frontiers in Genetics

METHODS
published: 03 March 2020
doi: 10.3389/fgene.2020.00155

Genetic Variants Detection Based on Weighted Sparse Group Lasso

Kai Che[1], Xi Chen[1], Maozu Guo[1,2,3]*, Chunyu Wang[1] and Xiaoyan Liu[1]

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, [3] Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing, China

Identification of genetic variants associated with complex traits is a critical step for improving plant resistance and breeding. Although the majority of existing methods for variants detection have good predictive performance in the average case, they can not precisely identify the variants present in a small number of target genes. In this paper, we propose a weighted sparse group lasso (WSGL) method to select both common and low-frequency variants in groups. Under the biologically realistic assumption that complex traits are influenced by a few single loci in a small number of genes, our method involves a sparse group lasso approach to simultaneously select associated groups along with the loci within each group. To increase the probability of selecting out low-frequency variants, biological prior information is introduced in the model by re-weighting lasso regularization based on weights calculated from input data. Experimental results from both simulation and real data of single nucleotide polymorphisms (SNPs) associated with *Arabidopsis* flowering traits demonstrate the superiority of WSGL over other competitive approaches for genetic variants detection.

Keywords: genome-wide association studies, genetic variants, single nucleotide polymorphisms, minimum allele frequency, sparse group lasso

OPEN ACCESS

Edited by:
Fa Zhang,
Institute of Computing Technology (CAS), China

Reviewed by:
Qi Ren,
Tianjin University, China
Hui Ding,
University of Electronic Science and Technology of China, China
Qin Ma,
The Ohio State University, United States

***Correspondence:**
Maozu Guo
guomaozu@bucea.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *19 November 2019*
Accepted: *10 February 2020*
Published: *03 March 2020*

Citation:
Che K, Chen X, Guo M, Wang C and Liu X (2020) Genetic Variants Detection Based on Weighted Sparse Group Lasso. Front. Genet. 11:155. doi: 10.3389/fgene.2020.00155

INTRODUCTION

Since completion of the sequencing-based structural genome project, the focus of life science research has gradually shifted from determining the composition of DNA sequences to elucidating the function of identified genes. However, the greatest challenge of functional genomics is to determine the risk genes associated with complex diseases or traits among the huge amount of DNA sequences. Approximately, 90% of all gene fragments in any two individuals of almost all organisms are identical; thus, the fragments affecting individual characteristics, diseases, or traits only appear in a small range of sequences (Tenaillon et al., 2001; Reich et al., 2002). Polygenic recombination or mutation can cause individual differences in genome sequences, resulting in genetic polymorphism. Single nucleotide polymorphisms (SNPs) are the most common form of such genetic variation. Therefore, identification and characterization of SNPs help to discover the underlying causes of various diseases or variable traits and to develop new therapeutic strategies and targets for drug development or crop improvement.

The goal of genome-wide association studies (GWAS) is to elucidate the relationship between millions of SNPs and complex traits (Klein et al., 2005). A single-locus association approach is typically used in GWAS; however, the "polygenic theory" proposes that complex traits are controlled by the action of multiple SNPs together rather than by individual genes or variants (Dudbridge, 2016). Since the number of SNPs far exceeds the number of samples in a multi-loci association study, the "curse of dimensionality" becomes the main challenge of this type of analysis (Waddell et al., 2005). Many machine-learning algorithms have been widely used to overcome this limitation and facilitate investigating the association between traits with SNPs. Based on current approaches, association studies can be divided into two main categories: one based on feature selection (FS) and the other based on statistical machine learning with regularizing penalty.

