RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

A Unified Sparse Representation for Sequence Variant Identification for Complex Traits

Shaolong Cao,[1,2] Huaizhen Qin,[2,3] Hong-Wen Deng,[2,3] and Yu-Ping Wang[1,2,3]*

[1]*Department of Biomedical Engineering, Tulane University, New Orleans, Louisiana, United States of America;* [2]*Center for Bioinformatics and Genomics, Tulane University, New Orleans, Louisiana, United States of America;* [3]*Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, Louisiana, United States of America*

Received 14 March 2014; Revised 23 June 2014; accepted revised manuscript 16 July 2014.
Published online 4 September 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21849

ABSTRACT: **Joint adjustment of cryptic relatedness and population structure is necessary to reduce bias in DNA sequence analysis; however, existent sparse regression methods model these two confounders separately. Incorporating prior biological information has great potential to enhance statistical power but such information is often overlooked in many existent sparse regression models. We developed a unified sparse regression (USR) to incorporate prior information and jointly adjust for cryptic relatedness, population structure, and other environmental covariates. Our USR models cryptic relatedness as a random effect and population structure as fixed effect, and utilize the weighted penalties to incorporate prior knowledge. As demonstrated by extensive simulations, our USR algorithm can discover more true causal variants and maintain a lower false discovery rate than do several commonly used feature selection methods. It can handle both rare and common variants simultaneously. Applying our USR algorithm to DNA sequence data of Mexican Americans from GAW18, we replicated three hypertension pathways, demonstrating the effectiveness in identifying susceptibility genetic variants.**
Genet Epidemiol 38:671–679, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: **sparse regression; relatedness; population structure; prior biological information; Mexican Americans**

Introduction

Complex traits are likely to be influenced by many rare and common genetic variants and environmental covariates. Next-generation sequencing technologies provide great potential for identifying both rare and common sequence variants. Single-marker association tests bear poor statistical power to identify associated rare variants due to their very low frequencies. Generalized linear model provides an effective approach to identify variant sets while adjusting for covariates of unrelated individuals [Lee et al., 2012; Yi et al., 2011]. However, the assumption of independence between individuals is frequently violated in sequence association studies. In the presence of complex pedigree structure or/and cryptic relatedness, it is challenging to correct for population structure [Price et al., 2010], especially for rare variants detection [Mathieson and McVean, 2012]. Most existing sequence association methods do not jointly model relatedness, population structure, and covariates.

Accurately pinpointing specific causal variants is necessary for elucidating genetic architecture of a complex disease. Sparse representation models were established to select a promising sparse set from a large number of variants [Wu et al., 2009; Zhou et al., 2010], e.g., those within a gene or a pathway. Such models allow the size of a testing set (gene or pathway) exceed the number of study participants [Fan and Li, 2001; Zou and Hastie, 2005] by the use of regularization terms (e.g., L_0 norm and L_1 norm). Although the L_0 norm penalty yields sparsest solution, its discontinuity makes the problem to be NP-hard [Natarajan, 1995], which is nearly infeasible for the regression model with a large number of predictors. The L_1 norm penalty or least absolute shrinkage and selection operator (Lasso) is a well-developed and computationally feasible method, with the relaxation of L_0 norm penalty. If a particular restricted isometric property holds, the solutions of Lasso and L_0 norm penalty are identical [Candes and Tao, 2005]. However, this restriction is too strong to have practical value. Recently, L_p norm ($0 < p < 1$), as an alternative relaxation, has aroused more interests, which yields more sparse solutions than does the Lasso. Despite these merits, existent sparse representation algorithms still suffer the limitations of aforesaid set (e.g., gene, pathway) based association methods.

Tremendous DNA sequence data are of complex population structure and relatedness, including known pedigree structure and cryptic relatedness. These confounders, if not appropriately adjusted for, may inflate false-positive rates or deflate false-negative rates. Incorporating prior biological information can boost statistical power. In this article, we developed a unified sparse regression (USR) as an effective solution to incorporate prior information and jointly adjust for relatedness, population structure, and environmental

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Yu-Ping Wang, Lindy Boggs Center Suite 500, New Orleans, LA 70118, USA. E-mail: wyp@tulane.edu

© 2014 WILEY PERIODICALS, INC.

covariates. Our algorithm adopts a modified kinship matrix to account for the confounding of complex relationship between pedigree members on a quantitative trait [Thompson and Shaw, 1990]. For the data of cryptic relatedness, we infer the kinship matrix by the REAP algorithm [Thornton et al., 2012]. Meanwhile, our USR models population structure and other environmental covariates as fixed effects. To allow proper sparsity and incorporate prior knowledge, our USR algorithm applies a weighted regularization with L_p norm ($0 < p < 1$) to select sparse representation—a sparse subset from a large number (>sample size) of markers. Our algorithm can automatically search for a sparse representation and allow users to determine the size of output set. As demonstrated by extensive empirical comparisons and real DNA sequence data analyses, our USR appears more effective than do many existent sparse regression models.

Methods

In our USR model, relatedness is treated as a random effect, and population structure is treated as a fixed effect. This model allows an arbitrary relatedness as captured by a corresponding kinship matrix. For an arbitrary $p \in (0,1)$, the L_p norm regularization is neither convex nor Lipchitz continuous. To solve the L_p problem, we first compute the explicit solution of the $L_{0.5}$ problem [Xu et al., 2012b] for an initial point. Next, we solve the smoothed surrogate model of L_p norm regularization by the smoothing conjugate method [Chen et al., 2010]. To improve accuracy, we modify the elastic-net regularization [Cho et al., 2010; Friedman et al., 2007, 2010] to adjust for relatedness and choose an optimal penalty parameter λ in terms of the Akaike information criteria (AIC). Lastly, we use the stability selection method to estimate the sparse regression coefficients.

The USR Method

Let n denote the total number of subjects, and m denote the number of independent variables. Let $Y = (y_1, y_2, \ldots, y_n)^T$ contain the trait values of the n subjects. We write $X = (x_1, x_2, \ldots, x_n)^T$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$ contains genotypic scores of subject *i at m markers, g*enotypic score x_{ij} is the individual copy number of the minor allele at marker *j*; and $W = (w_1, w_2, \ldots, w_L)$, where $w_i = (w_{i1}, w_{i2}, \ldots, w_{in})^T$ represents fixed-effect confounders, e.g., population structure surrogates, age, and gender.

