JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 21, Number 5, 2014
© Mary Ann Liebert, Inc.
Pp. 385–393
DOI: 10.1089/cmb.2014.0026

Finding Alternative Expression Quantitative Trait Loci by Exploring Sparse Model Space

ZHIYONG WANG[1], JINBO XU[1], and XINGHUA SHI[2]

ABSTRACT

Sparse modeling, a feature selection method widely used in the machine-learning community, has been recently applied to identify associations in genetic studies including expression quantitative trait locus (eQTL) mapping. These genetic studies usually involve high dimensional data where the number of features is much larger than the number of samples. The high dimensionality of genetic data introduces a problem that there exist multiple solutions for optimizing a sparse model. In such situations, a single optimization result provides only an incomplete view of the data and lacks power to find alternative features associated with the same trait. In this article, we propose a novel method aimed to detecting alternative eQTLs where two genetic variants have alternative relationships regarding their associations with the expression of a particular gene. Our method accomplishes this goal by exploring multiple solutions sampled from the solution space. We proved our method theoretically and demonstrated its usage on simulated data. We then applied our method to a real eQTL data and identified a set of alternative eQTLs with potential biological insights. Additionally, these alternative eQTLs implicate a network view of understanding gene regulation.

Key words: expression quantitative trait locus (eQTL) mapping, redundancy, sparse modeling.

1. INTRODUCTION

DISCOVERING THE RELATIONSHIP between genetic variation and molecular traits, such as gene expression levels, is an essential step for understanding cellular functions at a systems level. Identifying expression quantitative trait loci (eQTLs) through eQTL mapping is one of such endeavors in search of genetic variation that is associated with changes in gene expression levels. An association in eQTL mapping is primarily reflected by a statistical correlation between the genotypes of a genetic variant and the expression levels of the corresponding gene in the samples. These eQTL associations provide a hypothesis that there is some underlying regulatory mechanism for further investigation. (Montgomery et al., 2010; Pickrell et al., 2010; Schlattl et al., 2011; Shabalin, 2012; Stegle et al., 2010; Fusi et al., 2012; Stranger et al., 2007; Stranger et al., 2012).

Many methods have been proposed for eQTL mapping, including a recent propagation of machine-learning approaches (Chen et al., 2012; Kim and Xing, 2012; Lee and Xing, 2012; Fusi et al., 2012; Lee et al., 2010; Wang et al., 2011). These methods either separately examine if the correlation of each pair

[1]Toyota Technological Institute at Chicago, Chicago, Illinois.
[2]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina.

of traits and genetic variants is significant or characterize their associations as parameters in a machine-learning model. Least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) methods, as a type of commonly used machine-learning method, add an l_1 norm regularization term to the loss function, leading to a sparse model that favors a sparse solution with a small number of nonzero terms. The sparsity of these algorithms is justified with an assumption that there are only a small number of associations between genetic variant and traits, given an overwhelmingly large number of variant and trait pairs for genome-wide eQTL mapping. In a solution of Lasso model fitted on observed data, the parameters represent the effect of each genetic variant on a particular trait, and those nonzero terms correspond to the identified eQTL associations. However, neither the pair-wise correlation method nor the Lasso model captures the relationship among the expression levels of multiple genes. Hence, multitask Lasso models (Chen et al., 2012; Lee and Xing, 2012; Lee et al., 2010) were proposed to impose sparsity over all of the variant and trait pairs to take consideration of related genes.

In both Lasso and multitask Lasso models, the weights of genetic variants are computed from an optimization model with its original loss function and l_1 norm of the parameters. Solving the optimization problem results in a solution with the sparse property containing only a limited number of nonzero parameters. The sparseness can be controlled by changing the coefficient of the regularization term. These optimization-based methods have the same assumption that the optimization problem has a single global optimal solution. However, the solution from Lasso methods may be inconsistent in either theory (Zou, 2006) or real situation. In eQTL mapping, the data is usually high dimensional, where the number of genetic variants is much larger than the number of samples, and hence there exists multiple solutions with the same optimal value. In such situations, the solutions resulting from different initializations are inconsistent in terms of the nonzero support of the weights. Thus, the result from a single run of these methods is unreliable, although such a simplistic approach has been widely applied in eQTL mapping. The sparse assumption in these models is also problematic in modeling biological data, since biological systems typically contain redundant and backup mechanisms. (Fishman-Lobell et al., 1992; Korn et al., 2007). Nonetheless, biological redundancy is seldom taken into account in eQTL mapping.