FS is the process of selecting the most effective features among a set of original features so as to reduce the dimensionality of the dataset. There are two types of FS methods: the wrapper method as a dependent classifier (Hall and Smith, 1999), and the filter method as an independent classifier (Liu and Setiono, 1996). Typically, the wrapper and filter approaches are combined as the final selected method. When applying FS methods to GWAS, the SNPs are treated as the features, phenotypes are the labels, and the candidate SNPs are then selected according to their associations with phenotypes. Numerous FS methods have been applied in genetic association studies (Evans, 2010; Batnyam et al., 2013; Anekboon et al., 2014; Alzubi et al., 2017; An et al., 2017; Setiawan et al., 2018; Tsamardinos et al., 2019). For example, Evans (2010) combined two filter FS methods with classification methods in a machine-learning approach, and obtained strong association results. To further improve the accuracy of the selected SNPs, Batnyam et al. (2013) applied four popular FS approaches (Robnik-Šikonja and Kononenko, 2003; Liang et al., 2008; Seo and Oh, 2012; Lee et al., 2013) to select novel SNPs, which were then used to generate artificial features by applying a feature fusion method. Finally, the artificial features were classified by traditional classifiers. As an alternative combinational algorithm, Anekboon et al. (2014) proposed a correlation-based FS method as a filter to first select a portion of the SNPs, followed by a wrapper phase to sequentially feed each of these SNPs into k-nearest neighbor, artificial neural network, and Ridge regression classifiers. Alzubi et al. (2017) developed a hybrid FS method by combining conditional mutual information maximization and support vector machine-recursive feature elimination (SVM-RFE). An et al. (2017) used a hierarchical feature and sample selection framework to gradually select informative features and discard ambiguous samples in multiple steps to improve the classifier learning. Setiawan et al. (2018) firstly employed random forest algorithm to reduce the search space, then selected associated SNPs by sequential forward floating selection. Tsamardinos et al. (2019) applied p-values of conditional independence tests and meta-analysis techniques to select features, and made use of parallel techonology to increase the computing speed. Current methods based on FS have sufficient ability for selecting a candidate feature set. Nevertheless, it is important to use available biological information as prior knowledge in biocomputing. Since FS methods can only reflect the dataset itself, they are not suitable to screen features based on prior biological knowledge.

Regression models with penalty can also be used for GWAS. With this approach, the SNPs correspond to the independent variables, and phenotypes are mapped to dependent variables in the regression model. Since the number of SNPs typically far exceeds the number of samples, it is necessary to regularize the sparsity of coefficients in the regression model. As a representative example, the well-established lasso method proposed by Tibshirani (1996) can learn a sparse weight vector by penalizing the weight vector with a 1-norm loss while shrinking less important coefficients to zeros. Owing to this property, lasso and its extensions have been widely applied in the detection of genetic variants (Cao et al., 2014; Arbet et al., 2017; Tamba et al., 2017; Cherlin et al., 2018; Wang et al., 2019). For example, Cao et al. (2014) incorporated prior information in lasso to further increase the selection accuracy. Arbet et al. (2017) imposed a permutation method on lasso to improve the performance of the algorithm. Tamba et al. (2017) first reduced the number of SNPs to a moderate size, then used expectation maximization Bayesian lasso to detect the quantitative trait nucleotide (QTN). Cherlin et al. (2018) used lasso to explore the association between phenotype and SNP data and achieved good prediction. Wang et al. (2019) promoted a precision lasso that utilized regularization governed by the covariance and inverse covariance matrices of explanatory variables to increase sparse variable selection. However, SNPs (features) are generally found in groups, whereas lasso does not encourage sparsity between groups. Yuan and Lin, (2006) proposed the group lasso (GL) method, which sets a regularization of the sum of the ℓ_2 norm onto groups that encourages only a few groups to be selected. The GL approach has also been successfully applied in GWAS (Li et al., 2015; Lim and Hastie, 2015; Gossmann et al., 2017; Du et al., 2018). Gossmann et al. (2017) extended sorted L1 penalized estimation (SLOPE) in the spirit of Group LASSO to handle group structures between the predictor variables. Du et al. (2018) proposed the SCCA with truncated L1 penalized and GL to improve the performance and effectiveness of discovering SNPs or QTs in imaging genetics. However, once a group is chosen, all of its comprising features are also selected, which is not compliant with the actual biological situation in which SNPs are distributed sparsely across the genome in only a few groups. Simon et al. (2013) developed sparse GL (SGL) that uses the ℓ_2 penalty to select only a subset of the groups and the ℓ_1 penalty to select only a subset of the variables within the group. Indeed, SGL has been widely applied in detecting genetic variants (Rao et al., 2015; Li et al., 2017; Samal et al., 2017; Guo et al., 2019). Samal et al. (2017) proposed a method based on SGL to identify phenotype associated extreme currents decomposed from metabolic networks data. Combined SGL with group-level graph structure, which takes advantages of gene-level priors to penalty the nucleotide-level sparsity to identify the risk SNPs. Guo et al. (2019) proposed a method that combined SGL and linear mixed model (LMM) for multivariate associations of quantitative traits, and it obtained a good power. Despite this improvement, the limitation of this method is that SGL selects sparse features within a group, but gives the same penalty for all features within the group. Consequently, this approach can easily result in swiping out low-frequency features that may play an important role in influencing phenotypes. To overcome this obstacle, it is important to assign different penalties to different features. Ideally, candidate SNPs should have a smaller penalty weight while others would have a larger penalty weight. In this way, candidate SNPs will stand out among the data more readily. To achieve this goal, we here propose a novel approach termed weighted SGL (WSGL) by introducing biological prior information for more accurate genetic variants detection. Specifically, we compute the minimum allele frequency (MAF) among a dataset of SNPs and use those values to reweight as the ℓ_1 penalty of each SNP site, which can increase the chance of retaining low-frequency variants without loss of information. To validate this approach, we compared the performance of our model with simulation and real data against the three mainstream models discussed above.

