}_\Omega(\mathbf{Y})$. After solving

$$\min_{\mathbf{L}}\quad\frac{1}{2}\|\mathcal{P}_{\Omega_1}(\mathbf{Y}-\mathbf{L})\|_F^2+\lambda\|\mathbf{L}\|_* \tag{20}$$

if $\text{rank}(\mathbf{L})<\text{rank}_{\max}(\mathbf{L})$, we reduce λ by a factor $\eta=0.9$ and repeatedly solve (20) until $\text{rank}(\mathbf{L})\geq\text{rank}_{\max}(\mathbf{L})$. Using warm-start, this sequential optimization is efficient (Mazumder *et al.*, 2010).

Then we choose a λ value, which minimizes the prediction error

$$Err(\lambda) = \frac{1}{2}\|\mathcal{P}_{\Omega_2}(\mathbf{Y} - \mathbf{L}(\lambda))\|_F^2 \tag{21}$$

Let $\hat{\lambda}$ be the value corresponding to the minimal prediction error (21). Now we can perform a one dimensional search for a good value for ρ. We generate a sequence of ρ values with length n_ρ equally decreasing from $\rho_{\max}$ to $\epsilon\rho_{\max}$ on the log scale, where $\rho_{\max}$ is the smallest ρ value such that all entries of $\mathbf{B}(\rho, \hat{\lambda})$ are zero. Typically, we set $n_\rho = 20$ and $\epsilon = 0.05$. For each ρ value, we solve

$$\min_{\mathbf{B},\mathbf{L},\mu} \quad \frac{1}{2}\|\mathcal{P}_{\Omega_1}(\mathbf{Y} - \mathbf{1}\mu - \mathbf{XB} - \mathbf{L})\|_F^2 + \rho\|\mathbf{B}\|_1 + \hat{\lambda}\|\mathbf{L}\|_* \tag{22}$$

and evaluate the prediction error:

$$Err(\rho, \hat{\lambda}) = \frac{1}{2}\|\mathcal{P}_{\Omega_2}(\mathbf{Y} - \mathbf{1}\mu(\rho, \hat{\lambda}) - \mathbf{XB}(\rho, \hat{\lambda}) - \mathbf{L}(\rho, \hat{\lambda}))\|_F^2 \tag{23}$$

Then we choose the ρ value corresponding to the minimal prediction error (23). Now we can solve model (5) using $(\hat{\rho}, \hat{\lambda})$ as regularization parameters, and obtain a sparse matrix $\mathbf{B}(\hat{\rho}, \hat{\lambda})$ and a low rank matrix $\mathbf{L}(\hat{\rho}, \hat{\lambda})$.

5 DISCUSSION

5.1 Relationship between our method and other methods

To our knowledge, LMM-EH (Listgarten *et al.*, 2010) proposed the first framework, where multiple gene expression levels and confounder effects can be jointly analyzed in eQTL studies. For the *j*-th gene expression level in the LMM-EH model, it assumes the following structure:

$$\mathbf{Y}_j = \mathbf{XB}_j + \mathbf{u}_j + \mathbf{e}_j, j = 1, \ldots, q \tag{24}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times q}, \mathbf{X} \in \mathbb{R}^{n\times p}$ are the expression and SNP data matrices, respectively. Here $\mathbf{e}_j$ denotes Gaussian noise, i.e. $\mathbf{e}_j \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$, and $\mathbf{u}_j$ denotes a random effect, i.e. $\mathbf{u}_j \sim \mathcal{N}(0, \tau\Sigma)$, where τ is a scalar and $\Sigma \in \mathbb{R}^{n\times n}$. Assuming the independence among $\mathbf{Y}_j, j = 1, \ldots, q$, and integrate out $\mathbf{u}_j$ and $\mathbf{e}_j$, we arrive at the following form:

$$\Pr(\mathbf{Y}|\mathbf{X}, \{\mathbf{B}, \tau, \sigma_e\}, \Sigma) = \prod_{j=1}^{q} \Pr(\mathbf{XB}_j, \tau^2\Sigma + \sigma_e^2\mathbf{I}) \tag{25}$$

LMM-EH adopts the following strategy to estimate the covariance matrix Σ and other model parameters $\Theta = \{\mathbf{B}, \tau, \sigma_e\}$:

- First, it estimates Σ from the null model, which does not include any SNPs, denoted as $\widehat{\Sigma}$ (Kang *et al.*, 2008, 2010; Lippert *et al.*, 2011).
- Second, using $\widehat{\Sigma}$ in model (24) as a known covariance and estimate $\Theta = \{\mathbf{B}, \tau, \sigma_e\}$ for all gene-SNP pairs (one gene versus one SNPs at a time).

