Bioinformatics, 36(12), 2020, 3811–3817
doi: 10.1093/bioinformatics/btaa229
Advance Access Publication Date: 4 April 2020
Original Paper

Genetics and population analysis

Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning

Songshan Yang[1], Jiawei Wen[1], Scott T. Eckert[2], Yaqun Wang[3], Dajiang J. Liu[2], Rongling Wu[2], Runze Li[1] and Xiang Zhan[2,*]

[1]Department of Statistics, Pennsylvania State University, University Park, PA 16802, [2]Department of Public Health Sciences, Pennsylvania State University, Hershey, PA 17033 and [3]Department of Biostatistics, Rutgers University, New Brunswick, NJ 08901, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on September 30, 2019; revised on February 19, 2020; editorial decision on March 27, 2020; accepted on March 31, 2020

Abstract

Motivation: Large scale genome-wide association studies (GWAS) have resulted in the identification of a wide range of genetic variants related to a host of complex traits and disorders. Despite their success, the individual single-nucleotide polymorphism (SNP) analysis approach adopted in most current GWAS can be limited in that it is usually biologically simple to elucidate a comprehensive genetic architecture of phenotypes and statistically underpowered due to heavy multiple-testing correction burden. On the other hand, multiple-SNP analyses (e.g. gene-based or region-based SNP-set analysis) are usually more powerful to examine the joint effects of a set of SNPs on the phenotype of interest. However, current multiple-SNP approaches can only draw an overall conclusion at the SNP-set level and does not directly inform which SNPs in the SNP-set are driving the overall genotype–phenotype association.

Results: In this article, we propose a new permutation-assisted tuning procedure in lasso (plasso) to identify phenotype-associated SNPs in a joint multiple-SNP regression model in GWAS. The tuning parameter of lasso determines the amount of shrinkage and is essential to the performance of variable selection. In the proposed plasso procedure, we first generate permutations as pseudo-SNPs that are not associated with the phenotype. Then, the lasso tuning parameter is delicately chosen to separate true signal SNPs and non-informative pseudo-SNPs. We illustrate plasso using simulations to demonstrate its superior performance over existing methods, and application of plasso to a real GWAS dataset gains new additional insights into the genetic control of complex traits.

Availability and implementation: R codes to implement the proposed methodology is available at https://github.com/xyz5074/plasso.

Contact: xyz5074@psu.edu

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

1 Introduction

Over the last decade, genome-wide association studies (GWAS) have been extensively used to dissect the genetic architecture of many complex diseases and have identified thousands of common genetic variants, typically single-nucleotide polymorphisms (SNPs) that are associated with complex traits (Hardy and Singleton, 2009; MacArthur *et al.*, 2017; McCarthy *et al.*, 2008; Visscher *et al.*, 2017). Most current GWAS are primarily based on the paradigm of 'individual-SNP-based association analysis', which tests for marginal association between each individual SNP and a complex trait. The results are typically summarized in a Manhattan plot of all individual *P*-values (one for each SNP), and then SNPs with *P*-values less than a certain threshold (e.g. 5×10^{-8}, the multiple testing adjusted genome-wide significance level) are identified and investigated in further downstream studies, such as replication or validation studies and laboratory-based functional studies (Schaid *et al.*, 2018).

Despite being useful in identifying many disease-susceptibility genetic variants, more and more evidence has shown the limitations of the individual-SNP approach adopted in most current GWAS. First, continuing advances in high-throughput next-generation sequencing technologies have led to an explosion of novel genomics assays and platforms. As a consequence, a large number of SNPs (in millions) have been genotyped and tested in GWAS, and due to the heavy multiple-testing burden, threshold for genome-wide significance level used in individual-SNP approach can be difficult to attain, leading to few discoveries. This is especially true to those GWAS with fairly moderate effect size and relatively small sample size. Second, the genetic architecture of complex traits is often polygenic and complicated

in that multiple SNPs act in concert to accomplish tasks related to phenotypic changes, such as onset of a disease. Such joint effects or interaction/epistatic effects among multiple SNPs, however, are impossible to be captured using individual-SNP analysis. Due to these potential limitations of individual-SNP analysis, many authors proposed alternative, multiple-SNP (or rare variants) analysis strategies, such as burden tests (Li and Leal, 2008; Madsen and Browning, 2009) and variance component tests (Basu *et al.*, 2011; Ionita-Laza *et al.*, 2013; Lee *et al.*, 2014; Wu *et al.*, 2010, 2011; Zhan *et al.*, 2016, 2017) to examine the association between the trait and a group of genetic variants. The grouping is typically based on proximity of genomic location of these variants. These multiple-SNP approaches usually can have improved power over traditional individual-SNP analysis by reducing multiple-testing burden and by enabling capture of joint effects or epistatic effects of multiple SNPs (Wu *et al.*, 2010). Consequently, the multiple-SNP approaches have achieved a great success in identifying associations between genetic variants and complex traits (Auer *et al.*, 2014; Neale *et al.*, 2012).

A primary limitation of multiple-SNP association analysis is the interpretation of a significant result, since multiple-SNP approaches (both burden tests and variance component tests) model the cumulative effects of all SNPs in the set rather than the effect of each individual SNP. In other words, it only draws a global conclusion about existence of the overall association at the entire region level but does not directly inform which SNPs in the region are actually driving the overall genotype–phenotype association. This pitfall of multivariate-SNP approaches is a huge hurdle for downstream laboratory-based functional studies, and poses great challenges for making the leap from genetic association survey to rational genomics and genetics-based therapeutics (Schaid *et al.*, 2018).

