Noname manuscript No.
(will be inserted by the editor)

A regularized multivariate regression approach for eQTL analysis

Xianlong Wang* · Li Qin* · Hexin Zhang · Yuzheng Zhang · Li Hsu · Pei Wang

Received: date / Accepted: date

Abstract Expression quantitative trait loci (eQTLs) are genomic loci that regulate expression levels of mRNAs or proteins. Understanding these regulatory provides important clues to biological pathways that underlie diseases. In this paper, we propose a new statistical method, `GroupRemMap`, for identifying eQTLs. We model the relationship between gene expression and single nucleotide variants (SNVs) through multivariate linear regression models, in which gene expression levels are responses and SNV genotypes are predictors. To handle the high-dimensionality as well as to incorporate the intrinsic group

* Joint first authors

X. Wang
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, WA, USA

L. Qin
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, WA, USA

H. Zhang
Institute of Mathematics Sciences, Peking University, Beijing, China

Y. Zhang
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, WA, USA

L. Hsu
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, WA, USA

P. Wang
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, Seattle, WA, USA
Correspondence should be addressed to E-mail: pwang@fhcrc.org

structure of SNVs, we introduce a new regularization scheme to (1) control the overall sparsity of the model; (2) encourage the group selection of SNVs from the same gene; and (3) facilitate the detection of trans-hub-eQTLs. We apply the proposed method to the colorectal and breast cancer data sets from The Cancer Genome Atlas (TCGA), and identify several biologically interesting eQTLs. These findings may provide insight into biological processes associated with cancers and generate hypotheses for future studies.

Keywords GroupRemMap · remMap · eQTL Analysis · Regularization · Multivariate Linear Regression

1 Introduction

Understanding regulatory relationships between genetic variants and gene expression is important for deciphering biological mechanisms underlying a wide range of human diseases. The goal of expression quantitative trait loci (eQTLs) analysis is to identify not only cis-eQTLs when SNVs (single nucleotide variants) regulate the expression of their own genes, but also trans-eQTLs when SNVs regulate the expression of genes to which the SNVs do not belong. Despite the promising progress made in recent human eQTL studies (Morley et al., 2004), large scale identification of cis- and trans- eQTLs is still daunting. The challenges stem from the high dimensionality of the data and complex multiple-to-multiple relationships between SNVs and gene expressions. In addition to these challenges, eQTL signals are generally weak. It is therefore imperative that statistical methods for detecting eQTLs should utilize the data most efficiently.

A natural way to characterize the regularization network between a set of SNVs and a set of expressions is through multivariate regression models in which the expression levels are responses and the SNVs are predictors. However, both the number of responses and the number of predictors can be larger than the sample size. Moreover, the predictors are often highly correlated due to natural grouping structures (e.g., genes or linkage disequilibrium blocks) for SNVs. These challenges complicate the already difficult problem of model selection and parameter estimation in high dimensional data.

To tackle the challenge of high-dimension-low-sample-size in multivariate models, various regularization methods have been proposed, assuming model sparsity in the sense that only a few predictors are associated with outcomes. This sparsity assumption is believed to hold in many situations such as genetic regulatory networks and genome-wide associations with complex diseases. In Turlach et al.(2005), an L_∞ based penalty was employed to select a common subset of predictors for all outcomes. Lutz and Bühlmann (2006) introduced the L_2 row boosting method to generate a sparse predictive model. In addition, Yuan et al.(2007) proposed to impose the L_2 norm constraint on the loading matrix for multivariate linear factor regression models to reduce the dimensionality of the predictor space. Obozinski et al.(2008) proposed an L_1/L_2 penalty to identify the union support set. More recently, Peng et al.(2010)

proposed a remMap penalty to deal with the high dimensionality of both predictors and outcomes. remMap combines the L_1 norm of the whole coefficient matrix and the L_2 norm of the coefficient vectors corresponding to the same predictor. The L_1 penalty controls the overall sparsity of the coefficient matrix such that only a subset of predictors are selected, and each selected predictor only influences some but not all responses. The L_2 penalty induces group sparsity on coefficients for the same predictor and further limits the total number of predictors entering the model. Moreover, the L_2 penalty encourages the selection of master predictors through borrowing information across different regressions. Another recent work for penalized multivariate regression is Rothman et al. (2010), in which the authors modeled the residuels from different regressions using a joint Gaussian distribution to account for the correlation among the response variables. Other related work for jointly modeling gene expressions and genetic variants include Yin and Li (2011) and Li et al. (2012), both focusing on studying the conditional independent relationships among gene expressions adjusting for possible genetic effects.

However, none of these methods takes consideration of potential structures among predictors. SNVs from the same transcript, gene or linkage disequilibrium block are often correlated. Analyzing SNVs within the same group jointly as a unit could potentially increase power by aggregating signals, and enhance the interpretability (Neale and Sham 2004; Chen et al. 2010; Liu et al. 2010; Li et al. 2011). The latter aspect is attractive when assessing the association of SNVs with gene expression in the eQTL analysis, because the functional unit of gene expression is at transcript or gene level, which usually consists of 10s to 100s of SNVs. We, therefore, propose a new regularization penalty, called GroupRemMap, which encourages group selection of SNVs from the same group while controlling the overall sparsity of the model and facilitating the detection of trans-hub-eQTLs.

The paper is organized as follows. The GroupRemMap model and its implementation are described in Section 2. Section 3 and 4 demonstrate the performance of the new method with simulation studies and real data applications, respectively. The paper ends with a brief summary in Section 5.

2 The GroupRemMap Method

2.1 Model

Consider an eQTL study of n subjects where each subject has p SNVs and q gene expressions. Assume a multivariate linear regression model for the effects of the p SNVs on the q gene expressions:

$$(y_{j1}, y_{j2}, \cdots, y_{jq}) = (x_{j1}, x_{j2}, \cdots, x_{jp})\, B_{p\times q} + \epsilon_j, \quad 1 \le j \le n \qquad (1)$$

where $(y_{j1}, y_{j2}, \cdots, y_{jq})$ and $(x_{j1}, x_{j2}, \cdots, x_{jp})$ denote the q expressions and p SNVs for the jth subject, respectively, $B = (b_{ij})_{p\times q}$ is the coefficient matrix, and $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ are $i.i.d$ error vectors with mean 0. For simplicity, we use

$Y_{n\times q} := (Y_1, Y_2, \cdots, Y_q)$ to denote the gene expression data, and $X_{n\times p} := (X_1, X_2, \cdots, X_p)$ to denote the SNV data, where Y_i, $i = 1, \ldots, q$, is an $n \times 1$ vector of the ith gene expression levels, and X_i, $i = 1, \ldots, p$, is an $n \times 1$ vector of the ith SNV genotype. Throughout the paper, we assume that gene expression and SNVs have been centered with sample mean equal to 0.

We assume prior knowledge is available to group SNVs into J distinct groups and denote these groups by $A_1, A_2, \cdots, A_J \subseteq \{1, 2, 3, \cdots, p\}$. We note that model (1) can easily accommodate other covariates such as patient or tumor characteristics (e.g., age, sex and tumor stage) by adding the covariates as predictors on the RHS of model (1). For simplicity, we will not include this in the following presentation of the method.