PANAMA extends LMM-EH and allows joint analysis of all SNPs (Fusi *et al.*, 2012). Specifically, PANAMA models the relationship between gene expression levels and SNPs as follows:

$$\mathbf{Y} = \mu + \mathbf{XB} + \mathbf{HW} + \mathbf{e} \tag{26}$$

here μ is the intercept, $\mathbf{B}$ and $\mathbf{W}$ are the corresponding coefficients representing the effects of SNPs and hidden factors $\mathbf{H}$. PANAMA assigns independent Gaussian priors for $\mathbf{B}$ and $\mathbf{W}$:

$$\Pr(\mathbf{B}) = \prod_{i=1}^{p} \mathcal{N}(\mathbf{0}, \alpha_i^2\mathbf{I}), \quad \Pr(\mathbf{W}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{0}, \beta_k^2\mathbf{I}) \tag{27}$$

where K is the number of hidden factors. Assuming $\mathbf{Y}_j, j = 1, \ldots$, q are independent and integrating out $\mathbf{B}$ and $\mathbf{W}$, the model becomes

$$\Pr(\mathbf{Y}|\mathbf{H}, \Theta) = \prod_{j=1}^{q} \mathcal{N}(\mathbf{0}, \sum_{i=1}^{p} \alpha_i^2\mathbf{X}_i\mathbf{X}_i^T + \sum_{k=1}^{K} \beta_k^2\mathbf{H}_k\mathbf{H}_k^T + \sigma_e^2\mathbf{I}), \tag{28}$$

where the intercept term μ is dropped for notation convenience and $\Theta = \{\{\alpha_i^2\}, \{\beta_k^2\}, \sigma_e^2\}$. In principle, parameter estimation in (28) can be done by borrowing some computational tricks from Gaussian process model optimization (Rasmussen and Williams, 2006). Computation becomes prohibitive when all genome-wide SNPs are included. In this case, PANAMA adopts a heuristic strategy: PANAMA begins with the null model (i.e. the model does not include SNPs). It first uses principal components to initialize $\mathbf{H}$ and gradually adds significantly associated SNPs into model (28), and re-estimate model parameter Θ and $\mathbf{H}$. This process iterates until no significantly associated SNPs are added into the model. In summary, $\mathbf{H}$ and Θ are jointly optimized during the iterations.

Our model (1) can be regarded as an equivalent form of (26) because a low rank matrix can always be written as $\mathbf{L} = \mathbf{HW}$ with $\mathbf{H} \in \mathbb{R}^{n\times r}, \mathbf{W} \in \mathbb{R}^{r\times q}$, where r is the rank of $\mathbf{L}$. Unlike PANAMA, both our formulations (3) and (5) are joint convex w.r.t ($\mathbf{B}$, μ, $\mathbf{L}$). When the tuning parameters (λ and ρ) are given, our algorithms are guaranteed to converge to the optimal solution without any heuristic. Furthermore, we do not assume that $\mathbf{Y}_i, i = 1, \ldots, q$ are independent. This can be seen from Lemma 1 and Lemma 2: information among multiple gene expression is used jointly by singular value decomposition.

Compared with PANAMA, the proposed method LORS has its disadvantages. Using PANAMA's formulation, statistical significance of the associations can be evaluated. Currently, we can not provide a rigorous statistical significance test of the estimated coefficient matrix $\mathbf{B}$. The difficulty comes from the unknown statistical property of the nuclear norm. How to do statistical tests with the nuclear norm regularization needs to be investigated in the future. In this article, we use permutation to obtain a rough estimate of false discovery rate (FDR) for our method.

5.2 A screening method based on LORS

Although optimization of our LORS model (3) is a convex problem, it is still too computationally intensive to directly use it for analyzing human-size datasets (e.g. the number of genes $q \approx 20\,000$, the number of SNPs $p \approx 500\,000$). One can see the computational bottleneck in the ($\mathbf{B}$, μ) step of Algorithm 1 and 2. In this step, q Lasso problems need to be solved, each of which involves p variables. To overcome this computational difficulty, we propose to solve the following optimization problem:

$$\min_{\beta_j,\mu,\mathbf{L}} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{X}_j\beta_j - \mathbf{1}\mu - \mathbf{L}\|_F^2 + \lambda\|\mathbf{L}\|_* \tag{29}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times q}$ is the entire data matrix of gene expression, $\mathbf{X}_j \in \mathbb{R}^{n\times 1}$ is the j-th SNP, $\beta_j \in \mathbb{R}^{1\times q}$ is the coefficient of the j-th SNP corresponding to its effect size on q genes. Here we consider one SNP at a time, and thus, we do not add L_1 regularization. Clearly, it can be considered as the single-variable version of LORS. Thus, we call this screening method as 'LORS-Screening'. The algorithm to solve (29) is given in the Supplementary Documents. The computational time of LORS-Screening is given in the Supplementary Material.