MATERIALS AND METHODS

Materials

Simulation Data

We used *Arabidopsis thaliana* data from Atwell et al. (2010), downloaded from https://github.com/Gregor-Mendel-Institute/atpolydb for the simulation. We used a quality control protocol on the original data. The SNPs are eliminated by the standard that Minor Allele Frequency (MAF) is < 0.01, the missing rate is > 0.05, or the allele frequencies are not in Hardy-Weinberg (P < 0.0001). After data preprocessing, we chose 200 genes on chromosome 1 covering a total of 1,993 SNPs. Twenty of these SNPs were chosen as the associated variants.

Real Data

The genotype information was the same as that obtained from the simulation data. Ten phenotypes were selected among the 107 reported. First, from chromosome 1 to 5, we chose the first 1,000 genes, which were sorted according to sequence length, including 49,962 SNPs. Second, we selected 19 genes containing 367 SNPs, which have been verified to be associated with flowering time in *Arabidopsis*. Thus, a total of 50,329 SNPs were analyzed in our experiments.

Statistical Model and Methods

We first give a problem statement, followed by a brief overview of lasso and its extension for application in a genetic association study. Finally, we describe our new WSGL method.

Let $X = (x_1, x_2,\ldots, x_n)^T$ denote the $n \times p$ genotype matrix, where n is the number of samples and p is the number of genotypes. Let $Y = (y_1, y_2,\ldots, y_n)^T$ represent the $n \times 1$ phenotype vector, containing the phenotype values of the n samples. We then establish a linear model between X and y:

$$Y = X\beta + \epsilon \tag{1}$$

where $\beta = (\beta_1, \beta_2,\ldots, \beta_n)^T$ is a $p \times 1$ regression coefficients vector, and $\epsilon \sim N(0, 1)$.

Lasso and Its Extension for Association Mapping

Tibshirani (1996) proposed the popular lasso estimator,

$$\min_{\beta} \frac{1}{2} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \tag{2}$$

where β is the regression coefficients vector, and x, corresponding to the nonzero estimated coefficients in β, represents the candidate SNPs. $||\beta||_1$ is the ℓ_1 penalty item. λ is a regularization parameter, and its size determines the sparsity. When $\lambda = 0$, the lasso estimator is equivalent to ordinary least-squares regression.

However, the lasso applies to the situation in which the variables are independent of each other. For the situation in which the variables can be divided into m groups, Yuan and Lin (2006) proposed the GL estimator,

$$\min_{\beta} \frac{1}{2} \| y - \textstyle\sum_{l=1}^{m} X^{(l)} \beta^{(l)} \|_2^2 + \lambda \textstyle\sum_{l=1}^{m} \sqrt{p_l} \| \beta^{(l)} \|_2 \tag{3}$$

where m is the group of variables, the first part is OLS, the second part is the sum of the ℓ_2 penalty of the coefficients of each group, and λ is the regularization parameter. If the size of the group is 1, it will degenerate to the standard lasso.