Our goal is to develop a regularized method to incorporate the group strucuturs on the SNVs in selecting groups and important SNVs within the identified groups. Towards this goal, we propose to minimize the following objective function:

$$L(B; \lambda_1, \lambda_2) = \frac{1}{2}\sum_{l=1}^{q} \left|\left|Y_l - XB^l\right|\right|_2^2 + \lambda_1 \sum_{i=1}^{p} ||C_i \cdot B_i||_1 + \lambda_2 \sum_{j=1}^{J} w_j \left(\sum_{k \in A_j} ||C_k \cdot B_k||_2 \right)^{\gamma} \quad (2)$$

where $||\cdot||_1$ and $||\cdot||_2$ denote the L_1 and L_2 norms for vectors, respectively, B^l is the lth column of B, B_i is the ith row of B, and C_i is the ith row of $C = (c_{ij})_{p\times q}$, where c_{ij} is a 0-1 valued indicator for whether the corresponding coefficient b_{ij} should be penalized. For example, if we know in advance that the ith SNV has a significant effect on the jth gene expression, we can set $c_{ij} = 0$ and b_{ij} will not be penalized; otherwise, we let $c_{ij} = 1$. The values of tuning paramters $\lambda_1, \lambda_2 \geq 0$ control the model dimension. The weight w_j is a constant, which incorporates the dimensionality of group A_j. A simple choice is $w_j \propto |A_j|^{1-\gamma}$, where $|A_j|$ is the total number of SNVs in group A_j and $\gamma > 0$ is the bridge penalty (Frank and Friedman, 1993; Huang et al., 2009).

In the objective function (2), the second term is a Lasso penalty on the whole coefficient matrix with the turning parameter λ_1 to control the overall sparsity of the coefficient matrix B. This is similar to remMap proposed by Peng et al. (2010). In the third term, we impose a weighted bridge type of penalty on each group to incorporate the group structure on the SNVs. When $\gamma \in (0, 1)$, this term encourages a group selection effect (see section 2.2 for details). Within the same group, we use L_2 norm on the row vectors $C_k \cdot B_k$ to induce the row sparsity of B, i.e. some rows are penalized to be entirely zero, such that SNVs that have effects on a majority of gene expressions are more likely to enter the final model. This facilitates identification of master predictors, which are often of great interest in genetic regulatory network studies.

We refer to the combination of the L_1 penalty and the bridge penalty on L_2 norm of grouped predictors as the `GroupRemMap` penalty, and call the coefficient

estimator based on the GroupRemMap penalty the GroupRemMap estimator:

$$\hat{B}(\lambda_1, \lambda_2) = \text{argmin} L(B; \lambda_1, \lambda_2).$$

In the following section, we will introduce an iterative algorithm to solve the above minimization problem. We will also show that the new GroupRemMap penalty can conduct variable selection at both the group and predictor levels simultaneously if we let $0 < \gamma < 1$. Specifically, in the simulation and real applications, we set $\gamma = 1/2$, while selecting optimal values for tuning parameters λ_1 and λ_2 through cross validation (see Section 2.3 for details).

Both the remMap penalty in Peng et.al (2010) and the group bridge penalty in Huang et.al (2009) are special cases of the GroupRemMap penalty. Specifically, when $\gamma = 1$ and $w_j = 1$, the GroupRemMap penalty simplifies to the original remMap penalty. When $q = 1$ and $\lambda_1 = 0$ (i.e. univariate outcomes), the penalty function becomes the group bridge penalty.

2.2 Estimation

In this section, we introduce an iterative algorithm to obtain the GroupRemMap estimator $\hat{B}(\lambda_1, \lambda_2)$. Define an alternative objective function as follows,

$$S(B, \theta; \lambda_1, \tau) = \frac{1}{2} \sum_{l=1}^{q} \left|\left|Y_l - XB^l\right|\right|_2^2 + \lambda_1 \sum_{i=1}^{p} ||C_i \cdot B_i||_1 + \sum_{j=1}^{J} \theta_j^{1-\frac{1}{\gamma}} w_j^{\frac{1}{\gamma}} \left(\sum_{k \in A_j} ||C_k \cdot B_k||_2 \right) + \tau \sum_{j=1}^{J} \theta_j \qquad (3)$$

where $\theta = (\theta_1, \theta_2, \cdots, \theta_J)$ are nuisance parameters, and $\tau > 0$ is the tuning parameter on θ.

Similar to Proposition 1 in Huang et al. (2009), we have the following result:

Proposition 1. For $0 < \gamma < 1$, if we let $\lambda_2 = \tau^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}$, then

$$\hat{B}(\lambda_1, \lambda_2) \text{ minimizes } L(B; \lambda_1, \lambda_2)$$

$$\iff \left(\hat{B}(\lambda_1, \lambda_2), \hat{\theta}\right) \text{ minimizes } S(B, \theta; \lambda_1, \tau) \text{ subject to } \theta_j \geq 0, j = 1, \ldots, J.$$

The proof is elementary, and we brief the idea as follows. Let $\hat{\theta} = \text{argmin} S(B, \theta; \lambda$ then it is easy to show that $S(B, \hat{\theta}; \lambda_1, \tau) = L(B; \lambda 1, \lambda 2)$. The proposition follows immediately.

The above proposition motivates us to estimate B by minimizing $S(B, \theta; \lambda_1, \tau$ instead of directly minimizing $L(B; \lambda_1, \lambda_2)$. If θ is fixed, the penalties in (3) can be treated as a weighted version of remMap penalty, where predictors within the same group A_j share the same L_2 norm tuning parameter $\theta_j^{1-\frac{1}{\gamma}} w_j^{\frac{1}{\gamma}}$. For $0 < \gamma < 1$, a small θ_j leads to a large value of $\theta_j^{1-\frac{1}{\gamma}}$, which tends to shrink the

coefficients in group A_j entirely to 0, and induces the group selection. Within the same group, the combination of the L_1 norms and the weighted L_2 norms can induce sparse selection of individual predictors.

At iteration $s \geq 1$, given the previous estimate $B^{(s-1)}$ and fixed parameters λ_1 and τ, we propose the following two-step iterative algorithm to estimate the coefficients:

– **Step 1** Update θ by solving

$$\partial S(B^{(s-1)}, \theta; \lambda_1, \tau)/\partial\theta_j = 0,\ j = 1, 2, \cdots, J.$$

A simple calculation yields

$$\theta_j^{(s)} = w_j \left(\frac{1-\gamma}{\tau\gamma}\right)^{\gamma} \left(\sum_{k \in A_j} \left|\left|C_k \cdot B_k^{(s-1)}\right|\right|_2\right)^{\gamma}.$$

– **Step 2** Given current $\theta^{(s)}$, update B by solving

$$B^{(s)} = \arg\min \frac{1}{2}\sum_{l=1}^{q} \left|\left|Y_l - XB^l\right|\right|_2^2 + \lambda_1 \sum_{i=1}^{p} ||C_i \cdot B_i||_1 + \sum_{j=1}^{J} \left(\theta_j^{(s)}\right)^{1-\frac{1}{\gamma}} w_j^{\frac{1}{\gamma}} \left(\sum_{k \in A_j} ||C_k \cdot B_k||_2\right).$$

Repeat Step 1 and 2 until convergence.

For the minimization problem in Step 2, we adopt the strategy in Peng et.al (2010) to iteratively update each row of B until convergence. The detailed calculation for updating each row of B with all the other rows fixed is summarized below.