In this article, we propose a novel method to finding the alternative relationship between genetic variants and traits, which shows the redundancy of two genetic variants to a particular trait. Instead of studying a single optimization result, our method explores the space of the optimization solutions by repeatedly running the Lasso method with random initialization parameters. From this randomly sampled solution set, our method has the capability to extract not only the strength of each pair of genetic variant and trait, but also the relationship of two eQTLs regarding their effects on a given trait. We demonstrate the capability of our method both theoretically by a mathematical proof and practically using simulation data. We further apply our method to a real human eQTL data set and find a set of alternative eQTLs with potential biological insights. The remaining article is organized as follows: We describe our method in Section 2, present the results in Section 3, and conclude the article with discussions in Section 4.

2. METHOD

2.1. Problem definition

We denote X as the observed genotypes of the genetic variants, with D columns for D variants and N rows for N samples. We denote Y as the phenotype data, with P columns for P traits and N rows for N samples. For eQTL mapping, Y includes the expression profiles of all the genes in the samples under investigation. The method of sparse modeling is to find the optimal matrix B with D rows and P columns by minimizing the following square loss function plus a regularization term.

$$L(X, Y, B) = \|Y - XB\|_2 + \lambda \|B\|_q \tag{1}$$

The Lasso method is a special case with $q = 1$. With l_1-norm, a Lasso model favors a sparse result with a small number of nonzero terms (Candes et al., 2008). In practical situations, there are two cases when multiple solutions exist for a Lasso model, as illustrated in Figure 1.

As illustrated in Figure 1A, the loss function needs to be symmetric along the line of $x = y$. In eQTL mapping, the symmetry can be interpreted as two alternative genetic variants with redundant effect on gene expression. Therefore, we utilize this property to find alternative eQTLs, corresponding genes, or even

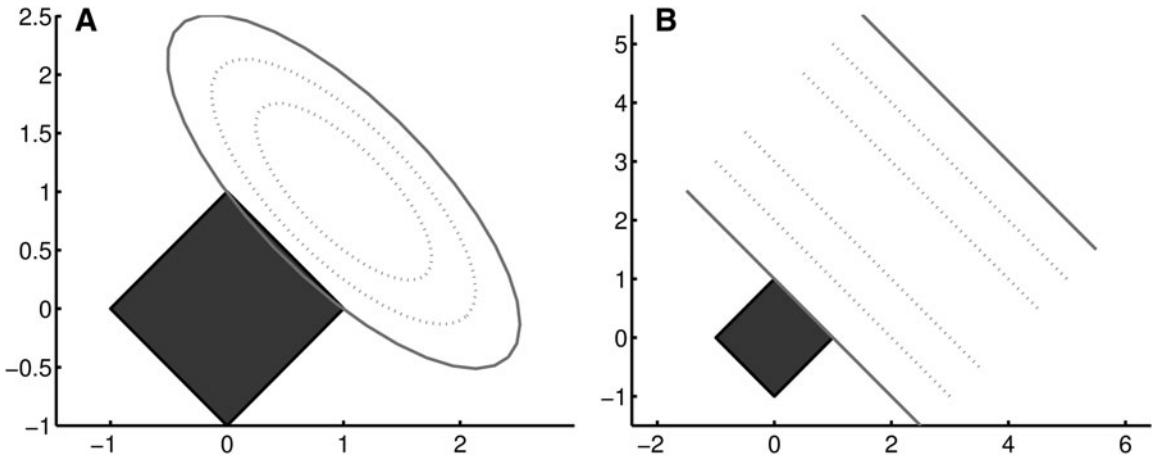

FIG. 1. Optimal solutions defined by the contour lines of the loss function and the regularization term. The axes of X and Y correspond to two weights of a solution. Dotted and solid ellipses and lines are the contour of the loss function, and gray boxes show the contour of l_1 regularization term. **(A)** A strict convex loss function and l_1 regularization term, with optimization error. **(B)** A nonstrict convex loss function and l_1 regularization term.

pathways that perform a backup function for system robustness. In this article, we aim to find such variant pairs that are redundant to each other, by investigating multiple solutions sampled from the solution space.

2.2. *Our method*

The existence of multiple optimization solutions in eQTL mapping prevents us from using a single optimized sparse model to explain the importance of features. In order to investigate the model solution space, we therefore propose a sampling method to generate a set of solutions that is expected to be a representative set covering all the multiple solutions. Consequently, we compute the covariance of the solution vectors as an approximation for the alternative relationship between two genetic variants in eQTL mapping. These identified alternative eQTLs represent a redundant relationship regarding their effect on gene expression.