One way to address the aforementioned limitation of the multiple-SNP approach is to apply variable selection methods to identify SNPs that are associated with the phenotype of interest. A wide range of variable selection procedures have been proposed and applied to the context of GWAS, most of which are built on the lasso-type penalized regression strategy (Arbet *et al.*, 2017; Ayers and Cordell, 2010; Basu *et al.*, 2011; Carbonetto and Stephens, 2012; Cho *et al.*, 2010; He *et al.*, 2016; Li *et al.*, 2011; Waldmann *et al.*, 2013; Wu *et al.*, 2009; Yi *et al.*, 2015). The lasso method has proven to be a versatile variable selection tool with appealing estimation and prediction properties (Tibshirani, 1996). However, what has been less studied in existing lasso-type methods in GWAS is the selection of the lasso tuning parameter, which determines the amount of shrinkage and can have a substantial influence on the selection performance of the method. Recently, it has been more and more popular to use pseudo-variables (e.g. permutation copies, knockoff copies) to facilitate variable selection (Barber and Candès, 2015; Candès *et al.*, 2018; Luo *et al.*, 2006; Wu *et al.*, 2007). In these procedures, selection of each pseudo-variable will be flagged as a false positive, and therefore it can reduce potential false positives of variable selection by using pseudo-variables. Motivated by this, we propose a new method that introduces pseudo-variables to assist the lasso tuning parameter selection aiming to reduce the false discoveries, which may cause costly but fruitless follow-up laboratory studies. Once the tuning parameter is selected by our procedure, the L_1 regularization problem can be easily solved to obtain the lasso estimator, and thus, SNPs with non-zero coefficients are selected. Finally, by discovering which SNPs are important, geneticists can design a more targeted follow-up investigation to study how these SNPs influence the phenotype of interest.

2 Materials and methods

2.1 Data and model

Phenotype and genotype $\{y_i, \mathbf{x}_i\}_{i=1,\ldots,n}$ from n subjects are measured in a GWAS, where $\mathbf{y} = (y_1, \ldots, y_n)$ is the phenotype of interest, such as height, blood pressure or a disease status, which can be either quantitative or dichotomous. $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T = (X_1, \ldots, X_p)$ is a $n \times p$ matrix with both genetic variants (SNP markers or rare variants) and non-genetic covariates. Examples of such covariates include clinical and demographic variables such as age, gender, race and top principal components of genetic relatedness (adjustment for population stratification). The number of covariates collected in a typical genetic association study is much smaller than the number of genetic markers, and by adjusting for these covariates, the genetic effects of SNP markers on the phenotype can be estimated more precisely. For simplicity, we use the term SNP markers for all these covariates without distinguishing the two from a methodology point of view. Finally, without loss of generality, we assume an additive model and code the SNP variables as 0/1/2 denoting the copies of minor allele throughout this article. The method developed in this article can be easily extended to other genetic models (e.g. dominant and recessive).

Despite the complicated genetic architecture of most complex traits, statistical geneticists agree in the usefulness of a multivariate (generalized) linear model to capture at least a preliminary approximation of the nature of the relation between genotypes and phenotypes (Wu *et al.*, 2010, 2011). Alternative to the individual-SNP analysis, we jointly consider multiple SNPs in the regression model to capture potential joint effects on phenotype and also to enhance both statistical power and biological interpretability. Specifically, for quantitative phenotypes, we use the linear model

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \tag{1}$$

and for dichotomous phenotypes, we model the conditional mean of phenotype y on genotypes $\mathbf{X}$ using logistic regression

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \tag{2}$$

where $\pi_i = Pr(y_i = 1)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is $p \times 1$ vector of regression coefficients. For any X_j with a non-zero coefficient β_j in either (1) or (2), we call it an active SNP. Yet, many geneticists also prefer the term causal SNPs or functional SNPs, and we will use these terms exchangeably throughout this article. The focus of most previous multiple-SNP association analysis is the global null hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$ to examine the overall effects of p SNPs on the phenotype. Once the global hypothesis is rejected, it is of extreme interest to identify SNPs that drive the overall association. To facilitate downstream research studies of these functional SNPs, we propose a new method in this article to identify which $\beta_j \neq 0$. In other words, the focus of the current article is to select important SNPs for the phenotype in a joint regression model (1) or (2).

The advancement of high-throughput techniques now allows the measurement of millions of SNPs selected across the entire genome. Despite the availability of massive SNP data, it is often the case that only a handful of these SNPs are expected to be relevant to the trait of interest. In other words, we assume that the number of important SNPs that influence the trait is sparse. This is the ideal scenario to apply the lasso-type methods (Tibshirani, 1996) due to its property of shrinking some of the model coefficients to exactly zero and thus identify SNPs with notable effects on the phenotype as those with non-zero coefficients. For quantitative phenotypes, the lasso-penalized least squares objective function is

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{3}$$

For dichotomous phenotypes, the objective function is

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\sum_{i=1}^{n}[y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)] + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{4}$$

where

$$\pi_i = \frac{e^{\sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\sum_{j=1}^{p} x_{ij}\beta_j}}.$$

In both penalized least squares (3) and (4), $\lambda > 0$ is a regularization parameter which controls the amount of penalization, and a larger λ would shrink more coefficients to zero. Once the tuning parameter λ is fixed, objective function (3) or (4) can be efficiently solved (Friedman *et al.*, 2010), and thus SNPs with non-zero β-coefficients are identified as active ones that affect the phenotype.

2.2 A new permutation-assisted lasso tuning procedure

The performance of lasso heavily hinges on an appropriate selection of the tuning parameter (Chand, 2012; Fan and Tang, 2013). For the purpose of this article, we refer to a tuning procedure as a method of selecting the lasso penalty parameter λ. As a common practice among statistical geneticists (Ayers and Cordell, 2010; Basu *et al.*, 2011; Cho *et al.*, 2010; Waldmann *et al.*, 2013), most existing tuning procedures used in GWAS focus on either maximizing the posterior model probability [e.g. Bayesian information criterion (BIC)] or minimizing the estimated prediction error via cross-validation (CV). Also, it is appealing to have a tuning procedure which directly focuses on variables rather than a surrogate model-based criterion such as prediction error or posterior model probability.