Proposition 2. For $k \in A_j$, when $\{B_i\}_{i \neq k}$ in Step 2 are fixed, the kth row $B_k = (b_{k1}, b_{k2}, \cdots, b_{kq})$ can be estimated by:

$$\hat{b}_{k,l} = \begin{cases} X_k^T \widetilde{Y}_l / \left|\left|X_k\right|\right|_2^2, & \text{if } c_{kl} = 0; \\ 0, & \text{if } c_{kl} = 1 \text{ and } ||\hat{B}_k^{lasso}||_{2,C} = 0; \\ \left(1 - \frac{\left(\theta_j^{(s)}\right)^{1-\frac{1}{\gamma}} w_j^{\frac{1}{\gamma}}}{||X_k||_2^2 \cdot ||\hat{B}_k^{lasso}||_{2,C}}\right)_+ \cdot \hat{b}_{kl}^{lasso}, & \text{if } c_{kl} = 1 \text{ and } ||\hat{B}_k^{lasso}||_{2,C} \neq 0, \end{cases}$$

for $l = 1, 2, \ldots, q$, where

$$\widetilde{Y}_l = Y_l - \sum_{i \neq k} X_i b_{il},$$

$$||\hat{B}_k^{lasso}||_{2,C} = \left(\sum_{l=1}^{q} c_{kl} \left(\hat{b}_{kl}^{lasso}\right)^2\right)^{1/2},$$

and

$$\hat{b}_{kl}^{lasso} = \begin{cases} X_k^T \widetilde{Y}_l / \left|\left|X_k\right|\right|_2^2, & \text{if } c_{kl} = 0; \\ \left(\left|X_k^T \widetilde{Y}_l\right| - \lambda_1\right)_+ \cdot \frac{sign(X_k^T \widetilde{Y}_l)}{||X_k||_2^2}, & \text{if } c_{kl} = 1. \end{cases}$$

2.3 Selection of λ_1 and τ

K-fold cross validation is a commonly used approach for selecting tuning parameters. The procedure can be summarized as the following steps:

1. Randomly partition the whole data (Y, X) into K non-overlapping subsets $(Y^{(i)}, X^{(i)})$ with approximately equal sample sizes. Let $D^{(i)} = (Y^{(i)}, X^{(i)})$ be the validation set, and $D^{(-i)} = (Y^{(-i)}, X^{(-i)})$ be the complementary set of $D^{(i)}$ representing the training set.
2. Given a pair of (λ_1, τ), obtain the estimator of B based on the training set $D^{(-i)}$ by using the two-step iterative algorithm in Section 2.2. Denote the estimator as $\hat{B}^{(i)}(\lambda_1, \tau)$.
3. Define the cross validation score for (λ_1, τ) based on the validation set $D^{(i)}$ for each outcome:

$$\text{CV.rss}^l(\lambda_1, \tau) := \sum_{i=1}^{K} \left|\left| Y_l^{(i)} - X^{(i)} \hat{B}^{(i),l}(\lambda_1, \tau) \right|\right|_2^2, \; l = 1, 2, \cdots, q$$

   where $\hat{B}^{(i),l}(\lambda_1, \tau)$ is the lth column of $\hat{B}^{(i)}(\lambda_1, \tau)$. Thus the residual sum of square across all outcomes can be defined as:

$$\text{CV.rss}(\lambda_1, \tau) := \sum_{l=1}^{q} \text{CV.rss}^l(\lambda_1, \tau).$$

4. Select the pair of (λ_1, τ) with the smallest CV.rss as the final tuning parameters.

To avoid overfitting, sometimes it may be helpful to calculate cross validation score using re-estimated coefficients based on the selected model (Peng et. al. 2010). Then the last two steps in the above procedure can be modified as follows:

3. In the training set $D^{(-i)}$, for the lth outcome, calculate the ordinary least square estimators for the predictors with non-zero estimators in the lth column of $\hat{B}^{(i)}(\lambda_1, \tau)$. Here, we need to assume that the least square estimators are well defined. Generally this is the case when the true model is sparse. Denote the new least square estimator by $\hat{B}^{(i)}_{OLS}(\lambda_1, \tau)$. The cross validation score for (λ_1, τ) can then be calculated as:

$$\text{CV.ols}(\lambda_1, \tau) := \sum_{l=1}^{q} \text{CV.ols}^l(\lambda_1, \tau),$$

   where

$$\text{CV.ols}^l(\lambda_1, \tau) := \sum_{i=1}^{K} \left|\left| Y_l^{(i)} - X^{(i)} \hat{B}^{(i),l}_{OLS}(\lambda_1, \tau) \right|\right|_2^2, \; l = 1, 2, \cdots, q.$$

4. Select the pair of (λ_1, τ) with the smallest CV.ols as the final tuning parameters.

In practice, we recommend to compute both CV.rss and CV.ols scores, and use the one that gives smaller minimum cross validation score to determine the tuning parameter. Based on our experience, when the signal in the data is moderate, the minimum CV.ols score is usually lower than CV.rss, and tuning parameters selected by CV.ols often produce models with lower false positive rate than CV.rss. However, when the signal is extremely weak and power is of main concern compared to false positive rate, the minimum CV.rss score is usually lower than that of CV.ols.

3 Simulation studies

In this section, we conduct simulation studies to evaluate the performance of `GroupRemMap`, and compare it with `remMap` and the univariate `group bridge` methods. We assess the performance of these methods based on two criteria: (1) group selection and individual predictor selection: average false positive number (FP) and false negative number (FN); (2) prediction error: average K-fold cross validation based prediction error under the identified model.

We consider four different simulation settings. For each setting, a total of 200 independent data sets are generated. We calculate both CV.ols and CV.rss for all methods. CV.ols consistently gives lower cross validation scores and selects models with less false positive counts than CV.rss under all scenarios (data not shown). Hence, we report the results corresponding to models selected by CV.ols in this section.

3.1 Simulation Setting I

We consider two cases: equal group sizes (G1) and varying group sizes (G2). For both cases $J = 60$ groups, $p = 300$ and the sample size $n = 100$. Specifically, we generate the data as follows:

S1. Simulate J latent random variables: Z_1, Z_2,$\cdots$,Z_J, from a multivariate normal distribution $N(0, \Sigma_Z)$, where $\Sigma_Z = \left[0.5^{|i-j|}\right]_{J\times J}$.

S2. Based on Z_1, Z_2,$\cdots$,Z_J, define J categorical variables for G1 and G2.
For G1:

$$X_i = \begin{cases} 1 & \text{if } Z_i < \Phi^{-1}(1/6), \\ 2 & \text{if } \Phi^{-1}(1/6) \le Z_i < \Phi^{-1}(1/3), \\ 3 & \text{if } \Phi^{-1}(1/3) \le Z_i < \Phi^{-1}(1/2), \\ 4 & \text{if } \Phi^{-1}(1/2) \le Z_i < \Phi^{-1}(2/3), \\ 5 & \text{if } \Phi^{-1}(2/3) \le Z_i < \Phi^{-1}(5/6), \\ 6 & \text{if } Z_i \ge \Phi^{-1}(5/6), \end{cases}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$.
For G2:

If i is odd, i.e. $1, 3, 5, \cdots, 59$, let

$$X_i = \begin{cases} 1 & \text{if } Z_i < \Phi^{-1}(1/4), \\ 2 & \text{if } \Phi^{-1}(1/4) \le Z_i < \Phi^{-1}(1/2), \\ 3 & \text{if } \Phi^{-1}(1/2) \le Z_i < \Phi^{-1}(3/4), \\ 4 & \text{if } Z_i \ge \Phi^{-1}(3/4). \end{cases}$$

If i is even, i.e., $2, 4, 6, \cdots, 60$, let

$$X_i = \begin{cases} 1 & \text{if } Z_i < \Phi^{-1}(1/8) \\ 2 & \text{if } \Phi^{-1}(1/8) \le Z_i < \Phi^{-1}(1/4) \\ 3 & \text{if } \Phi^{-1}(1/4) \le Z_i < \Phi^{-1}(3/8) \\ 4 & \text{if } \Phi^{-1}(3/8) \le Z_i < \Phi^{-1}(1/2) \\ 5 & \text{if } \Phi^{-1}(1/2) \le Z_i < \Phi^{-1}(5/8) \\ 6 & \text{if } \Phi^{-1}(5/8) \le Z_i < \Phi^{-1}(3/4) \\ 7 & \text{if } \Phi^{-1}(3/4) \le Z_i < \Phi^{-1}(7/8) \\ 8 & \text{if } Z_i \ge \Phi^{-1}(7/8) \end{cases}$$

S3. Based on $X_i (1 \le i \le J)$, define the grouped predictors:

$$X_{i.j} := I\left(X_i = j\right)$$

For G1, $j = 1, 2, 3, 4, 5$.
For G2, if i is odd, $j = 1, 2, 3$; otherwise, $j = 1, 2, 3, \cdots, 7$