Joint Adjustment of Confounders

For data with a known pedigree structure, we consider linear mixed-effect model:

$$Y = W\alpha + X\beta + \varepsilon \tag{1}$$

where $\varepsilon \sim N(0, \Sigma)$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_L)^T$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_m)^T$ are vectors of corresponding regression coefficients. The error term ε summarizes the random effect due to pedigree structure [Thompson and Shaw, 1990] and environmental residual. To be explicit, $\Sigma = \sigma_\Phi^2 \Phi + \sigma_e^2 I_{n\times n}$, where Φ is the kinship matrix; Φ_{ij} equals to twice the kinship coefficient between subjects *i* and *j*; and I is identical matrix of order n. We use EMMA software [Kang et al., 2008] to estimate the variance components in Σ under the null of $\beta = 0$ and fix them in our USR. For the data of cryptic relatedness, the kinship matrix can be inferred by extent algorithm, e.g., the REAP [Thornton et al., 2012]. For a given Φ, the likelihood can be formulated as

$$L(\alpha, \beta) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{(Y - W\alpha - X\beta)^T\Sigma^{-1}(Y - W\alpha - X\beta)}{2}\right)$$

The log-likelihood is

$$\ell(\alpha, \beta) = -\log(L(\alpha, \beta)) \propto (Y - W\alpha - X\beta)^T\Sigma^{-1}(Y - W\alpha - X\beta)$$

The Generic L_p Regularization

A general form of regularized regression is given by

$$(\alpha_{opt}, \beta_{opt}) = \arg\min_{\alpha,\beta}\{\ell(\alpha, \beta) + P_\lambda(\beta)\} \tag{2}$$

It is well known [Chen et al., 2010; Xu et al., 2012b] that L_p ($0 < p < 1$) norm regularization term can give more sparse solution than L_1 norm based regularization, also known as the famous Lasso. If we define the L_p norm based regularization term as

$$P_\lambda(\beta) = \lambda||\beta||_p^p = \lambda\sum_{i=1}^{n}|\beta_i|^p, \quad 0 < p < 1,$$

then the problem becomes to find the minimizer

$$\begin{aligned}(\alpha_{opt}, \beta_{opt}) &= \arg\min_{\beta\in R^m, \alpha\in R^L} f(\alpha, \beta) \\ &:= (Y - W\alpha - X\beta)^T\Sigma^{-1}(Y - W\alpha - X\beta) + \lambda||\beta||_p^p\end{aligned} \tag{3}$$

In particular, if the data only contain unrelated subjects, i.e.,$\Phi = I$, $\Sigma = (\sigma_\Phi^2 + \sigma_e^2)I$, Eq. (3) collapses to the classic least square sparse regression. Similar to other sparse regressions, we define the selected risk variants to be the set of nonzero regression coefficients, i.e., $\{i|\beta_i \neq 0\}$.

Incorporating Prior Information

The regularization term in Eq. (3) can be modified to incorporate prior knowledge. For this purpose, we introduce a weighted regularization term. To be explicit, the weighted L_p norm regularization is

$$\begin{aligned}(\alpha_{opt}, \beta_{opt}) &= \arg\min_{\beta\in R^m, \alpha\in R^L} \\ &(Y - W\alpha - X\beta)^T\Sigma^{-1}(Y - W\alpha - X\beta) + \lambda||\gamma\beta||_p^p\end{aligned} \tag{4}$$

where $\gamma\beta = (\gamma_1\beta_1, \gamma_2\beta_2, \ldots, \gamma_m\beta_m)^T$ and γ_j's (>0) represent marker-wise weights.

An appropriate choice of weights can improve statistical power. Each weight γ_j is prespecified, taking the genotypes, covariates and prior knowledge into account. The weight γ_j reflects the relative importance or preference of the jth variant. On one hand, we can assign a particular marker with small penalty weight, if we want to include the marker into the sparse representation. On the other hand, a marker with a large weight is more likely to be excluded from the sparse representation.

There are several ways to determine the weights. For example, we can give nonsynonymous SNPs or the SNPs in the risk gene lower weights to increase their chances to enter the model. Another way to assign weights is based on minor allele frequency. When analyzing rare and common variants together, we can assign lower weights to rare variants, in order to compensate for their low frequencies. Because in practice we do not know exactly which variants have high risk, the weights should be assigned prudently.

In particular, if all $\gamma_j = 1$, Eq. (4) collapses to Eq. (3), which is an unweighted one. The algorithm to solve Eq. (4) is almost the same as for Eq. (3). The only difference is to replace β_j by $\gamma_j \beta_j$. For the sake of simplicity, we just present the algorithm for solving Eq. (3). We assume all variants are equally weighted unless otherwise stated.

Solving the USR Problem

Generally, the L_p $(0 < p < 1)$ norm based regularization (Eq. (3)) is neither convex nor Lipschitz continuous, making the solution computationally difficult and time-consuming. We adopt the basic idea on nonconvex and noncontinuous optimization [Zhang and Chen, 2009] to solve the minimization problem of (3). To make the algorithm more stable and faster, we establish a lower bound to further regularize local optimal solution. Another issue with L_p norm regularization is that the iterative algorithm can be easily trapped at a local minimizer. Therefore, the choice of the initial point is crucial for the iterative algorithm. For this reason, we use the solution of $L_{0.5}$ norm regularization as the initial point for the L_p regularization problem. The details are discussed below.

Smoothing Method for L_p Norm Regularization

We use a smoothing approximation to the objective function in (3) [Chen et al., 2010]

$$f_\mu(\alpha, \beta) = (Y - W\alpha - X\beta)^T \Sigma^{-1} (Y - W\alpha - X\beta) + \lambda ||\psi_\mu(\beta)||_p^p$$

where $\psi_\mu(\beta) = (s_\mu(\beta_1), s_\mu(\beta_2), \ldots, s_\mu(\beta_n))^T$ and

$$s_\mu(x) = \begin{cases} |x| & |x| > \mu \\ \dfrac{x^2}{2\mu} + \dfrac{\mu}{2} & |x| \le \mu \end{cases}$$

The smoothing function $f_\mu(\alpha, \beta)$ is continuously differentiable and strictly convex on the set of $\{x | \max(x) \le \mu\}$. Moreover,

$$\lim_{\mu \downarrow 0} f_\mu(\alpha, \beta) = f(\alpha, \beta)$$

All these properties show that this smoothing function is a good approximation to the original one but makes the problem easy to solve.

Lower Bound Theory

Based on the first- and second-order necessary condition on the solution to the minimization problem, we derive a lower bound, and a sufficient and necessary condition to narrow the search of nonzero entries and guide the selection of causal variants. In our algorithm, we utilize the lower bound at each step to help refine the local minimizer.