Recently, it has been popular to use pseudo-variables to assist variable selection (Barber and Candès, 2015; Candès *et al.*, 2018; Luo *et al.*, 2006; Srinivasan *et al.*, 2019; Wu *et al.*, 2007; Yang *et al.*, 2019). These methods differ in the type of pseudo-variables being added to the original regression problem, yet they are common in that pseudo-variables are randomly generated to be independent of the response variable. Then, the basic idea of using these pseudo-variables for tuning variable selection is to avoid undesirable procedures that select too many pseudo-variables, as they are constructed to be inactive (independent of the response). Examples of such pseudo-variables include standard normal noises, permutations of original variables (Luo *et al.*, 2006; Wu *et al.*, 2007; Yang *et al.*, 2019) and knockoff copies (Barber and Candès, 2015; Candès *et al.*, 2018; Srinivasan *et al.*, 2019). On the one hand, the computational cost of constructing knockoff copies for SNPs in GWAS is huge considering the size of massive SNP data; and simple multivariate independent and identically normal distributed white noise largely ignore the complex correlation [or linkage disequilibrium (LD)] structure among different SNPs, and are often less powerful for genetic association studies (Wu *et al.*, 2010, 2011). On the other hand, permutations of original variables preserve the correlation structure of original variables and hence are used to facilitate the lasso tuning parameter selection in the plasso method described in the following.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T = (X_1, \ldots, X_p)$ be the original $n \times p$ SNP matrix. For ease of presentation, we refer to both non-genetic covariates and genetic markers as SNPs in this method description section. The pseudo-design matrix is constructed as $\mathbf{X}^\pi \equiv (x_{\pi(1)}, \ldots, x_{\pi(n)})^T$, where $\{\pi(1), \ldots, \pi(n)\}$ is a permutation of $\{1, \ldots, n\}$. Now, we define the augmented design matrix $\mathbf{X}^A = [\mathbf{X}, \mathbf{X}^\pi] = [X_1^A, \ldots, X_p^A, X_{p+1}^A, \ldots, X_{2p}^A]$, where $[X_1^A, \ldots, X_p^A] \equiv [X_1, \ldots, X_p]$ is the original design matrix and $[X_{p+1}^A, \ldots, X_{2p}^A]$ is the permuted design matrix (of pseudo-variables). After augmentation, the corresponding lasso regression problem [taking mode (3) as an example] is

$$\boldsymbol{\beta}^{\mathrm{A}}(\lambda) = \underset{\boldsymbol{\beta}^{\mathrm{A}}}{\operatorname{argmin}} \frac{1}{2n} ||\mathbf{y} - \mathbf{X}^A \boldsymbol{\beta}^{\mathrm{A}}||_2^2 + \lambda \sum_{j=1}^{2p} |\beta_j^A|, \tag{5}$$

where $\boldsymbol{\beta}^{\mathrm{A}} \equiv (\beta_1^A, \ldots, \beta_{2p}^A)^T$ is the regression coefficients for both the p original SNPs and p pseudo-SNP variables. Once the lasso tuning parameter λ is given, the estimated coefficients $\hat{\boldsymbol{\beta}}^A$ can be easily obtained (Friedman *et al.*, 2010; Wu *et al.*, 2009).

We use notation $\boldsymbol{\beta}^{\mathrm{A}}(\lambda)$ to denote the lasso solution path over tuning parameter λ. For a given value of tuning parameter λ, a certain number of variables with non-zero regression coefficients can be selected. Although minor exceptions occasionally occur, for a typical lasso solution path $\boldsymbol{\beta}^{\mathrm{A}}(\lambda)$, more variables can enter the model when λ decreases. Once a variable enters the model, it usually remains in the model as λ decreases (Wu *et al.*, 2009). Motivated by this, for each SNP variable in the augmented design matrix, one can find the maximum of tuning parameter λ such that the SNP variable remains in the model. That is, we consider the point λ on the lasso path at which variable j remains in (or first enters in terms of decreasing λ's) the model

$$W_j = \sup\left\{\lambda : \hat{\beta}_j^A(\lambda) \neq 0\right\}, \quad j = 1, \ldots, 2p. \tag{6}$$

Recall that the lasso tuning parameter $\lambda > 0$. We set $W_j = 0$ if a variable does not enter the model even without the lasso penalty [i.e. $\hat{\beta}_j^A(0) = 0$]. This new statistic W_j can be viewed as an importance metric for variable j, as an active variable tends to remain longer in the model as the penalty λ increases compared to an inactive variable.

Since these pseudo-variables are known to be inactive (as they are constructed without utilizing the phenotypic information), a preferred tuning procedure should be able to rule out those λ parameters that identify a pseudo-variable as active. Motivated by this, we define $C_\pi = \max_{(p+1) \leq j \leq 2p} W_j$ as a benchmark to separate active variables from inactive ones. That leads to the following variable selection procedure

$$\hat{S}_\pi = \{j : W_j > C_\pi, j = 1, \ldots, p\}. \tag{7}$$

In other words, we select those original variables which have importance metric W_j greater than C_π, the maximum of importance metrics of all p pseudo-variables. Here, $\hat{S}_\pi$ denotes the estimator of true active variables set S under a particular permutations π. As the selection results, $\hat{S}_\pi$ might be affected by permutations (occurs occasionally in finite samples). One straightforward way to stabilize the selection is to use multiple permutations. Let $\pi_1, \ldots, \pi_B$ be B permutations with selected variable set $\hat{S}_{\pi_1}, \ldots, \hat{S}_{\pi_B}$, respectively. For each SNP variable, we can define its selection stability (SS) as $\mathrm{SS}_j = \sum_{k=1}^B I[X_j \in \hat{S}_k]/B$, for $j = 1, \ldots, p$, where $I[\cdot]$ is the indicator function. Then, we can select SNPs with a high SS (e.g. SNPs with $\mathrm{SS} \geq 90\%$). Simulations have been conducted to evaluate how parameters (i.e. number of permutations and selection frequency) involved in the SS procedure would affect the performance of the proposed procedure and results are reported in Supplementary Material Section S2.2. Finally, it is possible that other more complicated approaches using either bootstrap or subsampling techniques are also feasible to achieve stability selection (Cho *et al.*, 2010; Meinshausen and Bühlmann, 2010).