For large datasets (e.g. human datasets), we recommend to use LORS-Screening to reduce the number of SNPs. After the screening process, we may select top d SNPs for each gene (based on the absolute value of the coefficients). Then we can fit the LORS model using the selected SNPs. This strategy is similar to the single-variable screening step followed by joint analysis in linear regression (Fan and Lv, 2008). According to the property of L_1 regularization, LORS can identify at most n non-zero coefficients for each gene. Here we may set $d = n$.

6 RESULTS

6.1 Synthetic data

To avoid the simulation setup favoring our own model, we use LMM-EH model (24). Specifically, we generate genetic effects, non-genetic effects and noises as follows:

- Genetic effects: each SNP is generated independently and the minor allele frequencies of these SNPs are uniformly distributed in the interval (0.1, 0.4). The coefficient matrix $\mathbf{B}$ is a sparse matrix with 1% non-zero entries. These non-zero coefficients are generated using standard Gaussian distribution. Let $\mathbf{G}$ denote the genetic effect $\mathbf{G} = \mathbf{XB}$.
- Non-genetic effects: The covariance matrix Σ is generated by $\mathbf{HH}^T$, where $\mathbf{H} \in \mathbb{R}^{n\times K}$ and $\mathbf{H}_{i,j} \sim \mathcal{N}(0, 1)$. Here K is the number of hidden factors. The random effect $\mathbf{u}_j$ is drawn from $\mathcal{N}(\mathbf{0}, \tau\Sigma)$. Let $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_q]$.
- $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$.

Now we have

$$\mathbf{Y} = \mathbf{XB} + \mathbf{u} + \mathbf{e} = \mathbf{G} + \mathbf{u} + \mathbf{e} \tag{30}$$

In the following simulation studies (Sections 6.2 and 6.3), we set $n = 100$, $p = 100$ and $q = 200$. To evaluate the performance under different signal-to-noise ratios, we define SNR_1 and SNR_2 as:

$$\mathrm{SNR}_1 = \sqrt{\frac{\mathrm{Var}(\mathbf{G})}{\mathrm{Var}(\mathbf{e})}}, \quad \mathrm{SNR}_2 = \sqrt{\frac{\mathrm{Var}(\mathbf{G})}{\mathrm{Var}(\mathbf{u})}} \tag{31}$$

Parameters τ and σ_e can be used to control SNR_1 and SNR_2. An example of synthesized datasets is given in the Supplementary Document.

6.2 Influence of parameter tuning

Before we compare LORS with some other methods, we would like to empirically evaluate our parameter tuning procedure. Given a dataset, we need to randomly partition the observed entries into two parts: Ω_1 and Ω_2. Basically, we train our model based on Ω_1 for different parameters and choose a good parameter configuration such that the trained model has an accurate prediction on Ω_2. There may be two concerns: (i) Because the random partition may introduce randomness in our modeling process, does this strategy provide a stable parameter selection? (ii) Can this strategy adapt to different noise level? To answer these questions, we do 100 random partitions of a synthetic dataset, and run our method based on each partition separately. The distribution of the selected parameters is shown in Figure 1. First, one can see that the selected parameters (λ, ρ) do not change a lot during 100 random partitions. The stability of our method should be attributed to the continuity property of the ℓ_1 norm (Fan and Li, 2001), that is, a small change of dataset will not cause a big change of the optimization solution. Second, when the signal becomes weaker, i.e. $\mathrm{SNR}_1 = 1$ (the left panel of Fig. 1) reduces to $\mathrm{SNR}_1 = 1/2$ (the right panel of Fig. 1), a larger ρ will be selected to prevent the noise from entering the model. This shows that our parameter tuning strategy can adapt to different noise levels. In Section 3.2 of the Supplementary Document, we provide more evidence to show that the random partition in our parameter tuning has little effects on eQTL mapping (i.e. the estimation of matrix $\mathbf{B}$).

6.3 Performance evaluation

We will mainly compare our method LORS with PANAMA. The reasons are: (i) PANAMA can be regarded as an extension of LMM-EH as we discussed above. (ii) Fusi *et al.* (2012) showed that PANAMA significantly outperforms other related methods, including SVA, PEER and ICE. Here we include the results from standard linear regression as reference, and compare LORS, LORS-Screening and PANAMA with the standard linear regression.

To compare our method with PANAMA under different settings, we vary SNR_1, SNR_2 and K. For each setting, we report the averaged result from 50 realizations. Figure 2 shows the comparison results for different combinations of SNR_1,