The lower bound theory for the unified sparse model

Denote X_p^* the set of local minimizers of objective formula (3)
For any $\beta^* \in X_p^*$ derived from initial point β^0, the following statements hold:
(i) Let $L_i = \max[\frac{\lambda p(1-p)}{2(A^T\Sigma^{-1}A)_{ii}}^{\frac{1}{2-p}}, (\frac{\lambda p\sqrt{K}}{2||A||||\Sigma^{-1}||\sqrt{f(\alpha,\beta^0)}})^{\frac{1}{1-p}}]$ for any
$\beta_i^* \in (-L_i, L_i) \Rightarrow \beta_i^* = 0$ where $A := X_\Lambda \in R^{n\times|\Lambda|}$ is a submatrix of X, which consists of the jth columns of X, with $j \in \Lambda$,
$\Lambda = \text{support}(\beta^*) = \{i | \beta_i^* \neq 0\}$, $K = ||\beta^*||_0$
(ii) The smallest eigenvalue of matrix $\tilde{A}\tilde{B}^{-1}$: $\lambda_{\min} \ge 1$; where $\tilde{A} = 2A^T\Sigma^{-1}A$ and $\tilde{B} = \lambda p(1-p)diag(|\beta_i^*|^{p-2})$, and $\beta_i^* \neq 0$

The detailed proof of this theory is described in the supplementary material.

$L_{0.5}$ Norm Regularization

The $L_{0.5}$ norm regularization has an analytical threshold operator [Xu et al., 2012b] compared with arbitrary L_p $(0 < p < 1)$ norm problem, which can be easily and fast solved. In addition, the $L_{0.5}$ regularization always yields more sparse solution than that of using L_p when $1/2 < p < 1$, and shows no significant difference from the one when $0 < p < 1/2$ [Xu et al., 2012a]. Thus in our algorithm, we first apply $L_{0.5}$ thresholding algorithm [Xu et al., 2012b] to obtain the solution of $L_{0.5}$ problem and then use the solution of $L_{0.5}$ as the initial point to search the minimizer of the L_p norm based regularization problem.

The $L_{0.5}$ regularization model is given by the following formulation (5)

$$\min_{\beta \in R^m, \alpha \in R^L} (Y - W\alpha - X\beta)^T \Sigma^{-1} (Y - W\alpha - X\beta) + \lambda ||\beta||_{1/2}^{1/2} \tag{5}$$

where $||\beta||_{1/2} = (\sum_{j=1}^m \sqrt{|\beta_j|})^2$

According to [Xu et al., 2012b], the solution of (5) can be obtained by the following thresholding operation

$$\beta^* = R_{\lambda\mu,1/2}(\beta^* + \mu X^T \Sigma^{-1} (Y - W\alpha^* - X\beta^*))$$

where $R_{\lambda\mu,1/2}(\bullet) : \mathbb{R}^n \to \mathbb{R}^n$ is the half thresholding operator. It is given as follows:

$$R_{\lambda\mu,1/2}((x_1, x_2, \ldots x_m)^T) = (f_{\lambda\mu,1/2}(x_1), f_{\lambda\mu,1/2}(x_2), \ldots, f_{\lambda\mu,1/2}(x_m))^T$$

where

$$f_{\lambda\mu,1/2}(x) = \begin{cases} \frac{2}{3}x\left(1+\cos\left(\frac{2\pi}{3}-\frac{2}{3}\varphi_{\lambda\mu}(x)\right)\right) & |x| > \frac{\sqrt[3]{54}}{4}(\lambda\mu)^{\frac{2}{3}} \\ 0 & otherwise \end{cases}$$

and $\varphi_{\lambda\mu}(\bullet)$ satisfies $\varphi_{\lambda\mu}(x) = \arccos(\frac{\lambda\mu}{8}(\frac{|x|}{3})^{-\frac{3}{2}})$, $\mu = ||X||_2$

We name our algorithm to solve the problem (5) as the hybrid $L_{0.5}$-SCG algorithm, where SCG stands for the smoothing conjugate gradient.

Unified sparse representation algorithm

Step 1: Data normalization:
$\sum_{i=1}^{n} x_{ij} = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$, for $j = 1,2,\ldots,m$.
Step 2: For any given λ, p, set iterative index $r = 0$, $\varepsilon = 0.0001$; initialize $\alpha_l^{(0)} = 0$, for $l = 1,2,\ldots,L$; $\beta_j^{(0)} = 0$, for $j = 1,2,\ldots,m$.
Step 3: Update $\alpha^{(r+1)} = (W^T W)^{-1} W^T (Y - X\beta^{(r)})$.
For $j = 1,\ldots,m$,
update $\beta^{(r+1)} = R_{\lambda\mu,1/2}(\beta^{(r)} + \mu X^T \Sigma^{-1}(Y - W\alpha^{(r+1)} - X\beta^{(r)}))$.
Step 4: Apply the lower bounds to regularize $\beta^{(r+1)}$ and use the SCG algorithm (Zhang et al., 2009) with the initial point $\beta^{(r+1)}$ to find the minimizer $\beta_p^{(r+1)}$ of objective function (3).
Step 5: Calculate $||\beta_p^{(r+1)} - \beta_p^{(r)}||_{l_2}$
If $||\beta_p^{(r+1)} - \beta_p^{(r)}||_{l_2} < \varepsilon$ stop; otherwise return to Step 3.

Then, $\beta_p^{(r+1)}$ is the final solution.

Tuning Parameter Selection

It is well known that the setting of regularization (tuning) parameter λ in Eqs. (3)–(5) controls the trade-off between data fitting fidelity and the use of prior knowledge. A larger λ results in a more sparse solution and vice versa.

The selection of optimal regularization parameters is a difficult problem. If computing time is not a concern, it is helpful to optimize the objective function over a grid of points and monitor how new predictors enter the model as λ decreases. Another way is to minimize either the Bayesian information criterion (BIC) or AIC as a function of λ. Also, we can use cross-validation to select optimal λ. After the comparison of these methods in our simulations, we choose the AIC as our variable selection criterion. For our model, we have the following form of AIC [Cetin and Erar, 2002]:

$$AIC = 2k + n(\log((Y - W\alpha - X\beta)^T \Sigma^{-1}(Y - W\alpha - X\beta)) + 1)$$

The goal is to find an optimal λ so that the AIC value can be minimized. Because λ is the key parameter to determine the sparsity level, it is crucial to understand the relationship between AIC and λ. However, there is no explicit expression of AIC(λ). So we use the discrete search in log-scale to find the optimal λ that yields the smallest AIC value.

However, a major drawback of AIC procedure is that it cannot control false-positive rate or family-wise error rate. So we use the idea of stability selection [Meinshausen and Bühlmann, 2010] to further control the false-positive rate based on the selected λ.