2.3 Comparison with other penalized regression methods in GWAS

A number of penalized regression methods have been proposed and applied to GWAS to perform fine-mapping (Schaid *et al.*, 2018). These methods, in general, belong to two main categories. One is to perform variable selection without post-selection inference (Cho *et al.*, 2010; Li *et al.*, 2011; Waldmann *et al.*, 2013; Wu *et al.*, 2009). The other performs both variable selection and post-selection inference (Arbet *et al.*, 2017; Ayers and Cordell, 2010; Yi *et al.*, 2015). The additional inference [e.g. type I error rate or false discovery rate (FDR) control] comes along with additional assumptions, which may or may not hold in reality. For example, the tuning procedure proposed in Ayers and Cordell (2010) assumes $p_0/p \approx 1$, where p_0 is the number of null SNPs (or inactive SNPs) and p is the number of total SNPs being modeled. When p is relatively small such that p_1/p cannot be ignored (where p_1 is the number of active SNPs), it has been observed that the proposed tuning procedure in Ayers and Cordell (2010) can have inflated type I error rate (Arbet *et al.*, 2017; Ayers and Cordell, 2010). The proposed plasso method belongs to the first category of GWAS penalized regression methods that does not consider rigorous post-selection inference. Despite not providing a measure of error for the selected SNPs, the rationale behind the methodology of plasso along with the SS consideration implemented in plasso provide a means to guarantee the statistical relevance of each selected SNP to the phenotype, which can also be

of interest to many geneticists who would not accept only a single *P*-value (or type I error) as definitive.

Compared to other similar GWAS penalized regression methods without inference (Cho *et al.*, 2010; Li *et al.*, 2011; Waldmann *et al.*, 2013; Wu *et al.*, 2009), one common thing among plasso and others is the computational scalability, which is largely determined by the underlying lasso method. Like the other methods (Cho *et al.*, 2010; Li *et al.*, 2011; Waldmann *et al.*, 2013; Wu *et al.*, 2009), plasso can typically handle tens of thousands of SNPs at a reasonable computational cost (see Supplementary Material Section S2.4). For the genome-wide SNPs data (e.g. millions of SNPs), preprocessing of data for dimension reduction is required. Examples of such data preprocessing include pre-screening based on GWAS summary statistics (e.g. marginal *P*-values) (Cho *et al.*, 2010; Waldmann *et al.*, 2013) and preconditioning (Li *et al.*, 2011). Finally, plasso has two main advantages over existing methods of its kind. First, plasso selects the lasso tuning parameter in a data-adaptive manner that preserves the LD structure among SNPs when generating permutations. On the other hand, existing methods (i) either select the tuning parameter in a rather ad hoc pattern [e.g. specifying the number of SNPs to be included in the model (Wu *et al.*, 2009) or assigning an appropriate hyperprior distribution on the tuning parameter (Li *et al.*, 2011)], which is difficult to generalize to real data as the underlying knowledge is never known in reality; (ii) or use CV (Cho *et al.*, 2010; Waldmann *et al.*, 2013), which could either have unstable results (fold number is small) or be computational expensive (fold number is large, e.g. leave-one-out CV). Second, plasso is more flexible in terms of accommodating different types of phenotype variables such as quantitative, dichotomous, counts and categorical variables with more than two levels. This is because plasso only depends on the solution path of the augmented lasso regression problem, which is calculated via the cyclic coordinate descent algorithm implemented in R package *glmnet* (Friedman *et al.*, 2010). And thus, plasso can handle many different types of phenotypes that are allowed in the *glmnet* package. On the other hand, most existing methods are typically developed under specific parametric models, such as linear regression (Cho *et al.*, 2010; Li *et al.*, 2011; Yi *et al.*, 2015) for quantitative traits or logistic regression (Ayers and Cordell, 2010; Basu *et al.*, 2011; Wu *et al.*, 2009) for dichotomous traits.

2.4 Design of simulations

Comprehensive simulation studies have been conducted to evaluate and compare the finite sample performance of plasso (denoted by P in numerical studies sections of this article) with other lasso λ-parameter tuning procedures. It has been popular among statistical geneticists to use CV for lasso tuning parameter value selection (Basu *et al.*, 2011; Cho *et al.*, 2010; Waldmann *et al.*, 2013; Yi *et al.*, 2015), and model-selection criterion, such as BIC, has been also widely used among statisticians for penalty parameter tuning in high-dimensional model (Fan and Tang, 2013). Due to their popularity, we compared plasso to CV (10-fold) and BIC in this simulation. For ease of presentation and results interpretation, we only simulated 0/1/2 coded SNP variables while ignoring other types of non-genetic covariates when simulated data matrix **X**. Two different sets of simulations were conducted to evaluate the SNP selection performance of P, CV and BIC: one for a quantitative phenotype (Simulation I) and the other for a dichotomous phenotype (Simulation II). Additional simulation studies to further evaluate the performance of different methods are also available in Supplementary Material Section S2.