Our method shown in Figure 2 can be summarized as follows. First, we search for an optimal l_1 regularization coefficient, λ, by carrying out a five-fold cross validation. Second, we sample M initialization,

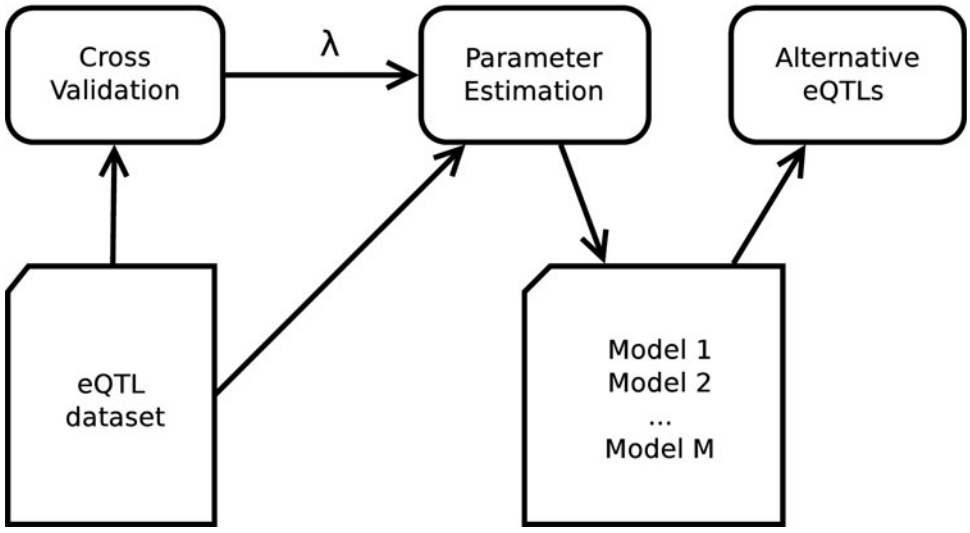

FIG. 2. The flowchart of our method. From a given eQTL data set with genotypes of genetic variants and expression profiles of genes, our method first finds the coefficient of the regularization term, λ, by a 5-fold cross validation. Then, our method samples M models through parameter estimation with different initializations. From the M models, our method finally screens for alternative eQTLs. eQTL, expression quantitative trait locus.

$B_0^{(1)}, B_0^{(2)}, \ldots, B_0^{(M)}$ independently from a uniform distribution. With these initialization values and λ, we then run Lasso M times and produce M models: $B^{(1)}, B^{(2)}, \ldots, B^{(M)}$. For each pair of genetic variants, we compute the correlation coefficient of their weights in M solutions. Finally, we extract the pairs whose correlation coefficients are equal to or greater than a cutoff c_1 and each of whom has more than c_2 nonzero values out of M models. These extracted pairs are considered as alternative eQTLs that influence the expression levels of a particular gene.

Our method can be justified in theory by the following claim. Here we describe the claim in a plain way, and the rigorous description and its proof can be found in Supplementary Material (Supplementary Material is available online at www.liebertpub.com/cmb).

Claim: Denote $b_i^{(1)}, b_i^{(2)}, \ldots b_i^{(M)}$ and $b_j^{(1)}, b_j^{(2)}, \ldots b_j^{(M)}$ as the weights of two features i and j. The correlation coefficient of $b_i^{(1)}, b_i^{(2)}, \ldots b_i^{(M)}$ and $b_j^{(1)}, b_j^{(2)}, \ldots b_j^{(M)}$ among all the M sampled models is negative when the two features have the same effect on the responding variable.

A sketch proof: The claim that two alternative features have the same effect on the responding variable can be viewed the same as the claim that we can exchange the value of the two features, but the responding value is kept the same. From this symmetry, we can derive the conclusion that if $b_1..b_i..b_j..b_p$ is an optimal solution, $b_1..b_i..b_j..b_p$ is also an optimal solution. Therefore, all the solutions of the optimization problem are paired, which implies that the correlation coefficient of b_i and b_j is negative.

We now justify our method by performing an analysis of suboptimal solution. Optimization methods may produce a solution with error caused by either the numerical precision or the convergence tolerance. We denote B^* as the optimal solution, and B^e as the solution with some small error e, such that $|L(X, Y, B^*) - L(X, Y, B^e)| < e$. We assume that the loss function $L(X,Y,B)$ has a lower bounded Lipschitz on the difference between a suboptimal solution and an optimal solution. The lower bound is denoted as u_1.

$$\frac{|L(X, Y, B^*) - L(X, Y, B^e)|}{\|B^* - B^e\|_2} \geq u_1 > 0$$

Then we have a bound for B^e dependent on the error of loss function.