For both G1 and G2, we generate the outcomes from:

$$\begin{aligned} Y_1 =& -1.77X_{1.1} - 1.87X_{1.2} - 2.07X_{2.3} + 2.41X_{3.1} - 1.70X_{3.2} + \epsilon_1 \\ Y_2 =& -2.40X_{1.1} + 2.44X_{1.2} - 2.16X_{1.3} - 2.13X_{3.1} + 1.56X_{3.3} + \epsilon_2 \\ Y_3 =& \ 1.71X_{1.1} + 1.68X_{1.2} - 2.19X_{1.3} + 1.88X_{2.1} - 2.27X_{3.2} + \epsilon_3 \\ Y_4 =& \ 2.00X_{1.1} + 2.22X_{1.3} + 2.49X_{2.1} + 1.88X_{2.2} + 2.28X_{3.3} + \epsilon_4 \\ Y_5 =& \ 2.43X_{1.1} - 1.71X_{2.1} - 2.15X_{2.2} + 1.63X_{2.3} + 1.77X_{3.3} + \epsilon_5 \end{aligned} \tag{4}$$

where $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5) \sim N(0, \Sigma_\epsilon)$ and $\Sigma_\epsilon = \left[0.5^{|i-j|}\right]_{5\times 5}$. All above non-zero coefficients are generated from $U(1.5, 2.5) \cup U(-2.5, -1.5)$. The results are shown in Table 1.

Under both G1 and G2 scenarios, `GroupRemMap` has smaller FP, FN and standard error than `remMap` and `group bridge` in both group selection and individual predictor selection. In addition, whether the group size is constant or not does not appear to affect the performance of `GroupRemMap`.

3.2 Simulation Setting II

We generate the data in the same way as in Simulation Setting I except that the noise level is higher in model (4). Specifically, we generate noise from $N(0, 4\Sigma_\epsilon)$, four times as large variance as in Setting I. For simplicity, we only consider the case of equal group sizes. The results are presented in Table 2.

Table 1 Summary of mean of FP (SE) and mean of FN (SE) over 200 data sets under setting I (n=100, J=60, p=300) when the noise level is moderate.

		Group.S (14) *		Indiv.S (25) *	
	Method	FP	FN	FP	FN
G1	G.remmap	4.97(3.8)	0.03(0.16)	7.28(4.35)	0.46(0.76)
	Remmap	8.38(6.02)	0.02(0.12)	13.05(5.98)	0.32(0.77)
	G.bridge	13.22(5.08)	2.19(1.43)	21.13(8.1)	3.54(2.43)
G2	G.remmap	4.94(3.4)	0.07(0.27)	7.94(3.92)	0.82(1.01)
	Remmap	11.38(6.45)	0.15(0.45)	16.37(7.95)	1.02(1.42)
	G.bridge	9.17(4.13)	1.18(0.99)	14.7(5.88)	2.61(1.85)

* Group.S(14): *Group Selection* and the number of true groups for all outcomes is 14. Indiv.S(25): *Individual Predictor Selecti*on and the number of true predictors for all outcomes is 25;
FP: *False Positive Number*; FN: *False Negative Number*.

Table 2 Summary of mean of FP (SE) and mean of FN (SE) over 200 data sets under setting II ($n = 100, J = 60, p = 300$) where the noise level is high.

		Group.S (14)*		Indiv.S (25)*	
	Method	FP	FN	FP	FN
G1	G.remmap	14.14(8.38)	1.46(2.09)	17.93(10.38)	4.98(3.49)
	Remmap	19.72(12.10)	0.68(1.29)	25.4(13.32)	4.06(3.36)
	G.bridge	14.02(5.92)	6.61(2.06)	20.42(9.85)	12.4(3.32)

* Group.S(14): *Group Selection* and the number of true groups for all outcomes is 14.Indiv.S(25): *Individual Predictor Selecti*on and the number of true predictors for all outcomes is 25;
FP: *False Positive Number*; FN: *False Negative Number*.

As expected, all three methods commit more FP and FN when the noise level increases compared to Simulation Setting I (see the top panel of Table 1 vs Table 2). However, `GroupRemMap` still gives more favorable results than `remMap` and `group bridge` methods. Specifically, compared to `remMap`, since `GroupRemMap` imposes an additional layer of regularization by using the group structure among predictors, it tends to have better control of FP than `remMap` with only slightly loss in detecting signals (less than 1 count). Thus, the overall performance of `GroupRemMap` is better than that of `remMap`. For `group bridge`, since it deals with each regression separately and ignores the dependence among different responses, it often has much higher FN than either `remMap` or `GroupRemMap`.

3.3 Simulation Setting III

We generate predictors using different numbers of groups $J = 30, 60, 100$ with equal group size of 5. We also consider a relatively larger linear model:

$$Y_1 = (X1.1, X1.2, X1.3, X2.1, X2.2, X4.1, X3.1, X3.2, X3.3, X8.1))^T B_1 + \epsilon_1$$
$$Y_2 = (X1.1, X1.2, X1.3, X2.2, X2.3, X4.2, X5.1, X5.2, X5.3, X10.1)^T B_2 + \epsilon_2$$
$$Y_3 = (X1.1, X1.2, X1.3, X2.3, X2.4, X4.3, X7.1, X7.2, X7.3, X12.1)^T B_3 + \epsilon_3 \ (5)$$
$$Y_4 = (X1.1, X1.2, X1.3, X2.4, X2.5, X4.4, X9.1, X9.2, X9.3, X13.1)^T B_4 + \epsilon_4$$
$$Y_5 = (X1.1, X1.2, X1.3, X2.1, X2.5, X4.5, X11.1, X11.2, X11.3, X14.1)^T B_5 + \epsilon_5$$

where $B_1, B_2, \cdots, B_5$ are non-zero coefficient vectors generated from $U(1.5, 2.5) \cup U(-2.5, -1.5)$, and $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5) \sim N(0, \left[0.5^{|i-j|}\right]_{5\times 5})$. The performance of the three methods for this simulation is shown in Table 3.

Table 3 Summary of mean of FP (SE) and mean of FN (SE) over 200 data sets under setting III($n = 100$).

		Group.S (25)*		Indiv.S (50)*	
	Method	FP	FN	FP	FN
$J = 30$	G.remmap	9.76(3.82)	2.48(1.17)	15.81(4.78)	4.84(2.18)
($p = 150$)	Remmap	21.4(5.54)	1.05(1.1)	32.54(7.19)	5.38(2.3)
	G.bridge	9.93(3.65)	3.52(1.6)	23.78(7.22)	4.79(2.49)
$J = 60$	G.remmap	15.29(5.85)	2.54(1.25)	21.86(7.12)	5.12(2.63)
($p = 300$)	Remmap	32.42(8.41)	1.19(1.06)	42.71(9.91)	6.06(2.82)
	G.bridge	14.53(4.92)	5.36(1.9)	26.15(8.36)	9.56(4.33)
$J = 100$	G.remmap	23.43(7.54)	2.63(1.35)	30.36(8.67)	5.18(3.13)
($p = 500$)	Remmap	43.17(12.7)	1.43(1.27)	53.44(15.1)	6.77(3.22)
	G.bridge	23.12(6.88)	6.61(2.23)	37.46(11.2)	12.73(5.16)

* Group.S(25): *Group Selection* and the number of true groups for all outcomes is 25. Indiv.S(50): *Individual Predictor Selection* and the number of true predictors for all outcomes is 50;
J: number of groups; FP: *False Positive*; FN: *False Negative*.