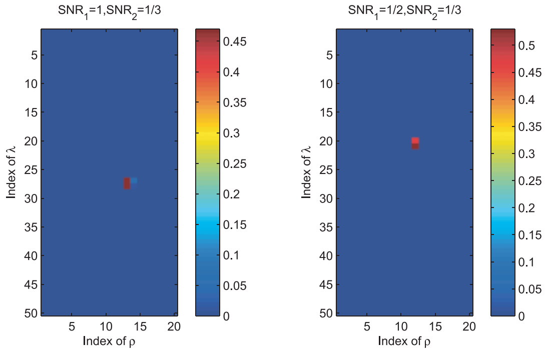

Fig. 1. The distribution of selected parameters (100 random partitions of training and testing data) for synthetic datasets. Left panel: the synthetic dataset is generated with $n = 100$, $p = 100$, $q = 200$, $\mathrm{SNR}_1 = 1$ and $\mathrm{SNR}_2 = 1/3$. For the parameter λ, λ_{27} and λ_{28} are selected in the sequence of λ values in most cases; for parameter ρ, ρ_{13} is selected in most cases. Right panel: the synthetic dataset is generated with $n = 100$, $p = 100$, $q = 200$, $\mathrm{SNR}_1 = 1/2$ and $\mathrm{SNR}_2 = 1/3$. For λ, λ_{20} and λ_{21} are often selected; for ρ, ρ_{12} is selected in most cases

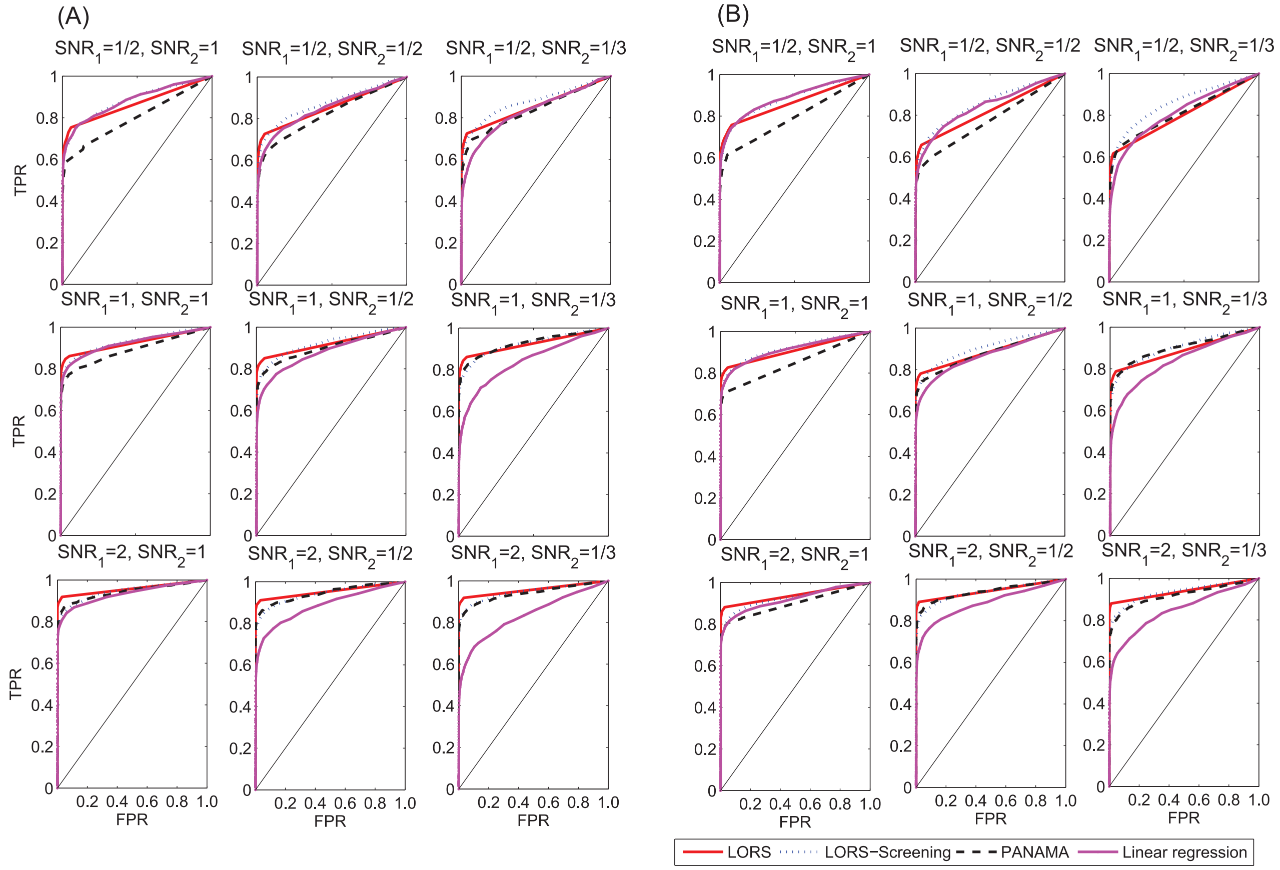

Fig. 2. The ROC curves for performance comparison. (**A**) The number of hidden factors $K = 10$. (**B**) The number of hidden factors $K = 30$. In each panel, we vary $\text{SNR}_1 = \{1/2, 1, 2\}, \text{SNR}_2 = \{1, 1/2, 1/3\}$ to compare the performance of LORS, PANAMA and standard linear regression