$$\|B^* - B^e\|_2 < \frac{e}{u_1}$$

Thus, suboptimal solutions are also bounded by u_1 and e, which implies that suboptimal solutions are close to the optimal solution. The error e will approach zero when we implement an optimization method with small tolerance and high precision. Therefore, with a small error e, the correlation coefficient of redundant features in suboptimal solutions should be close to that of the optimal solution. ■

3. RESULTS

3.1. Simulated data

We design a simulated data set to illustrate how our model can use the correlation of sampled solutions to capture the alternative relationships between two genetic variants. This data set is generated in the following way. $X^{(2)}$ and $X^{(11)}$ are randomly selected as a pair of alternative genetic variants. Y is the responding variable, which represents the gene expression level in eQTL mapping.

$$Y_k = c_1 \max(X_2^{(2)}, X_k^{(11)}) + c_2 X_k^{(1)} \ldots + c_{19} X_k^{(20)}$$

The max function of $X^{(2)}$ and $X^{(11)}$ indicates that the two variables play the same function in this model. We choose c_i, $i = 1.19$ randomly from a uniform distribution of $\{-1, +1\}$. We perform random sampling for 100 times, where every value of X is sampled from a uniform distribution of $\{0,1\}$. We then remove all the samples where the values of $(X^{(2)},X^{(11)})$ are (0,1) or (1,0). Finally, we solve the Lasso problem on this data set with $\lambda = 0.1$ in the regularization term of Equation (1), which is computed through a five-fold cross-validation. We compute 30 solutions using the Lasso model on this simulated data set and compute the correlation between w_2 and w_{11}, the weights of $X^{(2)}$ and $X^{(11)}$, respectively. The correlation is -0.999 in this simulated data set, as shown in Figure 3. This strong negative correlation between the two weights indicates that these two variants are alternative for their effect on responding variables.

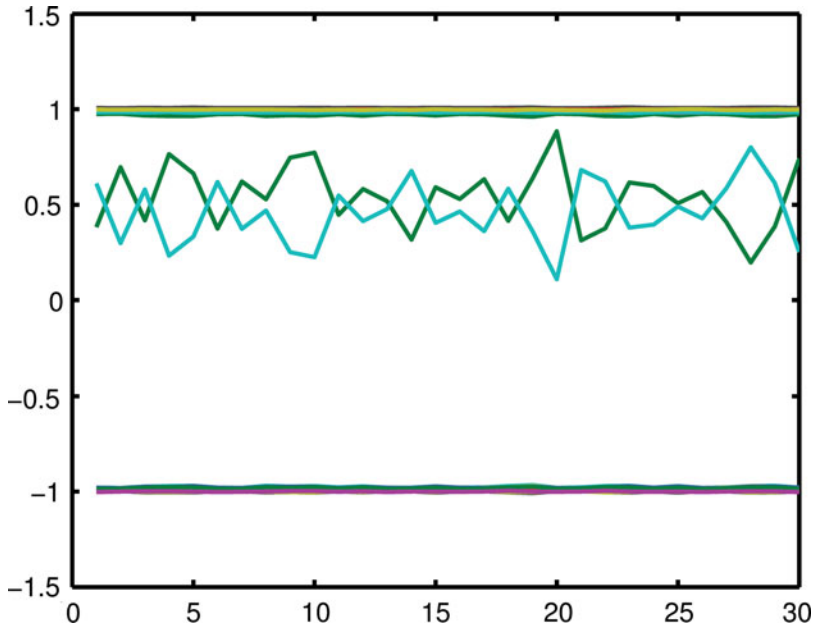

FIG. 3. Results on the simulated data. This figure shows the parameters of each model computed. *X*-axis represents the index of the model from 1 to 30. *Y*-axis shows the parameter value for each model. The weights w_2 and w_{11} correspond to the green line and the cyan line around $y = 0.5$. These two weights show a strong negative correlation with each other. The other weights are converged to -1 or $+1$ in our simulation.

3.2. Real data

We apply our method to a real data set on human eQTL mapping. Specifically, we extract the genotypes of copy number variants (CNVs) for 51 Yoruba individuals in Ibadan, Nigeria (YRI), from The 1000 Genomes Project Consortium (The 1000 Genomes Project Consortium, 2010; Mills et al., 2011). We collect gene expression levels of these same YRI individuals measured by mRNA sequencing (Pickrell et al., 2010). We apply our method to chromosome 20 of the data set, with genotypes of 139 CNVs and expression of 379 genes in 51 individuals. In total, we identify 79 CNV eQTL pairs that alternatively influence the expression of 69 target genes (Supplementary Table S1). These alternative eQTLs have high negative correlation coefficient values among models, and therefore they may alternatively affect the expression of target genes. In contrast, the correlation coefficients of their feature values are low, which prevents such pairs from being discovered by single models or by directly investigating the correlation between the genotypes of each genetic variant and the expression levels of each gene.