The GL can generate a sparse in groups; however, the variables in a group are not sparse. To solve this problem, Simon et al. (2013) proposed the SGL,

$$\min_{\beta} \frac{1}{2n} \| y - \textstyle\sum_{l=1}^{m} X^{(l)} \beta^{(l)} \|_2^2 + (1-\alpha)\lambda \textstyle\sum_{l=1}^{m} \sqrt{p_l} \| \beta^{(l)} \|_2 + \alpha\lambda \| \beta \|_1 \tag{4}$$

where λ still controls the overall penalty and α determines the ratio between ℓ_1 and ℓ_2. When $\alpha = 1$, it will be transformed into lasso, whereas when $\alpha = 0$, it will be GL. SGL can either select the variables in a group-by-group manner, or screen the individual variables in the remaining groups.

Our Method

With respect to the genetic association problem, the variables in a group have different effects on the independent variable. However, the SGL uses the same penalty coefficients for all variables, regardless of the relative importance among SNPs in the screened groups.

To tackle this problem, we introduce the prior information ω in the model to improve the statistical power, and propose the WSGL,

$$\min_{\beta} \frac{1}{2n} \| y - \textstyle\sum_{l=1}^{m} X^{(l)} \beta^{(l)} \|_2^2 + (1-\alpha)\lambda \textstyle\sum_{l=1}^{m} \sqrt{p_l} \| \beta^{(l)} \|_2 + \alpha\lambda \| \omega\beta \|_1 \tag{5}$$

The objective function in (5) is clearly convex; therefore, the optimal solution can be achieved by subgradient equations. Let$\hat{\beta}$ be the optimal solution of WSGL. For group $k = (1, 2,\ldots, m)$, the solution $\hat{\beta}^{(k)}$ satisfies

$$\begin{aligned} &\frac{1}{n} X^{(k)T} \left(y - \textstyle\sum_{l=1}^{m}) X^{(l)} \hat{\beta}^{(l)} \right) \\ &\quad = \sqrt{p_k}(1-\alpha)\lambda\mu^{(k)} + \alpha\lambda\omega^{(k)} \nu^{(k)} \end{aligned} \tag{6}$$

where $\mu^{(k)}$ and $v^{(k)}$ are subgradients of $\| \hat{\beta}^{(k)} \|_2$ and $\| \hat{\beta}^{(k)} \|_1$, respectively. According to Simon et al. (2013), $\mu^{(k)} = \hat{\beta}^{(k)} / \| \hat{\beta}^{(k)} \|_2$ if $\beta^{(k)} \neq 0$; otherwise, $\| \mu^{(k)} \|_2 \leq 1$. $v_j^{(k)} = sign(\hat{\beta}_j^{(k)})$ when $\hat{\beta}_j^{(k)} \neq 0$; otherwise, $\| v_j^{(k)} \|_2 \leq 0$.

Following the analysis in Simon et al. (2013), the condition for $\hat{\beta}^{(k)} = 0$ is

$$\| S\left(X^{(k)T}\gamma_{(-k)}/n, \alpha\lambda\omega^{(k)}\right) \|_2 \leq \sqrt{p_k}(1-\alpha)\lambda \tag{7}$$

where$\gamma_{(-k)} = y - \sum_{l\neq k} X^l \hat{\beta}^{(l)}$is the partial residual of y, and S is defined as $(S(a,b))_j = sign(a_j)(|a_j| - b_j)_+$.

If $\hat{\beta}^{(k)} \neq 0$, the subgradient condition for $\beta_i^{(k)}$ becomes

$$\frac{1}{n} X_i^{(k)T}\left(y - \sum_{l=1}^{m} X^{(l)} \hat{\beta}^{(l)}\right)$$

$$= \sqrt{p_k}(1-\alpha)\lambda \frac{\widehat{\beta}_i^{(k)}}{\| \hat{\beta}^{(k)} \|_2} + \alpha\lambda\omega_i^{(k)} v_i^{(k)} \tag{8}$$

This is satisfied for $\hat{\beta}^{(k)} = 0$, if $|X_i^{(k)T}\gamma_{(-k,i)}| \leq n\alpha\lambda\omega$, where $\gamma_{(-k,i)} = \gamma_{(-k)} - \sum_{j\neq i} X_j^{(k)} \hat{\beta}^{(k)}$ is the partial residual of y.