Again, `GroupRemMap` has better performance than `remMap` and `group bridge`. In addition, as the number of groups (and predictors) increases, the FP of all three methods increases. However, the FN of `GroupRemMap` and `remMap` appear to be less affected than `GroupBridge`. This suggests that jointly modeling through multiple regression helps enhance the power.

3.4 Simulation Setting IV

In this section, we generate data mimicking the setting of the colorectal cancer data set in Section 4.1. Specifically, we use the genotype data of 567 SNVs

from 202 colorectal tumor samples (see Section 4.1 for details) and generate the transcript levels of 67 genes based on a simulated eQTL network as shown in Figure 1. The 567 SNVs belong to $J = 26$ groups (genes), with mean size 21.8 and range from 1 to 101. There are a total of 121 eQTLs in the eQTL network, involving 46 SNVs and 36 transcripts. Eight out of 121 eQTLs are cis-regulation. In addition, there are 16 trans-hub (degree>5). The coefficients corresponding to eQTL edges are randomly generated from Uniform(1,4), and the mean noise-to-signal ratio is 1.216. Transcripts that don'thave eQTL edges are generated from N(0,1). A total of 200 independent data sets are generated. The results are presented in Table 4.

Similar to the previous results, `GroupRemMap` performs the best among three methods: it produces the lowest false positive and false negative errors in both individual predictor selection and group selection. `group bridge` has particularly high false positives in this example, suggesting that jointly modeling multiple transcripts is essential in the high dimensional eQTL analysis.

In practice, some predictors may be classified into wrong groups due to incomplete knowledge. To assess the impact of misclassification on the performance of `GroupRemMap`, we randomly assign the group labels of 10% of the SNVs from the groups that have eQTL regulations. The result is also presented in Table 4. Overall, the performance of `GroupRemMap` is robust against group label misclassficaition with a slight increase in FP and no obvious change in FN.

We also investigate the effect of different values of γ on the performance of `GroupRemMap`. The results for $\gamma = 0.25, \gamma = 0.50$ and $\gamma = 0.75$ are summarized in Table S-1 in the Supplementary Materials. While results of different γ are rather similar, $\gamma = 0.5$ gives the most favorable FP and FN. Thus, we choose to use $\gamma = 0.5$ in the real data analysis in Section 4.

Moreover, to evaluate the performance of our method under the setting where both the number of predictors and responses exceed the sample size, we perform another simulation using $q = 400$. Specifically, we added 333 new noise responses generated from independent $N(0, 1)$ to the original 67 transcripts. The corresponding B is of 567×400, with the last 333 columns being zeros. The performances of the three methods on this data set are summarized in Table S-2. The performance of `GroupRemMap` is still quite good, while the false positive of `remMap` and `group bridge` are at least doubled. This suggests that `GroupRemMap` is quite capable of handling cases with both p and q exceeding n.

4 Real data analysis

We apply `GroupRemMap` for eQTL analysis on two cancer data sets from the Cancer Genome Atlas (TCGA) consortium (`http://www.cancergenome.nih.`

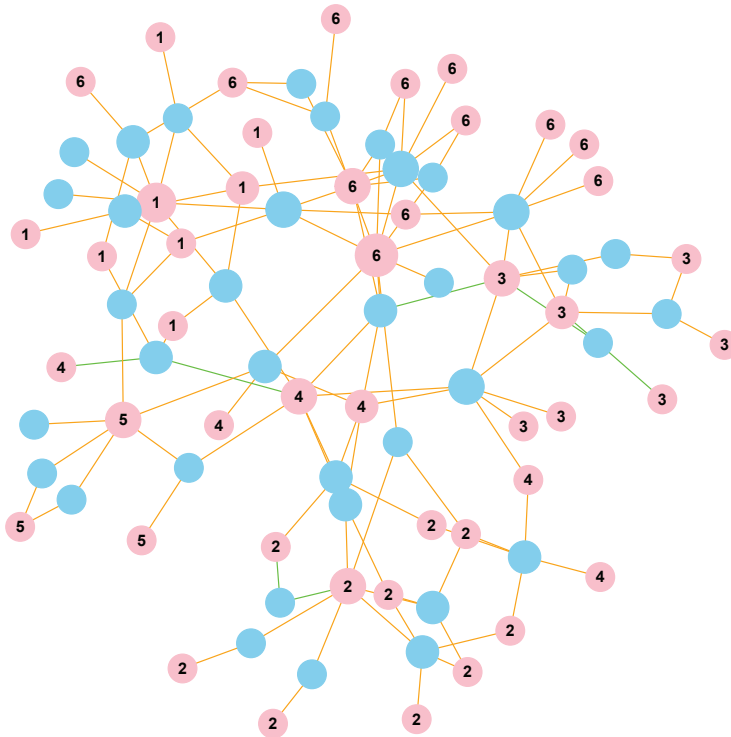

Fig. 1 Layout of the eQTL network used in Simulation IV. Each pink node represents a SNV and each blue node represents a transcript. Green and orange edges represent cis-eQTLs and trans-eQTLs respectively. Only those SNVs (46) and transcripts (36) with at least one eQTL edge are shown in the figure. The number of each SNV node represents the SNP-group label of that SNV.

gov/). The first data set consists of gene expression data and SNV genotype data for a group of colorectal tumor tissue samples; while the second data set is for breast tumor tissue samples.

Since genome-wide investigation of cis- and trans- eQTLs involves estimating thousands of millions of parameters, such analysis is extremely underpowered based on data sets with only a few hundred samples, as the ones considered in this section. To tackle this difficulty, we will make use of a priori knowledge about biological pathways and risk loci, and perform eQTL analysis for pre-selected gene sets from genome regions or pathways known to be relevant to the disease. Specifically, we will focus on 31 genome regions with known risk SNVs in the colorectal cancer study and the DNA repair pathway in the breast cancer study.

Table 4 Summary of mean of FP (SE) and mean of FN (SE) over 200 data sets under setting IV.

	Group.S (54)*		Indiv.S (121)*	
Method	FP	FN	FP	FN
G.remmap	25.90(0.36)	0.12(0.02)	140.44(0.92)	9.06(0.08)
G.remmap(mislab)*	30.25(0.39)	0.10(0.02)	146.16(0.88)	10.11(0.07)
Remmap	33.96(0.45)	0.13(0.02)	165.70(1.00)	9.16(0.07)
G.bridge	263.99(1.26)	2.80(0.03)	902.35(4.44)	12.52(0.08)

* Group.S(54): *Group Selection* and the number of true groups for all outcomes is 54. Indiv.S(121): *Individual Predictor Selection* and the number of true predictors for all outcomes is 121.
G.remmap(mislab): 10% of the group labels are mislabeled.
FP: *False Positive*; FN: *False Negative*. The numbers within the parentheses are estimated standard errors.

4.1 Colorectal cancer

As of September 13, 2012, in TCGA, tumor tissues from 224 colorectal cancer patients had been assayed on platform Agilent g4502a for their gene expression, and we downloaded the level 3 data[1] via Firehose from the Broad Institute's Genome Data Analysis Center (GDAC) website (`https://confluence.broadinstitute.org/display/GDAC/Home`). As of May 15, 2012, 584 had been assayed on platform Affymetrix® Genome-Wide Human SNP Array 6.0 for genotypes, and we used the level 2 data[2].

We focused on the 31 known colorectal cancer (CRC) susceptibility loci (Peters et al., 2013) and extracted SNVs/genes in their neighboring regions (defined as 2 genes upstream and 2 genes downstream). When pre-processing the genotype data, we sequentially removed SNVs with confidence score > 0.1; samples with missing rate $> 10\%$; SNVs with missing rate $> 10\%$; and SNVs with minor allele frequency $< 5\%$. For gene expression, we sequentially removed samples with probe missing rate $> 0.3\%$ and probes with at least one sample missing. This resulted in 567 SNVs and 67 transcripts on 202 samples. To incorporate both the dominance/recessive genetic models, we coded the genotype $X(= 0, 1, 2)$ with two variables: $(X_1, X_2) = (0, 1)$, if $X = 0$; $(X_1, X_2) = (1, 1)$, if $X = 1$; and $(X_1, X_2) = (1, 0)$, if $X = 2$. Each expression and each SNV were standardized to have mean 0 and standard deviation 1.