We use $M = 100$ models with two parameters $c_1 = -0.5$ and $c_2 = 0.3M$ on this real data set. We choose $c_1 = -0.5$ to find the negatively correlated alternative eQTLs. We choose $c_2 = 0.3M$, which implies that the selected alternative eQTLs have nonzero weight in at least 30 models. This parameter enables us to select alternative eQTLs that play significant contribution to the expression level of certain genes not by chance. The histograms of the background distribution of c_1 and c_2 are shown in Figure 4, with arrows indicating the values used to select the most significant alternative eQTLs in our method.

We perform experiments to demonstrate that the resultant set of alternative eQTLs will not change much with a small perturbation of c_1 and c_2. In Figure 5A, we illustrate the number of identified alternative eQTLs with different c_1 values ranging from -0.6 to -0.2 while fixed $c_2 = 0.3$. Since we are only interested in redundant CNV pairs with negative correlation coefficients, we choose -0.5 as a cutoff, and 79 alternative eQTLs are identified. With $c_1 = -0.4$, the number of identified alternative eQTLs is 196, which is also a small fraction (2.0%) of the total CNV pairs. With a larger cutoff $c_1 = -0.2$, which results in many eQTL pairs with low negative correlations, the number of identified alternative eQTLs is close to 800. Therefore, we suggest to carefully tune a cutoff on c_1 according to specific studies. To reflect this observation and keep our model flexible, we allow these parameters to be customized according to different data sets and studies. Figure 5A shows the number of identified alternative eQTLs with different c_2 ranging from 0 to 0.5 while fixed $c_1 = -0.5$. With c_2 larger than 0.05, the number of identified alternative eQTLs

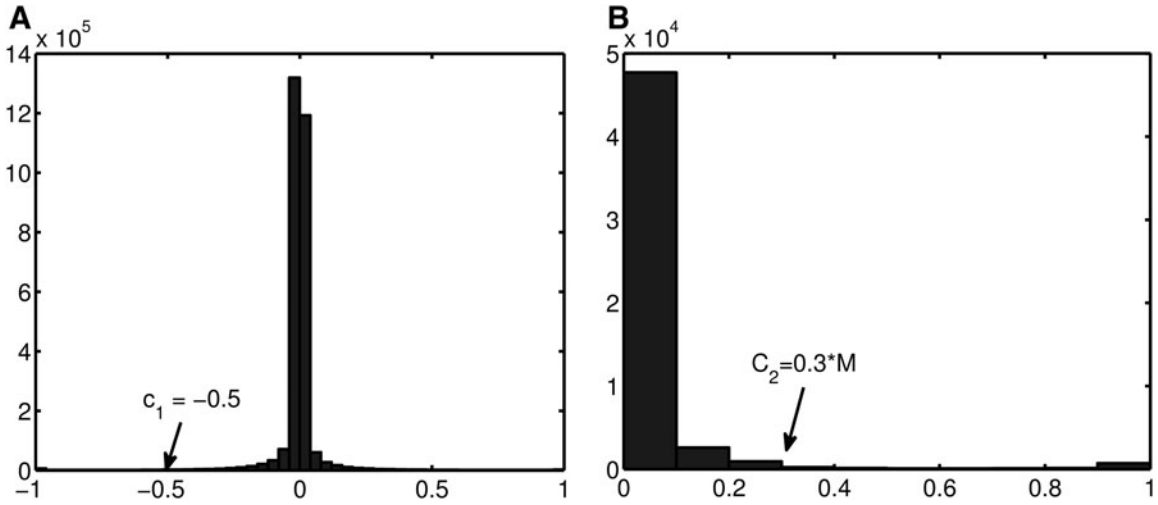

FIG. 4. The background distribution of c_1 and c_2. Both histograms are computed from 100 Lasso results with random initialization. *X*-axis represents the values of c_1 and c_2, and *Y*-axis represents the number of CNV pairs for each value interval of c_1 and c_2. **(A)** The distribution of the model weight correlation coefficient of all the alternative eQTLs. **(B)** The distribution of the fraction of the number of models in which a CNV eQTL has nonzero weight, among the total *M* models. CNV, copy number variant.

drops from 218 to 75. Considering the total of more than 10^5 CNV pairs, it is obvious that varying c_1 within the range of 0.05–0.5 would consistently select a small number of alternative eQTLs.