The basic idea about stability selection is to bootstrap the data, and then calculate the frequency of the variables to be selected. The higher frequency of the selected variables implies that they are more important. Hence, we can develop a new rank of importance of each variable (i.e., variants), and then a frequency threshold is applied to select final risk variants. An advantage of the stability selection over AIC selection is that the expected number of falsely selected variables or false-positive rate can be asymptotically controlled.

The detailed procedure of hybrid AIC and stability selection is described in the algorithm part of supplementary material. Despite the hybrid AIC and stability selection procedure, we also provide an adaptive method to make the solution to have predetermined k-sparsity.

Unified L_p algorithm with predetermined sparsity

Step 1: Data normalization:
$\sum_{i=1}^{n} x_{ij} = 0$, $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$, for $j = 1,2,\ldots,m$.
Step 2: For any predetermined sparsity level k, set iterative index $r = 0$, $\varepsilon = 0.0001$; initialize $\alpha_l^{(0)} = 0$, for $l = 1,2,\ldots,L$; $\beta_j^{(0)} = 0$, for $j = 1,2,\ldots,m$.
Step 3: Set $B^{(r)} = \beta^{(r)} + \mu X^T \Sigma^{-1}(Y - W\alpha^{(r)} - X\beta^{(r)})$, and denote $[|B^{(r)}|]_k$ to be the kth largest element of $|B^{(r)}|$.
Step 4: Update $\lambda^{(r)} = \frac{\sqrt{96}}{9\mu}([|B^{(r)}|]_k)^{\frac{3}{2}}$.
Step 5: Update $\alpha^{(r+1)} = (W^T W)^{-1} W^T (Y - X\beta^{(r)})$.
For $j = 1,\ldots,m$,
update $\beta^{(r+1)} = R_{\lambda^{(r)}\mu,1/2}(\beta^{(r)} + \mu X^T \Sigma^{-1}(Y - W\alpha^{(r+1)} - X\beta^{(r)}))$.
Step 6: Apply the lower bounds to regularize $\beta^{(r+1)}$ and use the SCG algorithm [Zhang and Chen, 2009] with the initial point $\beta^{(r+1)}$ to find the minimizer $\beta_p^{(r+1)}$ of objective function (3).
Step 7: Calculate $||\beta_p^{(r+1)} - \beta_p^{(r)}||_{l_2}$.
If $||\beta_p^{(r+1)} - \beta_p^{(r)}||_{l_2} < \varepsilon$ stop; otherwise return to Step 3.

The $\beta_p^{(r+1)}$ is the final outcome with k-sparsity.

Results

In this section, we empirically compared our USR algorithm with single-marker test (χ^2 test), elastic-net, orthogonal matching pursuit (OMP), focal underdetermined system solver (FOCUSS) [Rao and Kreutz-Delgado, 1999], Random Forest [Chen and Ishwaran, 2012; Goldstein et al., 2010], and Gemma [Zhou and Stephens, 2012]. We first compared these algorithms under our own simulation design with and without family structure. In addition, we compared the algorithms under the simulated data from Genetic Analysis Workshop 17 (GAW17).

Simulation I: Unrelated Individuals

To validate our USR, we performed simulation experiments based on the Encyclopedia of DNA Elements (ENCODE) data. This dataset contains 522 haplotypes and 1,688 SNPs. We used this haplotype pool to generate genotypes and the corresponding phenotypes, i.e., X and Y, respectively, in the linear mixed-effect model (3). The phenotypes are simulated based on the linear model with assigned causal SNPs under the controlled heritability. We implemented $L_{0.1}$, $L_{0.9}$, $L_{0.5}$, elastic-net, OMP, and FOCUSS methods, respectively. The elastic-net and weighted elastic-net programs were developed according to Friedman's papers [Friedman et al., 2007, 2010]; OMP and FOCUSS [Rao and Kreutz-Delgado,

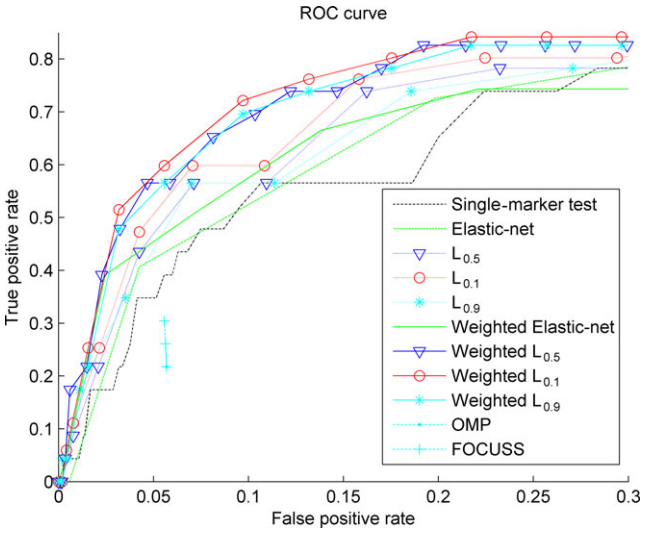

Figure 1. Partial ROC for methods comparison under population design of 1,000 unrelated individuals. Each point (FPR, TPR) corresponds to a specific λ value, where FPR is false-positive rate, and TPR is true positive rate.

1999] programs were downloaded from the link in their publications.

In this simulation, we generated 1,000 samples and give the weight for each marker as follows: $weight = 2\sqrt{MAF(1-MAF)}$ where MAF is the minor allele frequency.

The detailed procedure of our experiment is as follows:

Step 1: Set the risk haplotype ratio to be 25% (risk haplotypes/all haplotypes); set the iterative index $k = 0$, $I(0) = \emptyset$.

Step 2: $k = k + 1$; randomly select an SNP as causal variant $C(k)$; count the index of the haplotypes that contain $C(k)$, and denote this index set as $I(k)$ (risk haplotypes).

Step 3: $I(k) = I(k-1) \cup I(k)$; if $I(k) > 0.25$, jump to Step 4, otherwise return to Step 2.

Step 4: Generate 10,000 genotype samples from the pool randomly.

Step 5: Calculate each sample's genetic score S, i.e., how many risk haplotypes this sample has; $S = 0, 1, 2$.

Step 6: Generate each sample's phenotype: $y = b * S + \varepsilon$, $\varepsilon \sim N(0,1)$, $b = sqrt(0.01/(0.99 * var(S)))$.

To evaluate our methods, we compared them with the single-marker test (χ^2 test), elastic-net, OMP and FOCUSS respectively. We also extended our family adjustment and weighted model to the one with elastic-net penalty. For the numerical algorithm, we used the cyclical coordinate descent, computed along a regularization path.