We applied `GroupRemMap` with the tuning parameters selected by 5-fold cross validation based on CV.rss, which gave lower error score than CV.ols on this data set. The resulting eQTL network is shown in Figure 2. There are 4 cis-edges (i.e., the transcript is in the neighboring region of the SNV)

[1] `gdac.broadinstitute.org_COADREAD.Merge_transcriptome__agilentg4502a_07_3__unc_edu__Level_3__unc_lowess_normalization_gene_level__data.Level_3.2012091300.0.0.tar.gz`

[2] `gdac.broadinstitute.org_COADREAD.Merge_snp__genome_wide_snp_6__broad_mit_edu__Level_2__birdseed_genotype__birdseed.Level_2.2012051500.0.0.tar.gz`

and 11 trans-edges. Here, the number of detected cis-edges is proportionally much stronger than that of detected trans-edges, because the number of potential trans-regulations is more than 50 folds larger than that of cis-regulations. This is consistent to the common belief that signals of cis-eQTLs are usually higher than that of trans-eQTLs, and thus the power to detect the former should be higher. All three SNVs are in the neighbor of gene DIP2B and cis-regulate the expression of DIP2B, whose protein participates in DNA methylation (GeneCards), suggesting the functional relevance of these SNVs. This demonstrates that our method, which accounts for group structure, to some degree encourages the finding of multiple variants in the same gene region. In addition, some trans-regulations are also intriguing. For example, both rs3825402 and rs11169561 regulate the expression of GLDC, whose methylation status is a potential epigenetic biomarker for CRC (Ali, 2010). Such information revealed by the eQTL network can help to shed lights on the biological mechanism underlying CRC.

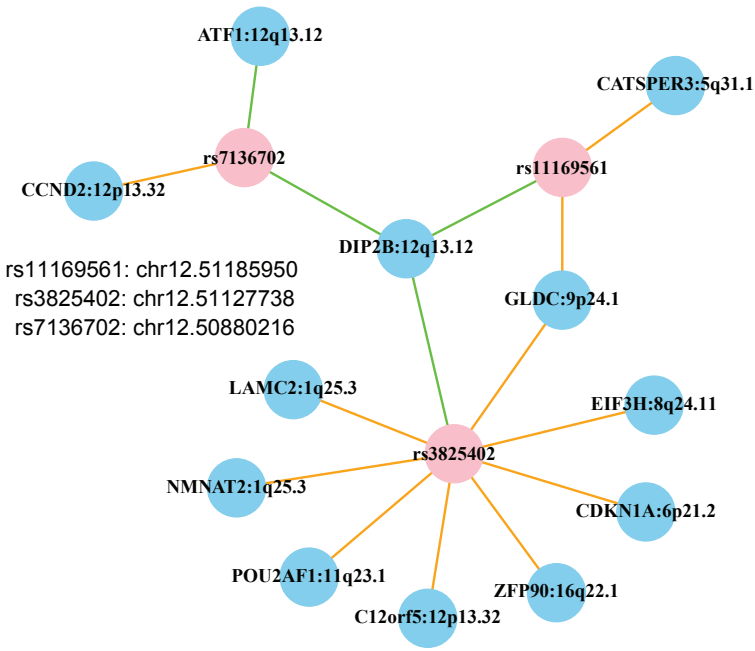

Fig. 2 eQTLs discovered through GroupRemMap from the colorectal cancer data set. Nodes: Pink-SNV (3); Blue-Exp(12). Edges: Green-Cis (4); Orange-Trans (11).

We also applied `remMap` and `Group bridge` to this data set. `remMap` identifies 4 cis-eQTLs and 31 trans-eQTLs (Figure S-1), and `group bridge` does not detect any eQTL. Although `remMap` identifies more eQTL edges than GroupRemMap, based on the observations from simulation studies, it is very

likely `remMap` yields more false eQTLs than `GroupRemMap`. Moreover, the result of `GroupRemMap` suggests a trans-hub eQTL rs3825402 while there is no clear hub-structure for `remMap`. Note that the cis-eQTL regulation between rs11169561 and *DIP2B*, and trans-eQTL regulation between rs3825402 and *GLDC*, are detected by both methods.

4.2 Breast cancer

As of September 13, 2012, in TCGA, tumor tissues from 777 breast cancer patients had been assayed on platform Illumina RNAseq for their gene expression, and as of August 25, 2012, 870 assayed on platform Genome-Wide Human SNP Array 6.0 for genotypes. The level 3 RNAseq data[3], and level 2 genotype data[4] were downloaded from Firehose website.

The pre-processing on the genotype data is the same as the procedure used for the colorectal cancer data. For the RNA-Seq data, we had the following pre-processing steps: 1) removed samples with missing values or 0 counts in more than 15% of the genes; 2) filtered out genes with missing rate $> 10\%$; 3) took log2 transformation on the read-count data; 5) normalized each sample to have median 0 and standard deviation 1; 6) removed the bottom 20% genes with small standard deviation. After the pre-processing, 614 samples had both the genotype and gene expression data. We then focused on 48 genes and 361 SNVs in the DNA damage repair pathway retrieved from the Gene Ontology database, because this pathway is essential in maintaining normal cell functions and proliferation and highly involved in breast cancer etiology (Gorgoulis et al., 2005; Bartkova et al., 2005; Di Micco et al., 2006; Bartkova et al., 2006). The number of SNVs each gene contains ranges from 1 to 63, with the average around 9.

We applied the three methods in the same way as we did in Section 4.1. Interestingly, both `GroupRemMap` and `remMap` select the same model corresponding to $\lambda_1 = 60$ and $\lambda_2 = 0$. This implies that the grouping/hub structure in the underlying eQTL network of this particular data set may be very weak such that the group penalty in `GroupRemMap` does not contribute. Figure 3 shows the discovered eQTL network, which consists of 92 SNVs, 43 transcripts, 16 cis-edges and 108 trans-edges. On the other hand, `group bridge` calls many more eQTLs than `GroupRemMap` and `remMap` (see Figure S-2 in the Supplementary Materials). However, based on our experience with simulated data, we expect a considerable number of them to be false positives.

The eQTL network in Figure 3 reveals interesting eQTLs for a set of genes implicated in tumorigenesis in breast cancer including *BRCA1*, *ATM* , *TP53* and *TERT* (Couch et al., 2013; Muraki et al., 2013; Beillerot et al., 2012). For example gene *TERT* is regulated by three SNVs (rs3136189, rs12602273

[3] `gdac.broadinstitute.org_BRCA.Merge_rnaseq__illuminahiseq_rnaseq__unc_edu_Level_3__gene_expression__data.Level_3.2012091300.0.0.tar.gz`

[4] `gdac.broadinstitute.org_BRCA.Merge_snp__genome_wide_snp_6__broad_mit_edu_Level_2__birdseed_genotype__birdseed.Level_2.2012082500.0.0.tar.gz`