Our results on the real human eQTL data set, as illustrated in Figure 6, reveal a network view of the relationships between genes overlapping with alternative CNV eQTLs, and the target genes for these loci. Particularly, one CNV (chr20:60004020–60004338), overlapping with *TAF4*, is an alternative CNV eQTL of many other eQTLs targeted for different genes. *TAF4* is a subunit of transcription initiation factor *TFIID* that has been shown to empower transcriptional activation by factors including retinoic acid receptors. *TAF4* is an important gene, with the role of mediating promoter responses to various activators and repressors (www.genecards.org) (Safran et al., 2010). One of the alternative CNV eQTLs (chr20:9097570–9097622) overlaps with *PLCB4*, which has a key role in the intracellular transduction of many extracellular signals in the retina (Safran et al., 2010). *MAVS*, the mitochondrial antiviral signaling protein, is one

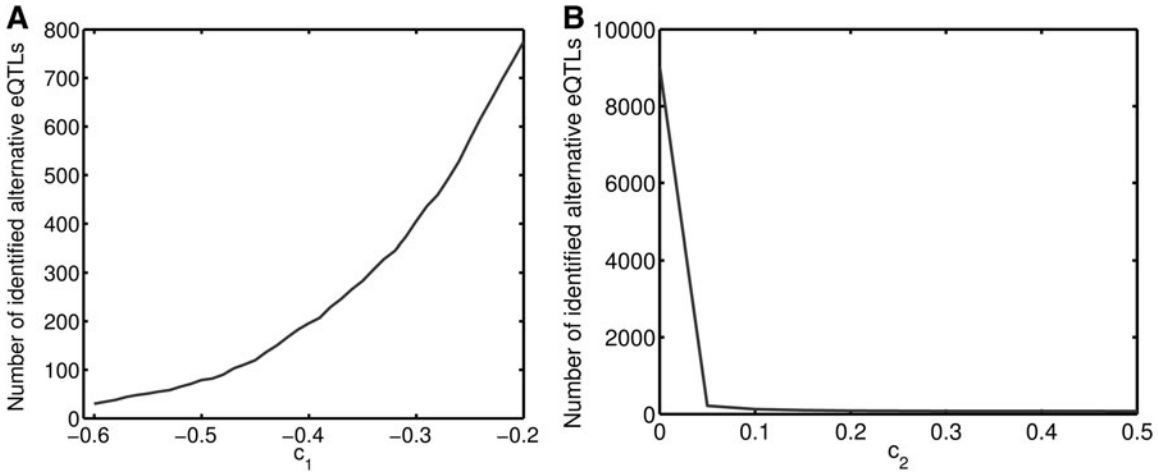

FIG. 5. The number of identified alternative eQTLs and different values of c_1 and c_2. The left panel shows that the number of identified alternative eQTLs changes with different c_1 ranging from 0 to 0.5 by fixing $c_2 = -0.5$. The right panel shows that the number of identified alternative eQTLs changes with different c_2 ranging from -0.6 to -0.2 by fixing $c_1 = 0.3$.

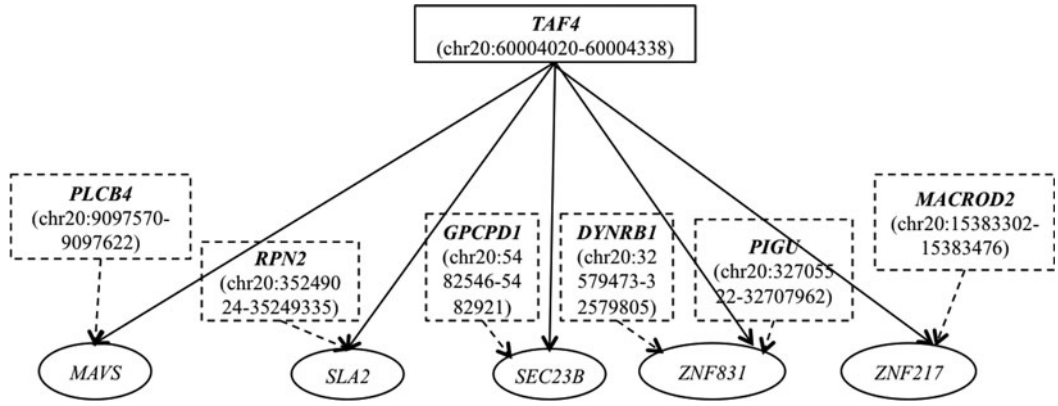

FIG. 6. A network view of the relationships between genes overlapping with alternative CNV eQTLs, and the target genes for these loci. The rectangles represent CNV eQTLs and the overlapping genes, with dotted rectangles, represent alternative CNV eQTLs. The ellipses represent target genes whose expression levels are associated with eQTLs. The arrowed lines represent the eQTL mapping identified by our method, where the CNVs are associated with target gene expression.