SNR_2 and K (more simulation results can be found in the Supplementary Document). From these simulation results, we can see the following:

- When the number of hidden factors increases, the performance of both LORS and PANAMA degrades slightly.
- When the genetic effects and non-genetic effects are small (compared with noises), e.g. $\text{SNR}_1 = 1/2$, $\text{SNR}_2 = 1$), LORS is only comparable with standard linear regression and PANAMA is even worse. This is because the noise plays a dominant role here, it is difficult to account for non-genetic effects under this situation.
- As the genetic effects and non-genetic effects become more apparent, both LORS and PANAMA perform better than standard linear regression. As we mentioned in Section 5.1, LORS and PANAMA share the same model structure, and they differ in how the model structure is inferred. We suspect that PANAMA may be trapped at a local optimum during its model optimization. As a result, LORS may have better performance than PANAMA.
- Regarding to LORS-Screening, it turns out that LORS-Screening is slightly worse than LORS but comparable with PANAMA. Because the computational cost is largely reduced, it is preferred in large data analysis.

6.4 Estimate of false discovery rate

Owing to lack of statistical tests, it is necessary to provide a way to estimate the FDR of our method. Because correlation exists among the rows and columns of $\mathbf{Y}$, exactly estimating FDR becomes difficult. Here we follow the strategy of (Tibshirani and Wang, 2008; Nowak *et al.*, 2011) to obtain a rough estimator of the true FDR, which may serve as a guideline when applying our method. We use

$$\widehat{FDR}_\tau = \frac{\mathcal{N}_\tau}{\mathcal{A}_\tau} \tag{32}$$

as a rough estimator for FDR, where $\mathcal{N}_\tau$ is the number of associations identified at threshold τ under the null distribution, $\mathcal{A}_\tau$ is the number of associations identified at threshold τ in the original dataset. We use permutation to obtain the number of associations identified under the null distribution. Specifically, for a given threshold τ, we do T permutations. At the t-th permutation, we permute the rows of the expression dataset $\mathbf{Y}$ to generate a null dataset, denoted as $\widetilde{\mathbf{Y}}^{(t)}$. Then, we run LORS on the permuted dataset $(\widetilde{\mathbf{Y}}^{(t)}, \mathbf{X})$ and obtain the number of associations by applying the threshold τ to the estimated matrix $\widehat{\mathbf{B}}$, denoted as $\widetilde{\mathcal{A}}_\tau^{(t)}$. After T permutations, the final estimation of $\widehat{FDR}_\tau$ is given by

$$\widehat{FDR}_\tau = \frac{\mathcal{N}_\tau}{\mathcal{A}_\tau} = \frac{\frac{1}{T}\sum_{t=1}^{T}\widetilde{\mathcal{A}}_\tau^{(t)}}{\mathcal{A}_\tau} \tag{33}$$