In this article, the TPR is defined by the number of selected true variants divided by the total number of true variants; and the FPR is defined by the number of selected false variants divided by the total number of false variants.

Table 1. The error rate of using optimal λ selected by the AIC

$N = 1{,}000$, $H^2 = 0.05$	TPR	FPR
Elastic-net	0.0745	0.0151
$L_{0.5}$	0.0805	0.0213
$L_{0.1}$	0.0021	1.5883×10^{-4}
$L_{0.9}$	0.0018	1.2202×10^{-4}

Table 2. The error rate of variables by the hybrid AIC and stability selection method

$N = 1{,}000$, $H^2 = 0.05$	TPR	FPR
Elastic-net	0.0729	0.0142
$L_{0.5}$	0.0818	0.0208
$L_{0.1}$	0.0337	2.0743×10^{-3}
$L_{0.9}$	0.0261	2.3173×10^{-3}

From Figure 1, by calculating the area under the receiver operating characteristic (ROC) curve (AUC), we can conclude that weighted models with the use of $L_{0.5}$ and $L_{0.1}$ regularization term perform best among all the methods listed above, and the classic single-marker test (χ^2 test) has the lowest power. The FPR and TPR of FOCUSS and OMP methods were stuck in a low range, which is difficult to perform a comparison of AUC. In addition, FOCUSS became unstable with the tuning parameter getting larger and its TPR decrease with FPR increase. For the sake of stability and efficiency, we did not perform OMP and FOCUSS methods in the following sections.

Tables 1 and 2 are generated by the average of 100 replicate simulations with 1,000 samples and 0.05 heritability. Apparently, the best method should have the highest TPR, whereas lowest FPR. However, there always exists a trade-off between TPR and FPR.

By comparing Table 1 with Table 2, we find that $L_{0.5}$ and elastic-net had quite similar performance under both AIC and stability selection. Under AIC, $L_{0.1}$ and $L_{0.9}$ appeared to be too conservative and yielded extremely low FPR and TPR. However, the stability selection rectified the conservativeness to make corresponding FPR closer to the preset type I error threshold (0.05). Therefore, we recommend hybrid stability selection with AIC as a better choice and just present results of using hybrid AIC and stability selection in the following sections. Furthermore, the $L_{0.9}$ was shown not as good as $L_{0.1}$ and $L_{0.5}$, which is also supported by the part of "comparison of different L_p norm regularization" in the supplementary material. So we mainly focused on $L_{0.1}$ and $L_{0.5}$ regularization methods for the remaining of the article.

Simulation II: Admixed Families

We downloaded the genotype data of region ENr113.4q26 from the ENCODE project Consortium. We inferred 180 CEU (Centre d'Etude du Polymorphisme Humain in Utah, USA) and 180 YRI (Yoruban in Ibadan, Nigeria) haplotypes. We observed 1,693 SNPs in total. At each SNP, we chose

the minor allele in the YRI haplotype data ($f_{YRI} \leq 0.5$) as the reference allele. Following previous association study on African Americans [Qin et al., 2010], we adopted $\omega = 0.8$ vs. $\varpi = 0.2$ as YRI-CEU admixture weights. To "genotype" one admixed subject in the ENr113.4q26 region, we randomly chose one and another haplotype from the YRI or CEU haplotype datasets with probabilities ω vs. ϖ. In this simulated admixture, the frequencies of reference alleles at the 1,693 SNPs ($f_{ADX} = \omega f_{YRI} + \varpi f_{CEU}$) range from 0.0011 to 0.5722, and 295 SNPs are of $f_{ADX} \leq 0.02$. This simulation design includes three major steps.

Step 1. Generate Parental Dataset

For each family, we generated father and mother independently. Each subject is composed of two haplotypes; each time we have 80% chance of randomly selecting a haplotype from YRI, and 20% from CEU. The local ancestry $a_i \in \{0, 1, 2\}$ for the ith subject is the number of haplotypes from YRI data. This design does not model recombination in the small region (ENr113.4q26). Therefore, ancestry a_i is the same for all SNPs in a specific subject.

Step 2. Generate Nuclear Family With Two Children

Two children are generated for each family. To generate one child, we randomly selected one haplotype from father and the other from mother. We simulated $N(=200)$ families with the same family structure, which is composed of two parents with two children.

Step 3. Generate Trait Values

To be explicit, for each person, we use the following model to generate trait values.

$$Y_i = bX_i\beta + \varepsilon_i, (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n)^T \sim N(0, \Sigma) \qquad (6)$$

where $\Sigma = diag(\Sigma_1, \Sigma_2, \ldots, \Sigma_N)$

In our simulation, the covariate matrix for each family is

$$\Sigma_j = \frac{2}{3}\Phi + \frac{1}{3}I$$

where $\Phi = \begin{pmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$

is the kinship matrix. To simulate heritability $H^2 = 0.05$, the true model we used is formula (6) where $b = \frac{H}{\sqrt{(1-H^2)Var(X\beta)}}$, and $\Sigma = diag(\Sigma_1, \Sigma_2, \ldots, \Sigma_N)$; in our simulation, the covariate matrix for each family is $\Sigma_j = \frac{2}{3}\Phi + \frac{1}{3}I$.

Using this model, we mainly compared the results with and without pedigree adjustment. Thus, we evenly assigned causal variants to include both rare and common variants and exclude the influence of weighted method. In this simulation, we did not consider any prior knowledge and set all the weight coefficients to be 1.