example of the target genes whose expression is associated with either of these two CNVs that might take effect by two different pathways. Both *TAF4* and *MAVS* are involved in the pathway of ''hsa05168,'' which is related to Herpes simplex infection, according to the Kyoto Encyclopedia of Genes and Genomes (www.genome.jp/kegg/). Hence, our results showing that *PLCB4* and *TAF4* are alternative eQTLs of *MAVS* may indicate that *PLCB4* might either involve the Herpes simplex infection pathway through an alternative route or has an alternative pathway that is different from Herpes simplex infection.

In addition, we observe that one CNV eQTL can have different alternative partners that are associated with the same target gene. For example, as shown in Figure 6, CNV chr20:60004020–60004338, overlapping with *TAF4*, is associated with the expression of *ZNF831*. This CNV has two alternative eQTL partners: CNV chr20:32579473–32579805 (overlapping with *DYNRB1*) and CNV chr20:32705522–32707962 (overlapping with *PIGU*). Specifically, the pair of (chr20:60004020–60004338, chr20:32579473–32579805) has high negative correlation values in the models running our method, as demonstrated in the scatterplot in Figure 7A. The model weights of these two CNVs regarding the expression level of *ZNF831* have a correlation coefficient of -0.574, while their genotypes are not correlated with a small correlation coefficient of 0.075. Comparatively, as shown in Figure 7B, the model weights of the

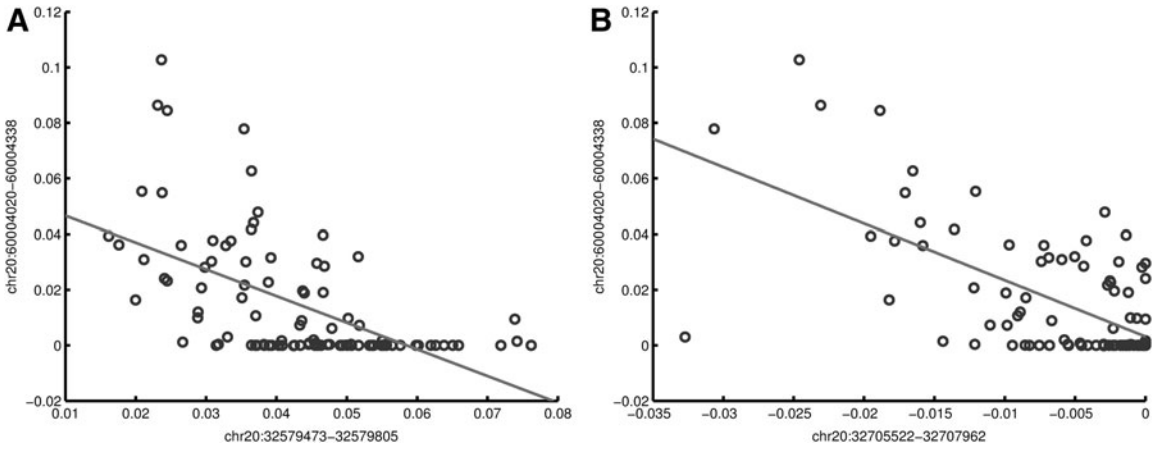

FIG. 7. Scatterplots showing model weight correlations of alternative CNV eQTLs. **(1)** chr20:60004020–60004338 and chr20:32579473–32579805. **(2)** chr20:60004020–60004338 and chr20:32705522–32707962. The axes represent the model weights of corresponding CNVs. The regression lines show the negative correlations between model weights of CNV pairs.

alternative CNV eQTLs (chr20:60004020–60004338, chr20:32705522–32707962) are highly correlated (the correlation coefficient is -0.645), whereas their genotypes have low correlation (the correlation coefficient is -0.174) over the expression of *ZNF831*. Such an example indicates that there exist multiple alternative CNV partners and corresponding pathways that affect the expression of certain genes.

4. CONCLUSION

Sparse modeling is a widely studied feature selection method in machine learning. Recent years have witnessed a trend of applying sparse modeling to genetic and genomic studies. However, the problems in such studies usually have different shapes compared with the problems in other machine-learning fields. Usually, biological systems not only desire sparsity for efficiency but also allow redundancy for robustness. The trade-off between efficiency and robustness reminds us to reconsider the sparsity in biological systems and requires careful handling in genomic studies.