When $\beta_i^{(k)} \neq 0$, we can get

$$\hat{\beta}_i^{(k)} = \frac{S\left(X_i^{(k)T}\gamma_{(-k,i)}/n, \alpha\lambda\omega\right)}{X_i^{(k)T} X_i^{(k)}/n + (1-\alpha)\lambda / \| \hat{\beta}^{(k)} \|_2} \tag{9}$$

For each locus, MAF indicates, to some degree, its rareness. The MAF of low-frequency variants is usually small, so the associated low-frequency variants are more susceptible to sparsity regularization than other common variants. With normal sparse group lasso, the pressure of being zeroed out on each locus within the same group is equally high. In this case, those low-frequency variants are more likely to be excluded during the process. So selection of an appropriate weight can help to filter out more accurate candidate low-frequency variants.

There are several approaches for deciding the weights. For example, a small penalty can be assigned to the loci in known susceptibility genes to ensure including them into the model. Alternatively, the weights can be dependent on the MAF. For a dataset including both low-frequency and common variants, low-frequency markers are assigned smaller weights to compensate for their low frequencies. Here, we assign each locus a weight as follows: $weight = 2\sqrt{MAF(1 - MAF)}$. Each weight ω_i is calculated in advance, which contains genotypes and biological explanations. The importance of the ith variable can be adjusted by the weight ω_i. Thus, to choose a locus, we can give it a relatively small penalty weight. Conversely, a larger weight can be assigned to exclude a locus. If $\omega_i = 1$, our model will be transformed to the SGL. Moreover, it is important to select an optimal regularization parameter λ, as a larger λ will generate a sparser result. For the present model, we chose cross-validation to select the optimal λ.

A brief algorithmic description of our method is shown in **Algorithm 1**. Let n represent the number of samples and p be the number of genotypes. The time complexity of subgradient step in each iteration is $O(np)$. In real data, p is usually supposed to be large, resulting in comparatively high time complexity. Therefore, in genome-wide association analysis, we suggest to analyze chromosomes individually for huge genome.

ALGORITHM 1 | Parameter estimation for weighted sparse group lasso.

Input: Genotype *X*, phenotype *y* ratio α, regularization hyperparameter λ

```
Output: Estimated $\hat{\beta}$
1: calculate $\omega = 2\sqrt{MAF(1 - MAF)}$;
2: while not converge do
3:     for $k$ from 1 to number_of_groups do
4:         for $i$ from 1 to length_of_groups ($k$) do
5:             update $\hat{\beta}_i^{(k)}$using equation (9);
6: return $\hat{\beta}$;
```

Performance Measurements

For performance evaluation of the new model, we treat the loci detection as a binary classification under class imbalance, in which associated loci are assigned the label 1, and all others are assigned the label 0. The testing frequency of each locus is then regarded as the predicted probability for label 1. The receiver operating characteristic (ROC) curve and the area under the precision-recall curve (AUPR) are typically used for performance assessments. The ROC curve is plotted based on the sensitivity and specificity, whereas AUPR is generated based on the precision and recall. In our problem, the number of variants is significantly lower than the number of all loci, resulting in an imbalanced dataset. In the ROC curve, the false positive rate cannot descend greatly when the true negative is huge. However, the AUPR is sensitive to false positive. Considering these factors, we chose the AUPR as the performance metric for this purpose.

RESULTS AND DISCUSSION

Experiments on Simulation Data

For assessing the performance of WSGL in selecting candidate SNPs associated with a trait of interest, its performance was compared with lasso, GL, and SGL. Two parameters needed to be controlled in this experiment: α, which is the proportion of ℓ_1 and ℓ_2 loss in SGL, and λ, which is the coefficient of the entire regularization term and influences the sparsity. We set α to 0.95. Based on the results of cross-validation, λ was set to 0.09.

Figure 1 shows the results of the four methods with the simulation data, which clearly exhibits the superior performance of WSGL. The AUPR of WSGL is 0.652, which outperformed lasso by 23.6%, GL by 50.8%, and SGL by 24.4%. Lasso uses ℓ_1 to guarantee the sparsity of selected SNPs, but does not consider the group information; therefore, the candidate SNPs may be selected from all groups equally. Although GL imposes group information on the model, it still lacks sparsity constraint within the group, which does not correspond with the biological assumption that only a small number of candidate SNPs are contained in a small number of groups. SGL considers the sparsity between and within groups, but can still easily exclude