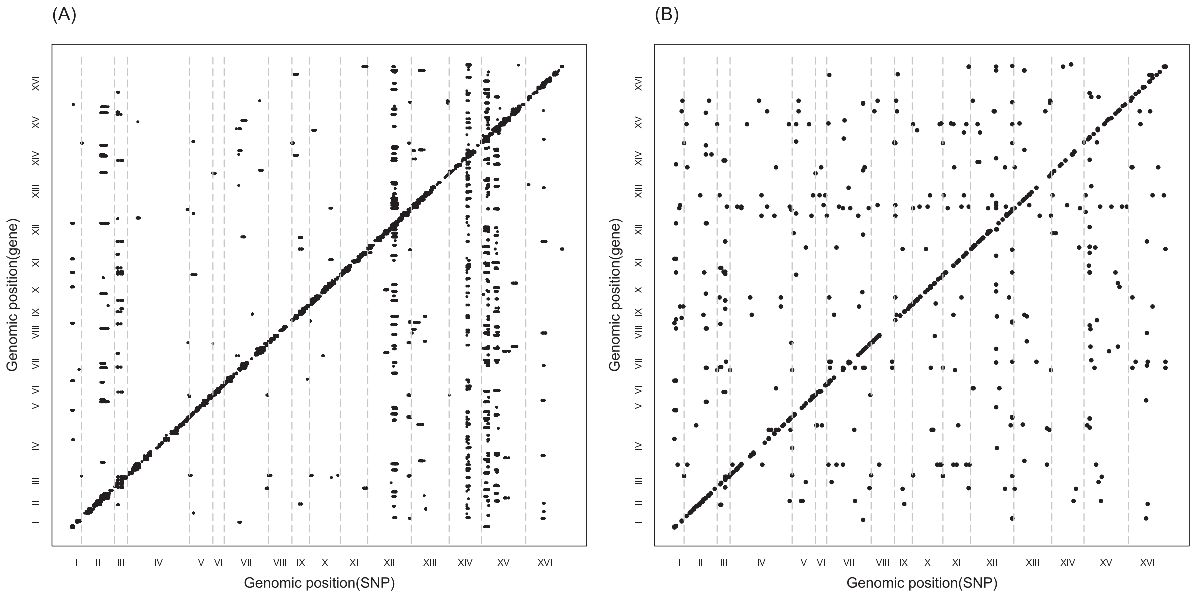

Fig. 3. (**A**) A plot of significant linkage peaks given by standard linear regression ($P < 0.01$ after Bonferroni correction) for expression QTL in the study (Smith and Kruglyak, 2008) by marker location (x-axis) and expression trait location (y-axis). (**B**) The plot of linkage peaks in the study (Smith and Kruglyak, 2008) given by LORS (Top 1000 associations based on abs (**B**) are shown here. The plot of All associations are given in the supplementary document)

We report estimated FDR based on the simulated data as described in Section 6.1 in Figure 9 of the Supplementary Document. The permutation strategy provides a reasonably good estimate for the true FDR. The estimated FDR often overestimates the true FDR (due to the correlation), and thus, it may serve as a conservative guideline.

6.5 Application to eQTL data from yeast

We applied our method to two yeast datasets for eQTL mapping. The first dataset is from Brem *et al.* (2002) (GEO accession number GSE 1990), which consists of 7084 probes and 2956 genotyped loci in 112 segregates. The second one comes from Smith and Kruglyak (2008), which includes 5493 probes measured in 109 segregates. Analysis of these two datasets provides us an opportunity to demonstrate the benefit of our method because the two expression data share the same genetic effect but different confounding effects.

The significant linkage peaks given by standard linear regression are shown in Figure 3 (A). We can clearly see the confounding effects that lead to spurious associations. We also show the result given by LORS in Figure 3 (B). In total, LORS has detected about 10 000 associations according to non-zero **B** values. Because LORS does not perform statistical significance tests, we are not able to report our result based on statistical significance. In practice, people may be more interested in the top signals that will be followed up for replications. Thus, we only show the top 1000 associations based on the absolute value of **B**. The plots of all associations are also given in the Supplementary Document for completeness. It can be seen that the confounding effects are successfully accounted by LORS, and thus spurious associations are greatly reduced. To quantitatively evaluate the ability of accounting for confounding effects, we compare the reproducibility of the results given by LORS and PANAMA. We examine the reproducibility based on the following two criteria:

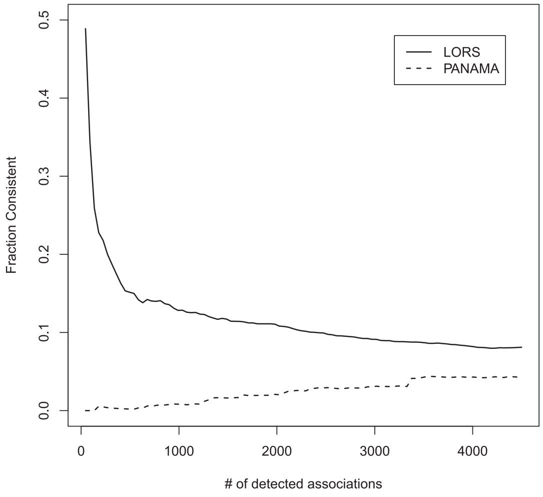

Fig. 4. Consistency of detected associations between two independent yeast eQTL datasets

- The consistency of detected SNP-gene associations. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the sets of SNP-gene associations detected in the two yeast datasets, respectively. The most T significant associations from the two datasets are denoted as $\mathcal{S}_1^T$ and $\mathcal{S}_2^T$. The consistency is defined as $\frac{|\mathcal{S}_1^T \cap \mathcal{S}_2^T|}{T}$, where $|\mathcal{S}_1^T \cap \mathcal{S}_2^T|$ denotes the size of $\mathcal{S}_1^T \cap \mathcal{S}_2^T$. For LORS, the ranking is based on the absolute value of **B**. For PANAMA, the ranking is based on the q-value.

- The consistency of detected hotspots. For a SNP, we can count the number of associated genes from the detected SNP-gene associations (for LORs, all SNP–Gene pairs with a non-zero B are defined as associations. For PANAMA, SNP–Gene pairs with a q-value <0.001 are defined as associations, we tried different cutoffs from 0.01 to 0.001, the results are similar), i.e. the regulatory degree of the SNP. SNPs with large degrees are often referred to as hotspots. According to SNPs' regulatory degrees, we sort them in a descending order and denote the sorted SNPs lists as $\mathcal{L}_1$ and $\mathcal{L}_2$ for the two yeast datasets. Let $\mathcal{L}_1^T$ and $\mathcal{L}_2^T$ be the top T SNPs in the sorted SNP lists, respectively. The consistency of detected hotspots is defined as $\frac{|\mathcal{L}_1^T \cap \mathcal{L}_2^T|}{T}$.