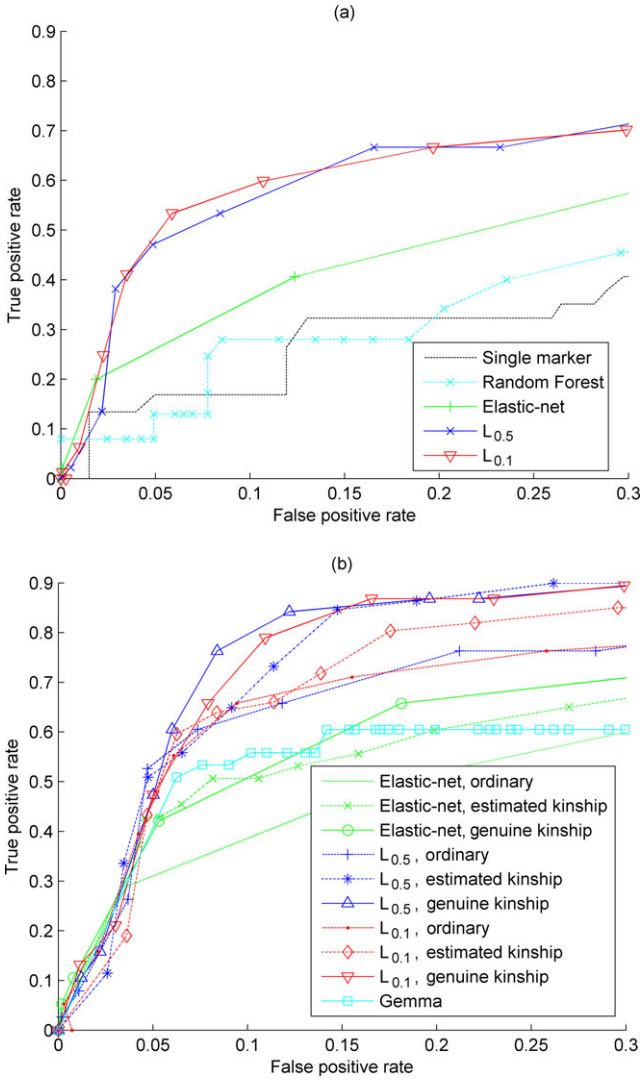

Figure 2. Methods comparison under family design of 200 unrelated nuclear families. (A) The partial ROC curves of five methods without adjusting for relatedness. (B) The partial ROC curves of two methods with adjusting for relatedness (by estimated kinship matrix and genuine kinship matrix) vs. the ordinary method without adjusting for relatedness.

First, we compared our USR algorithm with other feature selection methods (e.g., Random Forest and Gemma), using the genuine family structure. Second, to illustrate the capability of the USR to adjust for cryptic relatedness, we inferred the kinship matrix using the REAP [Thornton et al., 2012] and adopted the inferred kinship matrix when applying our USR.

Figure 2(A) shows the comparison among several methods without adjusting for family structure. For the Random Forest, we ranked the variables by their importance factors, and then selected different number of variables. Finally, we drew the corresponding ROC. The ROC and AUC indicate that $L_{0.1}$ is the best method and the single-marker test performs the worst.

Table 3. The error rate of variables selected by hybrid AIC and stability selection method

$N = 800$, $H^2 = 0.05$	TPR	FPR	Partial AUC	AUC
Elastic-net, ordinary	0.0811	0.0223	0.1012	0.6693
Elastic-net, estimated kinship	0.1291	0.0237	0.1152	0.6964
Elastic-net, genuine kinship	0.1351	0.0243	0.1316	0.7271
$L_{0.5}$, ordinary	0.0811	0.0169	0.1737	0.7789
$L_{0.5}$, estimated kinship	0.2320	0.0201	0.1758	0.7808
$L_{0.5}$, genuine kinship	0.2432	0.0205	0.1891	0.8126
$L_{0.1}$, ordinary	0.0435	3.623×10^{-3}	0.1714	0.7884
$L_{0.1}$, estimated kinship	0.0501	4.521×10^{-3}	0.1918	0.8063
$L_{0.1}$, genuine kinship	0.0526	4.753×10^{-3}	0.2125	0.8198
Gemma			0.1648	0.5941

In Figure 2(B), adjusting genuine relatedness and estimated relatedness outperformed the ordinary regression, which ignores the relatedness. The result of the USR using estimated kinship matrix is close to that of the USR using genuine kinship matrix. Hence, the estimated kinship based USR method is reliable for cryptic relatedness data analysis. The adjustment of a real kinship matrix appeared a bit better. The ROC also indicates that when the FPR or type I error is low, the $L_{0.1}$ solution is the best choice; when the FPR is higher, the $L_{0.5}$ solution is the best choice. The Gemma method and elastic-net sparse representation falls in between, indicating that it is a more stable solution. The Gemma method and elastic-net method are comparable and both outperform the single-marker test and Random forest.

Table 3 is generated by the average of 100 replicate simulations with 200 nuclear families (800 samples) and 0.05 heritability. The partial AUC is calculated based on the cutoff (FPR = 0.3). In terms of AUC and partial AUC, the $L_{0.1}$ and $L_{0.5}$ family adjustment models are the best models. The result confirms again that the model with adjusted family structure yields higher TPR whereas lower FPR and FDR (false discovery rate).

Analysis of the GWAS Data From GAW17

To further demonstrate the effectiveness of our USR, we compared it with competitors under an official simulation from the GAW17. This dataset contains real genotypes of 24,487 SNPs from 3,205 genes on 697 subjects, together with simulated phenotypes of these subjects. We chose replicate 1 of Q1 as outcomes and applied the algorithms to locate promising SNPs from genotype data. Both the weighted and unweighted versions of our USR detected five causal SNPs within two genes (*FLT1* and *KDR*). Three of the causal SNPs were rare variants but were missed by the single-marker test (Fig. 3, Table 4). Again, in this comparison, our USR inclined to discover rare casual variants with higher true positive than the single-marker test.

Analysis of the Sequence Data From GAW18

To illustrate effectiveness of our algorithm to locate rare genetic variants, we applied it to the analysis of Mexican Americans sequence data from the GAW18. This dataset contains next-generation sequencing data of 850 subjects within 21 large families.

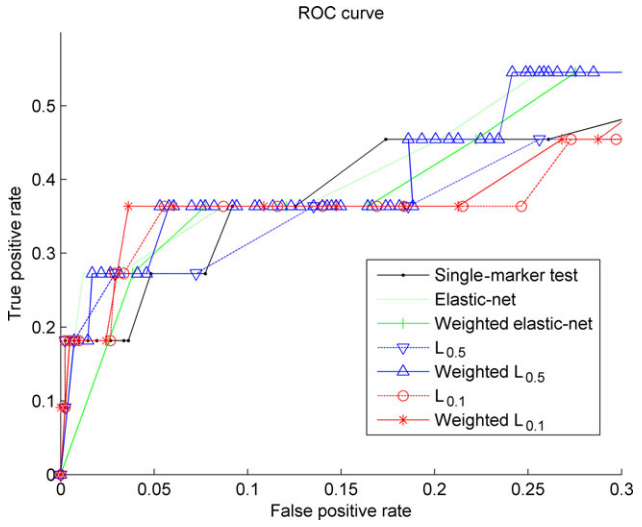

Figure 3. The partial ROC of chromosome 13 data. The weights are generated by the correlation coefficients between phenotypes and variants

Simulated Phenotype Analysis

First, we analyzed the simulated diastolic blood pressure (DBP), where DBP was set to be influenced by 1,243 variants of 245 genes and 1,040 variants of 205 genes, respectively. After quality control, we selected 504 subjects within the region of 1,244 variants from three genes (SLC35E2, TNN, and MAP4) that influenced the phenotypic data. All the data we used were from the first visit of the longitudinal data. We connected raw DBP data with covariates and population structure (adjusted by the first 10 principal components of the genotypic matrix) and pedigree structure by our generalized sparse regression model.