2.4.1 Simulation I

We followed the simulation design of a previous lasso-type method for GWAS paper (Wu *et al.*, 2009) to generate the p SNP variables from a latent multivariate Gaussian distribution. One advantage of using this multivariate Gaussian distribution is its ability to accommodate any general correlation structure Σ as described below. We first simulated n independent p-dimensional vector $(L_1, \ldots L_p)$ from multivariate normal distribution $(\text{MVN}(0, \Sigma))$, where Σ had the same structure as the one used previously (Wu *et al.*, 2009). That is, $\Sigma_{jj} = 1$, $\Sigma_{jk} = \rho$ if $j \neq k, j, k \leq p/20$, and $\Sigma_{jk} = 0$ otherwise. This correlation structure allows causal SNPs to be correlated with neighboring SNPs (as long as the index of causal SNP $j \in \{1, \ldots, p/20\}$). We considered different ρ values of 0, 0.4, 0.8 representing different levels of correlations among SNPs in this simulation. Next, we randomly generated p minor allele frequencies (MAFs) $m_1, \ldots, m_p$ from uniform distribution Unif (0.05,0.5). For the jth variable L_j, let c_1 and c_2 be the $(1-m_j)^2$-quantile and $(1-m_j^2)$-quantile of $\{L_{1j}, \ldots, L_{nj}\}$, respectively. Then, we set the genotype of the jth SNP for the ith individual X_{ij} equal to 0, 1 or 2 according to whether $L_{ij} \leq c_1$, $c_1 < L_{ij} < c_2$ or $L_{ij} \geq c_2$, respectively. After the p SNP variables were generated, we randomly picked $p_1 = 20$ of them as active ones that truly affect the phenotype and denoted the set of active SNPs as S. We standardized each SNP to have a mean of zero and a variance of one and then simulated the phenotype using the following quantitative phenotype model:

$$y_i = 1 + \sum_{j \in S} \beta_j x_{ij} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$. For each $j \in S$, we randomly set $\beta_j = 0.2$ or -0.2 with even probability, corresponding to a heritability (variance explained by the causal SNP over total variance) around 2.2% (ignoring correlation among SNPs, heritability $h_j^2 = \beta_j^2 / (\sum_{j \in S} \beta_j^2 + \sigma_\epsilon^2) = 0.04/1.8$) and a total heritability of 44%. We considered different GWAS with samples sizes n=1000 or 2500 and p=1000 or 5000 SNPs, and generated 1000 replications of genotype–phenotype $(y, \mathbf{X})$ datasets to compare different methods.

2.4.2 Simulation II

Compared to Simulation I, the procedure of generating SNPs was slightly modified in Simulation II to achieve case-control sampling. The number of cases and controls was set to be equal (i.e. $n_0 = n_1 = n/2$), and the phenotype y was generated according to the dichotomous phenotype model:

$$\log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \sum_{j \in S} \beta_j x_{ij},$$

where $y_i \sim \text{Binomial}(1, \pi_i)$ and $\beta_0 = \log(0.05/0.95)$ was determined to set the prevalence to 5% in the dichotomous phenotype model. We randomly set $\beta_j = \pm 1$ with even probability for $j \in S$ to mimic the scenario considered in a previous study (Wu *et al.*, 2009).

2.5 Real-world data

Following the Bayesian lasso paper (Li *et al.*, 2011), we applied the proposed plasso method to the GWAS data collected in the Framingham Heart Study, a long-term, ongoing cardiovascular cohort study on residents of the city of Framingham, Massachusetts sponsored by the National Heart, Lung, and Blood Institute (Jaquish, 2007; Mahmood *et al.*, 2014). Among the 550 000 genotyped variants in a cohort of 977 subjects (418 males and 559 females), about 350 000 SNPs with MAF greater than 0.1 were kept for further analysis. The phenotype of interest in our analysis is the body mass index (BMI). Besides genotype and phenotype, age and gender of each subject have been measured and will be adjusted when studying the association between BMI and SNPs. More details about this dataset is available in previous publications (Jaquish, 2007; Li *et al.*, 2011).

3 Results

3.1 Simulation studies

We applied the three lasso tuning procedures P, CV and BIC to each simulated genotype–phenotype dataset to calculate the selected set of causal SNPs $\hat{S}$. For the new procedure *P*, we determined the final selection set $\hat{S}$ by using 10 permutations and picked SNPs that were at least selected 9 times out of 10 (i.e. SS $\geq$ 90%). We evaluated the performance of P, CV, BIC by comparing the selected set $\hat{S}$ to

the true active SNPs-set S. In particular, the following two metrics were calculated to compare $\hat{S}$ and S: (i) precision $\mathcal{P} = \{S \cap \hat{S}\}/\hat{S}$: number of active variables selected/number of variables selected and (ii) recall $\mathcal{R} = \{S \cap \hat{S}\}/S$: number of active variables selected/number of active variables. Precision and recall describe different aspects of a procedure and are closely related to other metrics (i.e. precision equals 1−FDR and recall is true positive rate) used in previous studies (Arbet *et al.*, 2017; Auer *et al.*, 2014; Yi *et al.*, 2015). An ideal procedure is expected to have both a high precision rate and a high recall rate. However, there is usually an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other (similar to the relation between Type I error and Type II error in hypothesis testing). To balance these two conflicting metrics, it is often meaningful to combine them into a single measure, such as the F-score $\mathcal{F} = 2\times(1/\mathcal{P} + 1/\mathcal{R})^{-1}$, which is the harmonic mean of precision and recall. For each tuning procedure, we calculated its precision $\mathcal{P}$, recall $\mathcal{R}$ and F-score $\mathcal{F}$ over the 1000 replicates.