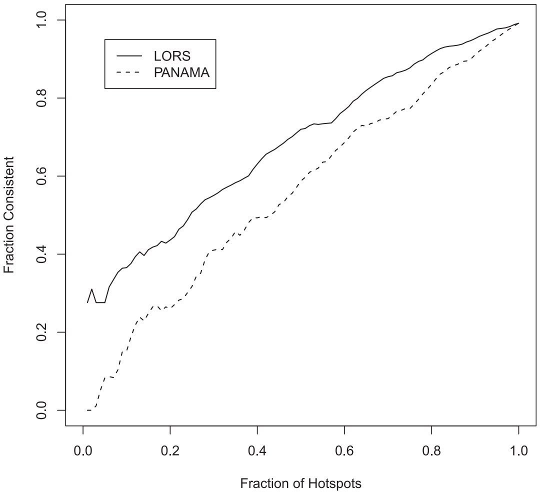

Fig. 5. Consistency of detected eQTL hotspots between two independent yeast eQTL datasets

For Brem's dataset (Brem *et al.*, 2002), the estimated sparse matrix **B** given by LORS has about 6000 non-zero entries in total. Among them, there are 4500 entries with abs $(\mathbf{B}) > 0.01$ and 2500 entries with abs $(\mathbf{B}) > 0.05$, respectively. For Simth's dataset (Smith and Kruglyak, 2008), the estimated **B** has about 10 000 non-zero entries in total. There are about 4500 entries with abs $(\mathbf{B}) > 0.05$. (To provide a meaningful guideline of the thresholds, we estimate FDR using 50 permutations. The estimated FDR corresponding to different thresholds are provided in the Supplementary Document. It tells us that FDR $\leq$ 0.01 when threshold $\tau \geq 0.01$). In Figure 4, we show the consistency of the top 4500 associations. The consistencies of hotspots are shown in Figure 5. It seems to be counter-intuitive that the fraction of consistency of PANAMA increases as the number of detected association increases. In fact, the consistency of PANAMA increases to 0.12 and then drops. We provide the detailed information in the Supplementary Document. From Figures 4 and 5, it can be seen that LORS achieves better consistency than PANAMA.

So far we have shown that spurious associations can be reduced by successfully accounting for non-genetic effects. Now we are going to show whether it could help to detect more biologically relevant associations. We take a closer inspection of the top 15 hotspots, as listed in Table 1. In most cases (12/15), associated genes are enriched with at least one GO category, which implies that they are biologically relevant findings. In particular, we detect two novel hotspots (NO. 9 and NO. 13), which can not be detected by standard linear regression (adjusted P >0.1). For these two hotspots, the associated genes are

Table 1. Summary of the detected hotspots

Hotspot index	Size[a]	Loci[b]	GO category[c]	Hits[d]	t-test (all)[e]	t-test (hits)[f]
NO. 1	32	Chr XII:1056103	Telomere maintenance via recombination***	5	20	5
NO. 2	27	Chr IV:1525327	Telomere maintenance via recombination***	4	5	3
NO. 3	26	Chr XII:662627	Sterol metabolic process***	7	25	7
NO. 4	24	Chr I:52859	Fatty acid metabolic process***	10	12	6
NO. 5	24	Chr XV:202370	Response to abiotic stimulus**	10	11	6
NO. 6	23	Chr III:201166	Response to pheromone**	7	16	6
NO. 7	21	Chr VII:402833	Protein folding**	8	4	3
NO. 8	19	Chr I:7298	Fatty acid beta-oxidation***	5	13	4
NO. 9	**18**	**Chr IV:33214**	**Response to toxin***	**5**	**0**	**0**
NO. 10	16	Chr II:562415	Cytokinesis***	8	15	8
NO. 11	16	Chr X:698149	—	—	3	—
NO. 12	16	Chr XV:132423	—	—	13	—
NO. 13	**15**	**Chr XIII:843356**	**Organic acid transport***	**5**	**0**	**0**
NO. 14	15	Chr V:395442	—	—	3	—
NO. 15	15	Chr XVI:486637	Sexual reproduction*	6	8	4

[a]Number of genes associated with the hotspot. [b]The chromosome position of hotspot. [c]The most significant GO category enriched in the associated gene set. The enrichment test was performed using DAVID (Huang da and Lempicki, 2008). The gene function is defined by GO Fat category. DAVID outputs the Benjamini–Hochberg adjusted P-value. Adjusted P-values are indicated by *, where *$10^{-2} \sim 10^{-3}$; **$10^{-3} \sim 10^{-5}$; ***$10^{-5} \sim 10^{-10}$. [d]Number of associated genes that are functional in the enriched GO category. [e]Number of associated genes that can also be identified using t-test. [f]Number of associated genes that are functional in the enriched GO category and can also be identified using t-test. Two novel hotspots (NO. 9 and NO. 13) which cannot be detected by standard linear regression are in bold.