In this analysis, our USR algorithm appeared to have better TPR and better AUC compared to the algorithms without adjusting pedigree structure, while maintaining almost the same FPR level (Table 5). The pedigree adjustment appeared to be both necessary and beneficial as shown by this set of results.

Real Data Analysis

Finally, we applied the proposed USR to analyze real DNA sequence data on DBP and systolic blood pressure (SBP) from GAW18. After quality control, we obtained GWAS data of 783 Mexican Americans with 438,790 SNPs and next generation sequencing data of 506 Mexicans with 6,824,165 SNPs. When analyzing the GWAS data using our USR algorithm, we obtained sparse representation for each chromosome by choosing the entire chromosome as a window. However, for

Table 4. Identified casual rare variants for phenotype Q1

Causal gene and SNPs	Single-marker test	Elastic-net	Weighted elastic-net	$L_{0.5}$	Weighted $L_{0.5}$	$L_{0.1}$	Weighted $L_{0.1}$	MAF
KDR/C4S1874	×	√	√	√	√	√	√	0.00717
KDR/C4S1877	√	√	√	√	√	√	√	0.164993
KDR/C4S1884	×	×	√	√	√	√	√	0.0208
KDR/C4S1887	×	×	×	√	√	√	√	0.00717
FLT1/C13S523	√	√	√	√	√	√	√	0.066714
FLT1/C13S523	×	×	√	√	√	√	√	0.004304

Note: A "√" indicates that the corresponding marker was detected as a causal marker by a particular method. A "×" indicates that the corresponding marker was not detected as causal marker by a particular method.

Table 5. The error rate of GAW18 data

N = 504, SNPs = 1,243	TPR	FPR	Partial AUC	AUC
Elastic-net	0.1739	0.0704	0.1064	0.6436
Elastic-net family	0.2609	0.1605	0.1095	0.6786
Elastic-net family and weight	0.2174	0.0573	0.1247	0.6816
$L_{0.5}$	0.1739	0.0459	0.1131	0.6947
$L_{0.5}$ family	0.2174	0.0524	0.1170	0.7074
$L_{0.5}$ family and weight	0.2174	0.0745	0.1426	0.7125
$L_{0.1}$	0.2609	0.0983	0.1233	0.7008
$L_{0.1}$ family	0.2174	0.0524	0.1366	0.7117
$L_{0.1}$ family and weight	0.2609	0.0983	0.1417	0.7142
Gemma			0.1107	0.5638

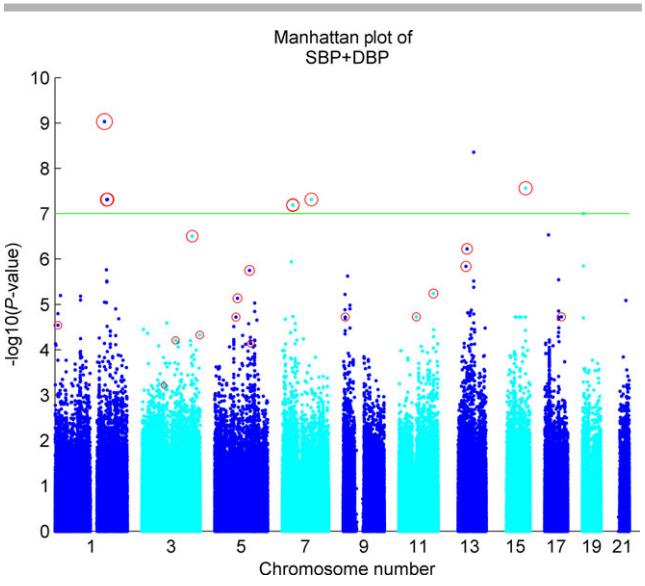

Figure 4. The Manhattan plot for SNPs on odd numbered chromosomes. The *P*-values were computed from single-marker tests. The red circles stand for the markers selected by our USR. We used SBP+DBP as the phenotype. The genome-wide nominal significance level was set to be 10^{-7}, as shown by the green horizontal line.

the sequence data set, it is too large to be analyzed as a whole window. Thus, we divided each of the large chromosomes (1, 3, 5, 7, 9) into two equal parts and obtained their sparse representations separately (Fig. 4).

Based on above algorithm, we analyzed GWAS and sequence data by our USR separately to find the susceptible genetic variants. Combining the significant variants selected by both GWAS and sequence data, we identified 23 promising genes (supplementary Table 3S). We also identified three significant pathways relevant to hypertension by pathway-wise SKAT [Wu et al., 2011]. The most significant pathway ($P = 3.24 \times 10^{-8}$) was *Glioma*, including *BRAF*, *SHC3*, *CAMK2B*, *EGFR*, and *PDGFRB*. An independent study [Houben et al., 2004] suggested that *Glioma* pathway would be associated with hypertension through potentially neurocarcinogenic effects of antihypertensive medication. The second most significant pathway ($P = 2.74 \times 10^{-7}$) is the regulation of *actin cytoskeleton* pathway, including *GNA12, BRAF, EGFR, PDGFRB* and *PIP5K1B*. This pathway was identified to be associated with hypertension by an independent study [Tripodi et al., 1996]. The third most significant pathway ($P = 3.87 \times 10^{-6}$) is *chronic myeloid leukemia* pathway, including *BRAF*, *RUNX1*, *SHC3*, and *MECOM*. This pathway, as suggested by independent studies [Dumitrescu et al., 2011; Guymer et al., 1993], would highly influence benign intracranial hypertension and pulmonary arterial hypertension.

Furthermore, we found some new candidate genes and pathways that were not reported in the previous independent study. For example, *FMO1* ($P = 9.81 \times 10^{-5}$) is a risk gene of cardiovascular disease [Mendelsohn and Larrick, 2013], which is usually associated with hypertension. Another susceptible gene is *AGBL1* ($P = 8.16 \times 10^{-4}$), which is associated with carotid plaque [Dong et al., 2012], and prehypertension is associated with significantly increased carotid atherosclerotic plaque [Hong et al., 2013]. We also report *long-term depression* pathway ($P = 4.21 \times 10^{-6}$) as a significant pathway. It might cause depression that is a risk factor of hypertension [Meng et al., 2012].