The precision and recall of P, CV and BIC under the quantitative phenotype model are reported in Table 1. the precision of P is much higher than that of BIC, which in turn is much higher than that of CV (all differences are highly significant due to the small standard errors). For both P and BIC, it is more difficult to select the causal SNPs as the correlation (ρ) among SNPs are increasing. On the other hand, CV tends to be more robust to the correlation, yet it has the worst performance of the three in terms of precision. This is probably because CV is choosing the λ that provides the best predictive accuracy and tends to favor those λ's with less sparse solutions (as adding more predictors into the model would not hurt the predicting performance). On average, only about 20% of signals selected by the CV method are true signals, while the majority of the selected set by CV are false positives. This large amount of false discoveries detected by CV and BIC could probably lead to costly and fruitlessly downstream validation and functional studies.

Table 1. Performance of three lasso tuning procedures under the continuous phenotype model

		$\rho=0$		$\rho=0.4$		$\rho=0.8$	
(n, p)	Method	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$
I	P	0.992	0.916	0.986	0.910	0.966	0.913
		(0.001)	(0.003)	(0.001)	(0.003)	(0.002)	(0.003)
	CV	0.222	1	0.226	1	0.219	0.997
		(0.002)	(0)	(0.002)	(0)	(0.002)	(0)
	BIC	0.708	0.999	0.693	0.998	0.681	0.991
		(0.003)	(0)	(0.003)	(0)	(0.003)	(0.001)
II	P	0.990	0.803	0.982	0.807	0.943	0.793
		(0.001)	(0.004)	(0.001)	(0.004)	(0.003)	(0.004)
	CV	0.160	1	0.157	1	0.158	0.991
		(0.001)	(0)	(0.002)	(0)	(0.001)	(0.001)
	BIC	0.670	0.993	0.653	0.991	0.641	0.979
		(0.003)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)
III	P	0.995	1	0.989	1	0.971	0.992
		(0)	(0)	(0.001)	(0)	(0.002)	(0.001)
	CV	0.228	1	0.231	1	0.227	1
		(0.002)	(0)	(0.002)	(0)	(0.002)	(0)
	BIC	0.769	1	0.756	1	0.741	0.998
		(0.003)	(0)	(0.003)	(0)	(0.003)	(0)
IV	P	0.993	1	0.983	1	0.946	0.991
		(0.001)	(0)	(0.001)	(0)	(0.002)	(0.001)
	CV	0.176	1	0.170	1	0.171	1
		(0.002)	(0)	(0.002)	(0)	(0.002)	(0)
	BIC	0.755	1	0.741	1	0.712	0.997
		(0.003)	(0)	(0.003)	(0)	(0.003)	(0)

Note: Numbers listed without parentheses are average values over 1000 replicates and numbers listed within parentheses are the corresponding standard errors. (n, p)-scenario I, II, III, IV corresponds to (1000, 1000), (1000, 5000), (2500, 1000), (2500, 5000), respectively.

In practice, a higher precision typically comes along with a lower recall. As can be seen from Table 1, the recall of P is slightly lower than that of BIC and CV when n=1000. This loss in recall of the proposed P method is small compared to its gain in precision. When the sample size is large (n=2500, i.e. scenario III and IV), the recall of all three methods are close to one. That is, all methods can asymptotically recover the true model. To better evaluate the overall performance of three lasso tuning procedures, we report the combined F-score in Figure 1, where one can see that our proposed P method has the best overall performance under each scenario. The differences observed in Figure 1 are again highly significant after considering their standard errors. Similar results are observed under the dichotomous phenotype model (see Supplementary Table S1 and Fig. S1).

To summarize, on the one hand, the proposed plasso method consistently has a much higher precision rate than both CV-based lasso and BIC-based lasso. The introduced pseudo-variables in plasso can facilitate variable selection by reducing false positive rate to a large extent. On the other hand, these noise pseudo-variables can also dilute the association signal between genotype and phenotype, making it more difficult to recover all true signals. That is, plasso tends to have a slightly smaller recall rate than CV and BIC as observed in Table 1. Overall, plasso usually has a much better variable selection performance than CV and BIC in terms of F-score. In practice, the gain in precision rate of plasso in GWAS discovery stage can largely facilitate follow-up laboratory studies by avoiding costly and fruitlessly downstream validation and functional studies on false positive SNPs.

3.2 Real data analysis

Following the preprocessing procedures used in the Bayesian lasso paper (Li *et al.*, 2011), we focus our analysis on 1837 SNPs which has a single SNP analysis (marginal) P-value less than $10^{-3.5}$ in at least one gender. For each phenotype, we first fit a linear model with age and gender as covariates, and then the conditional residuals from this fit were used as the new outcome for all subsequent genetic association testing. Besides plasso, both CV-based and BIC-based lasso were applied to test the association between age and gender-adjusted BMI and 1837 SNPs. To enhance reproducibility of selecting results of plasso, only SNPs with SS no less than 0.9 (9 out of 10 times) were selected by plasso. For fair comparison, we also implemented the CV-based lasso 10 times and selected those SNPs that

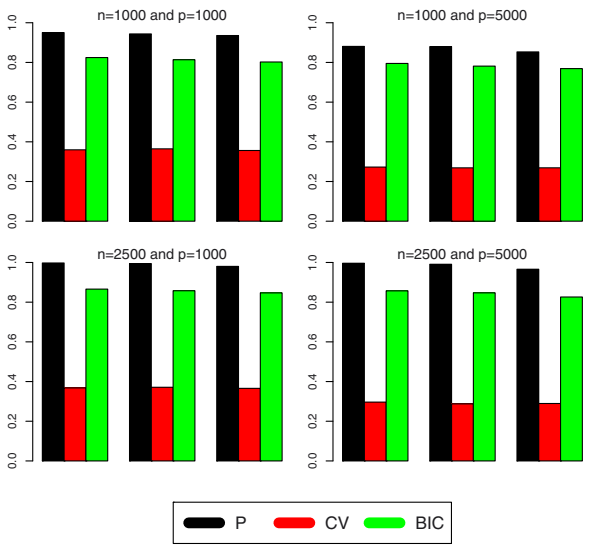

Fig. 1. F-score of P, CV and BIC-based lasso under the continuous phenotype model. The y-axis represents the F-score and the x-axis represents the correlation levels corresponding to $\rho = 0, 0.4, 0.8$, respectively