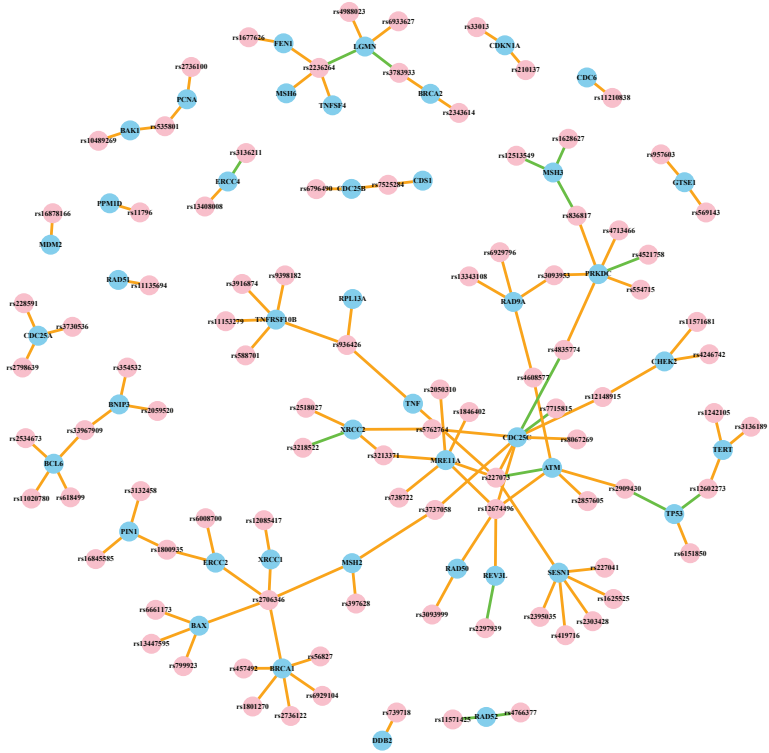

Fig. 3 eQTLs discovered through GroupRemMap and remMap from the breast cancer data set. Nodes: Pink-SNV (92); Blue-Exp(43). Edges: Green-Cis (16); Orange-Trans (108).

and rs1242105) from three different genes. Notably, variant rs12602273 is contained in gene *TP53* which is a well known tumor suppressor gene. Gene *TP53* encodes a protein that regulates the expression of genes in a wide ranges of biological functions. Mutations in *TP53* are associated with a variety of human cancers (`http://www.ncbi.nlm.nih.gov/gene/7157`). In addition, gene *TERT* is involved in the reverse transcriptase activity of telomerase. Telomerase expression plays a crucial role in cellular senescence, as it is normally repressed in postnatal somatic cells resulting in progressive shortening of telomeres. Deregulation of telomerase expression in somatic cells may be involved in oncogenesis (`http://www.ncbi.nlm.nih.gov/gene/7015`). Haiman et al. (2011) identified a common variant at the TERT-CLPTM1L locus associated with estrogen receptor-negative breast cancer. Our result suggests that these regulatory elements may coordinate to mediate the transcriptional activity of *TERT*.

5 Summary

In this paper, we proposed a novel statistical method, `GroupRemMap`, for eQTL analysis. `GroupRemMap` models the dependent relationship between gene expressions and genetic variants through multivariate linear regression, and regularizes the regression coefficients while accounting for the correlation among SNVs on the genome. By design, the new method is able to control the overall sparsity of the model, encourage the group selection of SNVs from the same gene, and facilitate detection of trans-hub-eQTLs. We applied the proposed method to TCGA data on colorectal and breast cancer, and were able to identify a few biological relevant eQTLs. The regulatory mechanism underlying these findings is worth further investigation, which could potentially enhance our understanding of the underlying biological processes of both cancers.

We implement the main algorithm of `GroupRemMap` in C programming language, and an R package will be available through CRAN. Regarding computation time, one run of `GroupRemMap` on one pair of parameter (λ_1, λ_2) takes $\sim$2.0 seconds on a Dell R710 computer with Intel® Xeon® X5680 3.3GHz processors and 128 GB RAM.

Acknowledgements This work is supported by NIH grants R01CA138215(XW, PW), SUB-CA160034(XW, YZ, PW), R01GM082802 (PW), P01CA53996 (LH, PW), R01AG014358 (LH), P50CA138293 (LH, PW), and U24CA086368 (PW).

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at "http://cancergenome.nih.gov". The TCGA colorectal and breast cancer data sets were accessed from the dbGaP website(http://www.ncbi.nlm.nih.gov/gap) through accession number phs000178_v3_p3.

The authors declare no conflict of interest.

References

1. Ali, Deeqa Ahmed Mohamed. Identification of novel epigenetic biomarkers in colorectal cancer, gldc and ppp1r14a. Thesis of masters degree, Department of Molecular Biosciences(University of Oslo), 2010.
2. Bartkova, J., Hořejší, Z., Koed, K., Krämer, A., Tort, F., Zieger, K., Guldberg, P., Sehested, M. et al.(2005). DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature*, **434**:864-70.
3. Bartkova, J., Rezaei, N., Liontos, M., Karakaidos, P., Kletsas, D., Issaeva, N., Vassiliou, L. V., Kolettas, E. et al., (2006). Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature*, **444**:633-7, Nov. 30.
4. Beillerot A, et al. (2012). Protection of CDC25 phosphatases against oxidative stress in breast cancer cells: evaluation of the implication of the thioredoxin system. *Free Radic Res*, 2012 May. PMID 22360685.
5. Bellam, N., Pasche, B., (2010). Tgf-beta signaling alterations and colon cancer. *Cancer Treat Res.* **155**:85-103.
6. Blank, M., Tang, Y., Yamashita, M., Burkett, S. S., Cheng, S. Y., Zhang, Y. E., (2012). A tumor suppressor function of Smurf2 associated with controlling chromatin landscape and genome stability through RNF20. *Nat Med.* Jan 8; **18(2)**:227-34.
7. Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U., Hsu, L., (2010). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am J Hum Genet.* **86(6)**:860-71.
8. Couch FJ, et al. (2013). Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet*, 2013. PMID 23544013.
9. Di Micco, R., Fumagalli, M., Cicalese, A., Piccinin, S., Gasparini, P., Luise, C., Schurra, C., Garre', M. et al., (2006). Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. *Nature*, **444**:638-42, Nov. 30.
10. Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with Discussion). *Technometrics*, **35**, 10948.
11. Gorgoulis, V. G., Vassiliou, L. V., Karakaidos, P., Zacharatos, P., Kotsinas, A., Liloglou, T., Venere, M., Ditullio, R. A. et al, (2005). Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature*, **434**:907-13.
12. Haiman, C. A., Chen, G. K., Vachon, C. M., Canzian, F., Dunning, A., Millikan, R. C., Wang, X., Ademuyiwa, F., Ahmed, S., Ambrosone, C. B. et al, (2011). A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet.* **43(12)**:1210-4. doi: 10.1038/ng.985.
13. Huang, J., Ma, S., Xie, H., Zhang, C., (2009). A group bridge approach for variable selection, *Biometrika*, **96(2)**:339-355.
14. Li, B., Chun, H. and Zhao, H., (2012). Sparse Estimation of Conditional Graphical Models with Application to Gene Networks. *J of Am Stat Assoc*, **107(497)**, 152-167.
15. Li, M. X., Gui, H. S., Kwan, J. S., Sham, P. C. (2011). GATES: a rapid and powerful gene-based association test using extended Simes procedure. The American Journal of Human Genetics, **88(3)**, 283-293.
16. Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U., Hsu, L. (2010). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. The American Journal of Human Genetics, **86(6)**, 860-871.
17. Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., ... & Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. The American Journal of Human Genetics, **87(1)**, 139-145.
18. Lutz, R. and Bühlmann, P., (2006). Boosting for high-multivariate responses in high-dimensional linear regression, *Statist. Sin.*, **16**, 471-494.
19. Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., Cheung, V. G., (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743747.
20. Muraki K, et al. (2013). The role of ATM in the deficiency in nonhomologous end-joining near telomeres in a human cancer cell line. *PLoS Genet*, 2013 Mar. PMID 23555296.