Conclusion

Many existent sparse regression algorithms assume unrelated subjects. Such algorithms fail to adjust for complex pedigrees and cryptic relatedness as often occur in the genomic data. In this article, we have proposed the USR algorithm for variant selection from DNA sequence data with an arbitrary intraindividual relationship and population structure. Our USR algorithm allows informative weighting to incorporate prior knowledge. This approach provides a flexible way to adjust for preference or risk variants. Extensive simulation results indicated that a properly predetermined weighting scheme can notably improve selection accuracy of causal variants.

Our algorithm can handle both rare and common variants simultaneously. The ability of our algorithm to pinpoint causal variants, especially rare causal variants, was clearly demonstrated by intensive simulations (see the details in supplementary material). We suggest using L_p norms ($0.1 < p < 0.5$) in the model because these regularization terms provide better performance in terms of AUC, TPR, and FPR. For the sake of computational speed, $L_{0.5}$ norm is a better choice. In particular, our algorithm can solve the low sample size but high dimensional feature problem, i.e., sample size is less than the number of variants, as often happens in genomic studies.

Like existent methods, our algorithm has some limitations. First, it focuses on a single variant effect on a trait of interest. A more powerful strategy would be to group multiple variants and incorporate group wide information into the model. Doing so, however, would scarify single-marker resolution. Second, our algorithm assumes linear relationship between phenotype and genotype, which may be unrealistic for many scenarios in practice. Extension to nonlinear regression models calls for additional efforts. Last, it deserves further investigation on how to choose the optimal tuning parameter and the optimal set of features.

Acknowledgments

We are grateful to Junbo Duan, Dongdong Lin, Jingyao Li, and Tao Yang for their useful discussions. We thank Weiwei Ouyang for providing Kinship matrix of the GAW18 data. Funding was provided by National Institutes of Health. YPW was partially supported by NIH grants (R01MH104680, R01GM109068). HWD was partially supported by NIH grants (P50AR055081, R01AG026564, R01AR059781, R01AR050496, and R01AR057049) and Edward G. Schlieder Endowment.

References

Candes EJ, Tao T. 2005. Decoding by linear programming. *IEEE Trans Inform Theory* 51(12):4203–4215.

Cetin MC, Erar A. 2002. Variable selection with Akaike information criteria: a comparative study. *Hacet J Math Stat* 31:89–97.

Chen X, Ishwaran H. 2012. Random forests for genomic data analysis. *Genomics* 99(6):323–329.

Chen X, Xu F, Ye Y. 2010. Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization. *SIAM J Sci Comp* 32(5):2832–2852.

Cho S, Kim K, Kim YJ, Lee JK, Cho YS, Lee JY, Han BG, Kim H, Ott J, Park T. 2010. Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann Hum Genet* 74(5):416–428.

Dong C, Beecham A, Wang L, Blanton SH, Rundek T, Sacco RL. 2012. Follow-up association study of linkage regions reveals multiple candidate genes for carotid plaque in Dominicans. *Atherosclerosis* 223(1):177–183.

Dumitrescu D, Seck C, ten Freyhaus H, Gerhardt F, Erdmann E, Rosenkranz S. 2011. Fully reversible pulmonary arterial hypertension associated with dasatinib treatment for chronic myeloid leukaemia. *Eur Respir J* 38(1):218–220.

Fan J, Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96(456):1348–1360.

Friedman J, Hastie T, Höfling H, Tibshirani R. 2007. Pathwise coordinate optimization. *Ann App Stat* 1(2):302–332.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1–22.

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. 2010. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet* 11(1):11–49.

Guymer R, Cairns J, O'Day J. 1993. Benign intracranial hypertension in chronic myeloid leukemia. *Aust NZ J Ophthalmol* 21(3):181–185.

Hong H, Wang H, Liao H. 2013. Prehypertension is associated with increased carotid atherosclerotic plaque in the community population of Southern China. *BMC Cardiovasc Disord* 13(1):13–20.

Houben M, Louwman W, Tijssen C, Teepen J, Van Duijn C, Coebergh J. 2004. Hypertension as a risk factor for glioma? Evidence from a population-based study of comorbidity in glioma patients. *Ann Oncol* 15(8):1256–1260.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178(3):1709–1723.

Lee S, Wu MC, Lin X. 2012. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13(4):762–775.

Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44(3):243–246.

Meinshausen N, Bühlmann P. 2010. Stability selection. *J R Stat Soc Ser B (Stat Method)* 72(4):417–473.

Mendelsohn AR, Larrick J. 2013. Dietary modification of the microbiome affects risk for cardiovascular disease. *Rejuvenation Res* 16(3):241–244.

Meng L, Chen D, Yang Y, Zheng Y, Hui R. 2012. Depression increases the risk of hypertension incidence: a meta-analysis of prospective cohort studies. *J Hypertens* 30(5):842–851.

Natarajan BK. 1995. Sparse approximate solutions to linear systems. *SIAM J Comput* 24(2):227–234.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 11(7): 459–463.

Qin H, Morris N, Kang SJ, Li M, Tayo B, Lyon H, Hirschhorn J, Cooper RS, Zhu X. 2010. Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics* 26(23):2961–2968.

Rao BD, Kreutz-Delgado K. 1999. An affine scaling methodology for best basis selection. *IEEE Trans Signal Process* 47(1):187–200.

Thompson E, Shaw R. 1990. Pedigree analysis for quantitative traits: variance components without matrix inversion. *Biometrics* 46:399–413.

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. 2012. Estimating kinship in admixed populations. *Am J Hum Genet* 91(1):122–138.

Tripodi G, Valtorta F, Torielli L, Chieregatti E, Salardi S, Trusolino L, Menegon A, Ferrari P, Marchisio P-C, Bianchi G. 1996. Hypertension-associated point mutations in the adducin alpha and beta subunits affect actin cytoskeleton and ion transport. *J Clin Invest* 97(12):2815–2822.

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25(6): 714–721.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82–93.

Xu Z-B, Guo H-L, Wang Y, Zhang H. 2012a. Representative of L1/2 regularization among Lq Lq Lq ($0 < q \leq 1$) regularizations: an experimental study based on phase diagram. *Acta Autom Sin* 38(7):1225–1228.

Xu Z, Chang X, Xu F, Zhang H. 2012b. L1/2regularization: a thresholding representation theory and a fast solver. *IEEE Trans Neural Netw Learn Syst* 23(7):1013–1027.

Yi N, Liu N, Zhi D, Li J. 2011. Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. *PLoS Genet* 7(12):e1002382.

Zhang C, Chen X. 2009. Smoothing projected gradient method and its application to stochastic linear complementarity problems. *SIAM J Optim* 20(2):627–649.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44(7):821–824.

Zhou H, Sehl ME, Sinsheimer JS, Lange K. 2010. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26(19):2375–2382.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Stat Method)* 67(2):301–320.