Table 2. List of SNPs selected by plasso

Name	Chromosome	Reference allele	Effect
Ss66227971	3	G	−0.632
ss66047592	5	T	1.0237
ss66045603	7	T	1.1089
ss66429034	10	T	0.6834
ss66383647	11	C	0.7024
ss66366940	20	A	0.9234

Note: The effect column corresponds to the regression coefficient of the reduced linear model with age, gender-adjusted BMI as outcome and selected SNPs as covarites. All effects are highly significant in the final reduced regression model.

have been picked at least 9 times. The selection results of BIC-based lasso are deterministic and replications are not needed. All non-essential parameters (e.g. λ-sequences) used in the R *glmnet* function were set as default in all plasso, CV-based and BIC-based lasso, and 10-fold CV was used in CV-based lasso by default.

A total of 6 414 717 SNPs were selected by plasso, CV-based lasso and BIC-based lasso, respectively. The six selected SNPs by plasso are presented in Table 2. The plasso identified SNPs are quite different from those identified in Bayesian lasso (Li *et al.*, 2011) probably because different methods model different aspects of data. In plasso, we additively coded SNP as a 0/1/2 predictor variable in a regression model, while Bayesian lasso considers both the additive effects and dominant effects in the regression model (Li *et al.*, 2011). On the other hand, both CV-based lasso and BIC-based lasso select much more SNPs than plasso. It is not surprising for CV-based lasso considering its low precision performance in simulation studies (Table 1). According to the simulations, BIC-based lasso, in general selects more false positives than plasso but less than that of CV-based lasso according to Table 1 of simulations. The discrepancy of BIC-based lasso in simulation and in real data analysis is due to the number of null SNPs being analyzed. In simulation, only 20 of either 1000 or 5000 SNPs are true signals while the rest majority are null SNPs. On the other hand, the 1837 SNPs in the real data analysis all have marginal P-values smaller than $10^{-3.5}$ and hence, it is more likely to have much more true signals. As observed in previous publications (Arbet *et al.*, 2017), a smaller number of null SNPs favors smaller λ which corresponds to a less penalized model with more predictors (as BIC-based lasso in this real data analysis).

By producing a much more parsimonious model, plasso enhances the model interpretability in the sense that interpretability decreases if the response is dependent on many predictors and increases if we can reduce the number of features as well as maintain the accuracy. Besides model interpretability, another feature that many geneticists often care is the model predictability. To compare the prediction performance of plasso, CV-based lasso and BIC-based lasso, we randomly split the 977 subjects into a training set with $n_1 = 500$ subjects (indexed by I) and a testing set with the rest $n_2 = 477$ subjects (indexed by II). For each lasso tuning method, we first used the observations $(\mathbf{x}_i, y_i), i \in I$ to build a reduced linear model with selected SNPs and then applied the reduced model to observations $\mathbf{x}_i, i \in II$ to obtain predictions $\tilde{y}_i, i \in II$. Finally, we calculated the squared prediction error (SPE) $= \sum_{i \in II} (y_i - \tilde{y}_i)^2 / n_2$ and the relative squared prediction error (RSPE), which is the ratio of SPE and the sample variance of responses in the testing set. We randomly repeated this sample-splitting procedure 1000 times and compared SPEs and RSPEs of three methods in Figure 2. Clearly, each prediction error of plasso is significantly smaller than that of both CV-based lasso and BIC-based lasso. The large number of SNPs selected in CV-based and BIC-based lasso indicates model overfitting and negatively impacts the model's ability to generalize to new data.

To conclude, we illustrated the potential usefulness of the newly proposed plasso method in terms of model interpretability and predictability. On the one hand, plasso can largely enhance the model interpretability by selecting a more parsimonious model with fewer

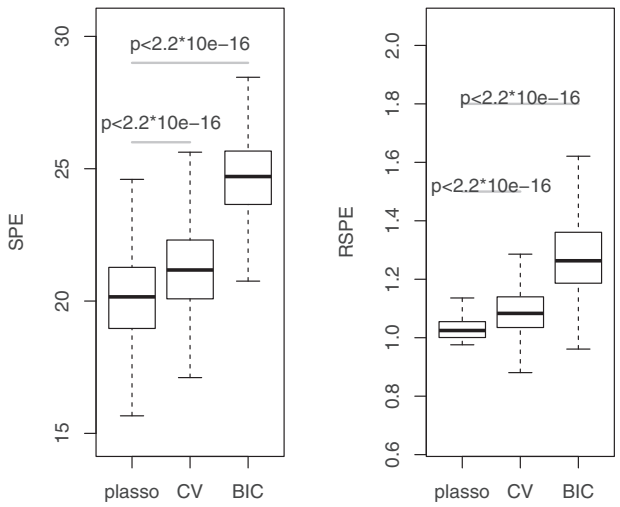

Fig. 2. Prediction performance of P, CV and BIC-based lasso. The boxplot represents the distribution of 1000 SPEs of each method and the number above the gray bar represents P-value of paired two-sample t-test

SNPs compared to traditional lasso tuning methods such as CV and BIC. On the other hand, the model selected by plasso offers better predictability than other lasso methods by avoid overfitting. Putting together, the proposed plasso method provides a useful and powerful GWAS fine-mapping tool which tends to select SNPs in a more reliable and reproducible manner, which can further facilitate downstream functional studies by avoiding costly and potentially fruitless downstream studies on those false positive SNPs.