21. Neale, B.M., and Sham, P.C. (2004). The future of association studies: Gene-based analysis and replication. Am. J. Hum. Genet. 75, 353362.
22. Obozinski, G., Wainwright, M. J., Jordan, M. I., (2011). Union support recovery in high-dimensional multivariate regression, *Ann. Statist.*, **39(1)**, 1-47.
23. Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., Wang, P., (2010). Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer, *Ann.Appl.Stat*, **4(1)**, 53-77.
24. Peters, U., et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastoenterology, 144(4):799807, 2013.
25. Rothman, A., Levina, L., and Zhu, J., (2010). Sparse multivariate regression with covariance estimation. *J. Comp Graph Stat.* 19:947-962.
26. Slattery, M. L., Lundgreen, A., Herrick, J. S., Wolff, R. K., (2011). Genetic variation in RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, and PDK1 and risk of colon or rectal cancer. *Mutat Res.* **706(1-2)**:13-20.
27. Turlach, B., Venables, W., Wright, S., (2005). Simultaneous variable selection, *Technometrics*, **47**, 349-363.
28. Yin, J. and Li, H., (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data, *Ann. Appl. Stat.*, **5(4)**, 2630-2650.
29. Yuan, M., Ekici, A., Lu, Z., and Monteiro, R., (2007). Dimension reduction and coefficient estimation in multivariate linear regression, *J. R. Statist. Soc. B*, **69(3)**, 329-346.

Supplementary Materials

Table S-1 Simulation results under setting IV.

	Group.S (54)*		Indiv.S (121)*	
Method	FP	FN	FP	FN
G.remmap($\gamma = 0.25$)	26.90(0.36)	0.13(0.02)	157.62(1.09)	9.58(0.09)
G.remmap($\gamma = 0.50$)	25.90(0.36)	0.12(0.02)	140.44(0.92)	9.06(0.08)
G.remmap($\gamma = 0.75$)	39.02(0.72)	0.12(0.02)	154.91(1.21)	9.38(0.06)

Indiv.S(121): *Individual Predictor Selection* and the number of true predictors for all outcomes is 121.
Group.S(54): *Group Selection* and the number of true groups for all outcomes is 54.
FP: *False Positive*; FN: *False Negative*. The numbers within the parentheses are estimated standard errors.

Table S-2 Summary of mean of FP (SE) and mean of FN (SE) over 200 data sets with 400 responses ($> n = 202$).

	Group.S (54)*		Indiv.S (121)*	
Method	FP	FN	FP	FN
G.remmap	16.81(0.33)	0.38(0.03)	105.51(0.75)	11.39(0.07)
Remmap	92.77(0.73)	0.10(0.02)	252.03(1.27)	9.57(0.08)
G.bridge	1031.71(2.91)	2.93(0.03)	2322.17(7.04)	13.24(0.08)

* Group.S(54): *Group Selection* and the number of true groups for all outcomes is 54.
Indiv.S(121): *Individual Predictor Selection* and the number of true predictors for all outcomes is 121.
FP: *False Positive*; FN: *False Negative*. The numbers within the parentheses are estimated standard errors.

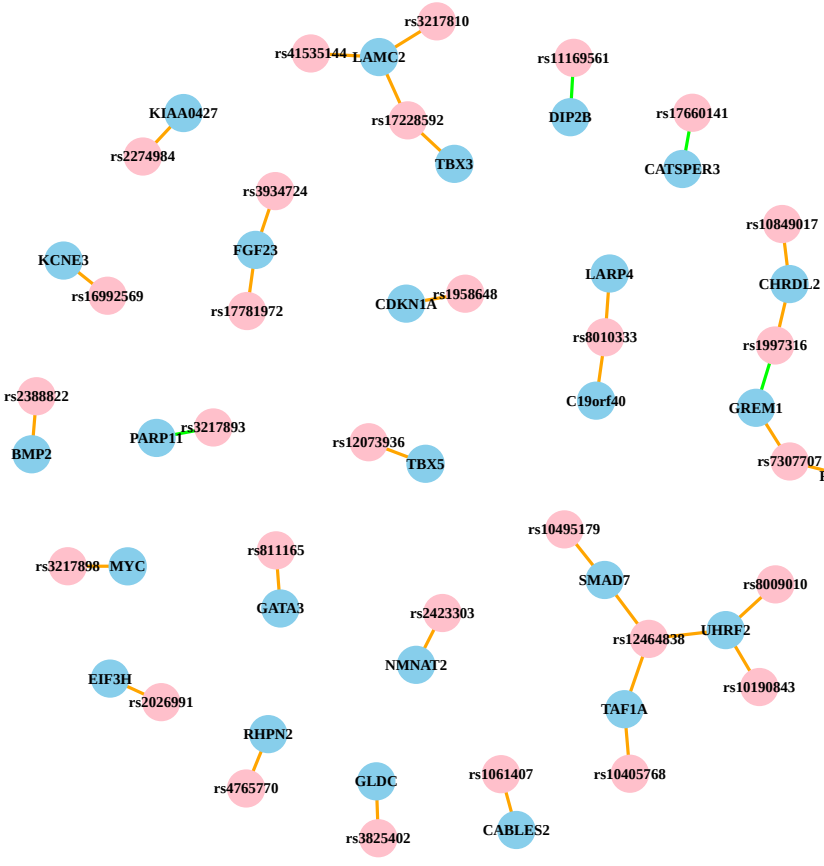

Fig. S-1 eQTLs discovered through `remMap` from the colorectal cancer data set. Nodes: Pink-SNV (29); Blue-Exp(26). Edges: Green-Cis (4); Orange-Trans (31).

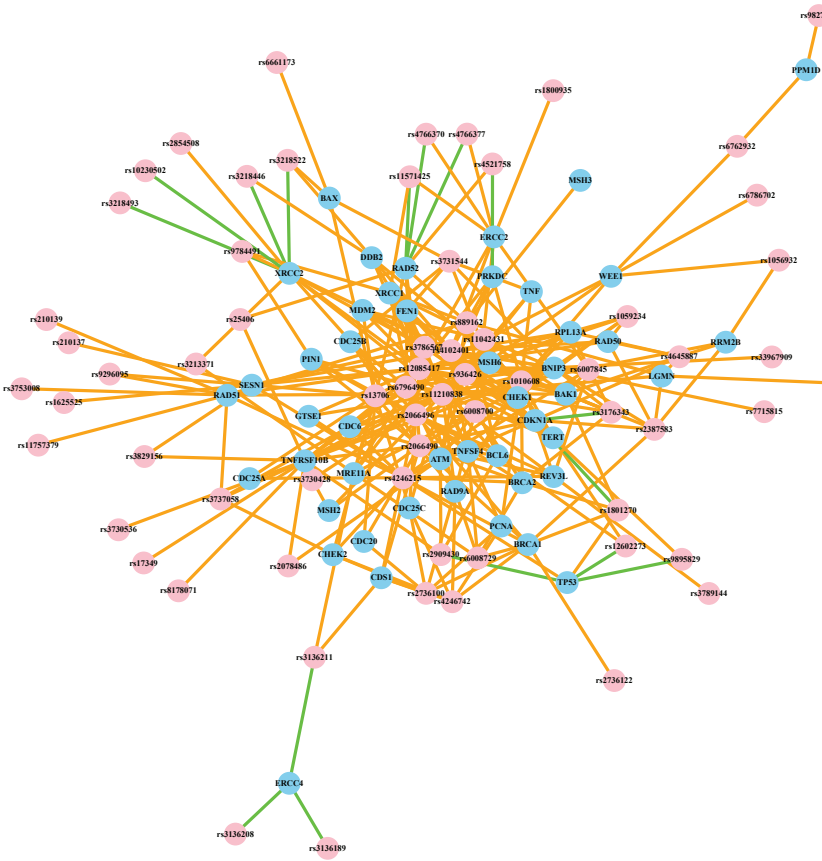

Fig. S-2 eQTLs discovered through `group bridge` from the breast cancer data set. Nodes: Pink-SNV (66); Blue-Exp(47). Edges: Green-Cis (17); Orange-Trans (321).