enriched in GO categories. In detail, for hotspot NO. 9, five of the 18 associated genes are functional in response to toxin, they are *AAD4*, *YDL218W, YLL056C*, *AAD6* and *SPS100*. The hotspot eQTL is cis-linked to one of them, AAD4, which apparently explains the detected association. Hotspot NO. 13 locates at transcription factor (TF) *CAT8*. Based on the transcriptional regulation information in yeast from both direct (Chip-chip) or indirect (Microarrays—wild type versus TF mutant) evidence (Teixeira *et al*., 2006), 9 (they are *ADY2*, *PUT4*, *GAP1*, *ATO3*, *ALP1*, *YDR222w*, *CWP1*, *ADH2* and *LPX1*) of the 15 associated genes can be potentially regulated by *CAT8* (adjusted $P < 10^{-10}$, the details of the p-value calculation is given in the Supplementary Document). Interestingly, five genes (i.e. *ADY2*, *PUT4*, *GAP1*, *ATO3*, *ALP1*) are functional in organic acid transport and *CAT8* is known to regulate acid transport pathway (Young *et al*., 2003). Identification of this hotspot provides a positive example and indicates that, when non-genetic effect has been successfully accounted for, we may be able to detect more biologically relevant *trans* eQTL.

7 CONCLUSIONS

In this article, we have introduced a method named 'LORS' to account for non-genetic effects in eQTL mapping. LORS provided a unified framework in which all SNPs and all gene probes can be jointly analyzed. The formulation of LORS is a convex optimization problem and thus its global optimum can be achieved. We also developed an efficient algorithm to solve this problem and guaranteed its convergence. We demonstrated its performance using synthetic datasets and real datasets.

A limitation of LORS is that we do not provide a rigorous statistical significance test of the estimated coefficient matrix **B**. Here we simply rank associations based on the estimated sparse matrix abs(**B**) and estimate FDR.

Funding: This work was supported in part by NSF (DMS 1106738), NIH (R01 GM59507) and NSFC (10901042, 10971075 and 91130032).

Conflict of Interest: none declared.

REFERENCES

Alter,O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Brem,R.B. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Cheung,V. and Spielman,R. (2009) Genetics of human gene expression: mapping dna variants that influence gene expression. *Nat. Rev. Genet*., **10**, 595–604.

Cookson,W. *et al*. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet*., **10**, 184–194.

Huang da,W. *et al*. (2008) Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc*., **4**, 44–57.

Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc*., **96**, 1348–1360.

Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Series B Stat. Methodol*., **70**, 849–911.

Friedman,J. *et al*. (2007) Pathwise coordinate optimization. *Ann. Appl. Stat*., **1**, 302–332.

Friedman,J. *et al*. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw*., **33**, 1–22.

Fusi,N. *et al*. (2012) Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol*., **8**, e1002330.

Gagnon-Bartsch,J. and Speed,T. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.

Gibson,G. (2008) The environmental contribution to gene expression profiles. *Nat. Rev. Genet*., **9**, 575– 581.

Kang,H. *et al*. (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**, 1909–1925.

Kang,H. *et al*. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet*., **42**, 348–354.

Leek,J. and Storey,J. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*., **3**, e161.

Leek,J. *et al*. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet*., **11**, 733– 739.

Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Li,L. *et al*. (2012) eQTL. *Methods Mol. Biol*., **871**, 265–279.

Lippert,C. *et al*. (2011) Fast linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

Listgarten,J. *et al*. (2010) Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 16465–16470.

Mazumder,R. *et al*. (2010) Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res*., **11**, 2287–2322.

Nica,A. and Dermitzakis,E. (2008) Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet*., **17** (R2), R129–R134.

Nielsen,T. *et al*. (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*, **359**, 1301–1307.

Nowak,G. *et al*. (2011) A fused lasso latent feature model for analyzing multi-sample aCGH data. *Biostatistics*, **12**, 776–791.

Pastinen,T. *et al*. (2006) Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet*., **15** (**Suppl. 1**), R9.

Rasmussen,C. and Williams,C. (2006) *Gaussian Processes in Machine Learning*. The MIT Press, Cambridge, MA, USA.

Recht,B. *et al*. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev*., **52**, 471–501.

Shi,L. *et al*. (2006) The microarray quality control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol*., **24**, 1151–1161.

Smith,E.N. and Kruglyak,L. (2008) Gene-environment interaction in yeast gene expression. *PLoS Biol*., **6**, e83.

Stegle,O. *et al*. (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol*., **6**, e1000770.

Teixeira,M. *et al*. (2006) The yeastract database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res*., **34** (**Suppl. 1**), D446–D451.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B*., **58**, 267–288.

Tibshirani,R. and Wang,P. (2008) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, **9**, 18–29.

Young,E. *et al*. (2003) Multiple pathways are co-regulated by the protein kinase Snf1 and the TFs Adr1 and Cat8. *J. Biol. Chem*., **278**, 26146– 26158.