In this article we make the first move to address this problem by proposing a novel method to capture the alternative relationships between two genetic variants in eQTL mapping. Our method investigates the space of all the sparse models and prioritizes potential alternative eQTLs. We present both theoretical proof and numerical experiments to support our method. The results on a real eQTL data set further provide evidence that our method serves its goal. In the real eQTL data, we use CNVs from The 1000 Genomes Project as an example. Nonetheless, our method can be applied to study other types of genetic variants, including SNPs and small insertions and deletions.

The alternative genetic variants in eQTL mapping identified by our method point to a systematic or network view of the relationship between genetic entities. These alternative eQTLs may locate at different alternative pathways or regulatory networks that lead to gene expression changes of particular genes. Therefore, a network view of gene regulation could be provided by extensive analysis of such alternative eQTLs using our method.

Our proposed method is focused on identifying redundant relationships among genetic variants for their contribution to gene expression. Such redundant relationship is straightforward yet ubiquitous in biology. In the future, we plan to extend our method to investigate more complex phenomena, including high-order relationships among genetic features.

Although we demonstrate our method in eQTL mapping in this article, this method is not limited to eQTL analysis. Our method can be extended to many other genetic studies as long as optimization is involved. For instance, we can use the same rationale to perform genome-wide association studies where the goal is to find genetic variants associated with various diseases.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Grant AF-1149811 (CAREER award), the Alfred P. Sloan Research Fellowship, and NIH R01GM089753 grants to J.X., as well as the Wells Fargo Foundation Fund for Faculty Excellence from Charlotte Research Institute and University of North Carolina at Charlotte to X.S.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press, Cambridge.

Candes, E.J., Wakin, M.B., and Boyd S.P. 2008. Enhancing sparsity by reweighted ℓ 1 minimization. *J. Fourier Anal. Appl.* 14, 877–905.

Chen, X., Shi, X., Xu, X., et al. 2012. A two-graph guided multi-task lasso approach for eQTL mapping. *J. Mach. Learn. Res. W&CP* 22, 208–217.

Fishman-Lobell, J., Rudin, N., and Haber, J. 1992. Two alternative pathways of double-strand break repair that are kinetically separable and independently modulated. *Mol. Cell. Biol.* 12, 1292–1303.

Fusi, N., Stegle, O., and Lawrence, N.D. 2012. Join modelling of confounding factors and prominent genetic regulators provide increased accuracy in genetical genomics studies. *PLoS Comp. Biol.* 8(1), p.e1002330.

Kim, S., and Xing, E.P. 2012. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *Ann. Appl. Stat.* 6, 1095–1117.

Korn, T., Bettelli, E., Gao, W., et al. 2007. IL-21 initiates an alternative pathway to induce proinflammatory TH17 cells. *Nature* 448, 484–487.

Lee, S., and Xing, E.P. 2012. Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs. *Bioinformatics* 28, i137–i146.

Lee, S., Zhu, J., and Xing, E. 2010. Adaptive multi-task LASSO: with application to eQTL detection. *Adv. Neural Inf. Process. Syst.* 23, 1306–1314.

Mills, R.E., Walter, K., Stewart, C., et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* 470, 59–65.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., et al. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–777.

Pickrell, J.K., Marioni, J.C., Pai, A.A., et al. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772.

Safran, M., Dalah, I., Alexander, J., et al. 2010. GeneCards Version 3: the human gene integrator. *Database (Oxford)* 2010:baq020.

Schlattl, A., Anders, S., Waszak, S.M., et al. 2011. Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res.* 21, 2004–2013.

Shabalin, A.A. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358.

Stegle, O., Parts, L., Durbin, R., and Winn, J. 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.

Stranger, B.E., Forrest, M.S., Dunning, M., et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848–853.

Stranger, B.E., Montgomery, S.B., Dimas, A.S., et al. 2012. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* 8, e1002639.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. *J. Roy. Statis. Soc., Series B (Methodological)* 58, 267–288.

Wang, P., Dawson, J.A., Keller, M.P., et al. 2011. A model selection approach for expression quantitative trait loci (eQTL) mapping. *Genetics* 187, 611–621.

Zou, H. 2006. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101, 1418–1429.

Address correspondence to:
Dr. Xinghua Shi
Department of Bioinformatics and Genomics
University of North Carolina at Charlotte
9201 University City Boulevard
Charlotte, NC 28223

E-mail: x.shi@uncc.edu

and

Dr. Jinbo Xu
Toyota Technological Institute at Chicago
6045 S. Kenwood Ave
Chicago, IL 60637

E-mail: jinbo.xu@gmail.com