4 Discussion

In this article, we have proposed a novel method for selection of the lasso tuning parameter λ and further performed the corresponding lasso using the selected λ to GWAS data to select causal SNPs that are of importance to the trait. Besides linear regression model and logistic regression model presented in this article, the proposed plasso method can also accommodate Poisson and Cox models that are allowed in the *glmnet* package. Previously, it is common to select the lasso tuning parameter based on either its predictive performance via CV (Basu *et al.*, 2011; Cho *et al.*, 2010; Waldmann *et al.*, 2013) or likelihood-based model-selection criterion (Chand, 2012; Fan and Tang, 2013). It is interesting that the numerical results in this article suggest that CV may not be a reliable method for selecting SNPs in GWAS. Despite the fact that CV is capable of recovering all true causal SNPs (recall close to one), the price paid by CV is that it selects many false positives (low precision), which can lead to more costly and time-consuming downstream laboratory-based functional studies (Schaid *et al.*, 2018). The newly proposed plasso method has a much lower false discover rate (higher precision); however, it has only slightly lower recall than CV. Overall, the new procedure improves the variable selection performance of lasso, which can largely facilitate biomedical and biological studies (e.g. validation studies, function studies, eventually genetic therapeutics and personalized medicine in the post-GWAS era).

Through this article, we have illustrated the permutation-assisted tuning parameter selection using lasso as an example. The method proposed in this article is very flexible and can be easily extended to other penalized regression models such as SCAD (Fan and Li, 2001) and elastic net (Zou and Hastie, 2005), which have also been widely used in GWAS (Ayers and Cordell, 2010; Yi *et al.*, 2015). We demonstrated our method in the framework of prioritizing genetic variants in GWAS and showed improved prediction performance of plasso over traditional penalized regression models tuned by CV or BIC. The proposed methodology for improving a penalized regression procedure by selecting its underlying penalty parameter is quite general, which also makes it a useful tool in many

other settings such as in the predicting transcriptome step of prediXcan (Gamazon et al., 2015).

Acknowledgements

The authors thank Dr Jiahan Li for providing the data used in the Section Real data analysis of this article. They also want to thank the editor and two reviewers for their insightful comments and suggestions that have significantly improved the article.

Funding

This work has been supported by the National Institutes of Health [grant R21AI144765 to X.Z.; R01CA229542 and P50DA039838 to R.L.; and R01GM126479 and R01HG008983 to D.J.L.] and National Science Foundation [grant DMS-1820702 to R.L.].

Conflict of Interest: none declared.

References

Arbet,J. *et al.* (2017) Resampling-based tests for Lasso in genome-wide association studies. *BMC Genet.*, **18**, 70.

Auer,P.L. *et al.* (2014) Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat. Genet.*, **46**, 629–634.

Ayers,K.L. and Cordell,H.J. (2010) SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.*, **34**, 879–891.

Barber,R. and Candès,E. (2015) Controlling the false discovery rate via knockoffs. *Ann. Stat.*, **43**, 2055–2085.

Basu,S. *et al.* (2011) Multilocus association testing with penalized regression. *Genet. Epidemiol.*, **35**, 755–765.

Candès,E. *et al.* (2018) Panning for gold: 'model–X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B*, **80**, 551–577.

Carbonetto,P. and Stephens,M. (2012) Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.*, **7**, 73–108.

Chand,S. (2012) On tuning parameter selection of lasso-type methods—a Monte Carlo study. In: 9th International Bhurban Conference on Applied Sciences and Technology (IBCAST). IEEE, Islamabad, Pakistan, pp. 120–129.

Cho,S. *et al.* (2010) Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann. Hum. Genet.*, **74**, 416–428.

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

Fan,Y. and Tang,C.Y. (2013) Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B*, **75**, 531–552.

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

Gamazon,E.R. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.

He,Q. *et al.* (2016) Prioritizing individual genetic variants after kernel machine testing using variable selection. *Genet. Epidemiol.*, **40**, 722–731.

Hardy,J. and Singleton,A. (2009) Genomewide association studies and human disease. *N. Engl. J. Med.*, **360**, 1759–1768.

Ionita-Laza,I. *et al.* (2013) Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.*, **92**, 841–853.

Jaquish,C.E. (2007) The Framingham Heart Study, on its way to becoming the gold standard for Cardiovascular Genetic Epidemiology? *BMC Med. Genet.*, **8**, 63.

Lee,S. *et al.* (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Li,J. *et al.* (2011) The Bayesian lasso for genome-wide association studies. *Bioinformatics*, **27**, 516–523.

Luo,X. *et al.* (2006) Tuning variable selection procedures by adding noise. *Technometrics*, **48**, 165–175.

MacArthur,J. *et al.* (2017) The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.

McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Mahmood,S.S. *et al.* (2014) The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*, **383**, 999–1008.

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

Neale,B.M. *et al.* (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, **485**, 242–245.

Schaid,D.J. *et al.* (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, **19**, 491–504.

Srinivasan,A. *et al.* (2019) Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *bioRxiv*, 851337.

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

Visscher,P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.

Waldmann,P. *et al.* (2013) Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.*, **4**, 270.

Wu,Y. *et al.* (2007) Controlling variable selection by the addition of pseudovariables. *J. Am. Stat. Assoc.*, **102**, 235–243.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Wu,M.C. *et al.* (2010) Powerful SNP-set analysis for case-control genome--wide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Yang,S. *et al.* (2019) ET-lasso: a new efficient tuning of lasso-type regularization for high-dimensional data. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, Anchorage, Alaska, pp. 607–616.

Yi,H. *et al.* (2015) Penalized multimarker vs. single-marker regression methods for genome-wide association studies of quantitative traits. *Genetics*, **199**, 205–222.

Zhan,X. *et al.* (2016) A novel copy number variants kernel association test with application to autism spectrum disorders studies. *Bioinformatics*, **32**, 3603–3610.

Zhan,X. *et al.* (2017) Powerful genetic association analysis for common or rare variants with high dimensional structured traits. *Genetics*, **206**, 1779–1790.

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