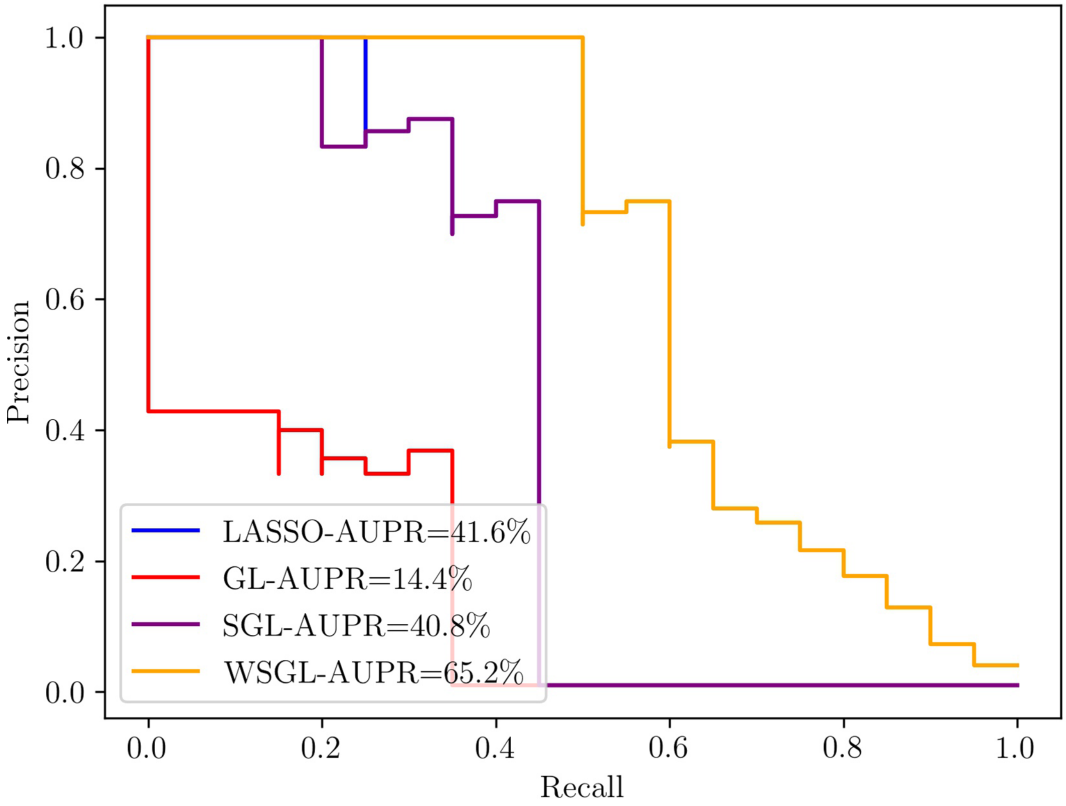

FIGURE 1 | Precision-recall (PR) curves of WSGL and the other methods.

important SNPs with a lower MAF. By introducing biological information to adjust the penalty of SNPs in the selected groups, WGSL places less weight on the low-frequency variants and thus increases their chance of being kept out. Despite its simplicity, the simulation results demonstrated the effectiveness of this approach for screening out important SNPs.

To further compare the performance of the four algorithms, we computed their AUPR values by fixing α at 0.95 and varying λ from 0.01 to 0.1 by steps of 0.01. As shown in **Figure 2**, with smaller λ, the model shows lower sparsity. When λ is 0.01 or 0.02, the model will include more SNPs, which may include more non-candidate SNPs that would cause a high false positive rate. Conversely, as λ increases, the number of selected SNPs decreases, which might result in the loss of some candidate SNPs, leading to a low TP rate. However, WSGL will include more candidate low-frequency loci by introducing prior knowledge to adjust the weight. Accordingly, WSGL keeps the highest position starting from λ = 0.03. When λ increases from 0.02 to 0.05, the AUPR of WSGL increases significantly from 58% to 64.5%, whereas the AUPR of lasso decreases from 59.2% to 53.2%, and that of SGL decreases from 59.9% to 56.1%. Surprisingly, the AUPR of GL decreases even more sharply from 51.1% to 34.1%. When λ reaches 0.05, the AUPR of WSGL tends to be stable, and the peak of 65.2% occurs at λ = 0.09. The AUPR of both lasso and SGL gradually decreases, and finally drops to around 40%. When λ is 0.1, the AUPR of GL drops to 13.3%. These results were consistent with our expectation that the performance of WSGL would be the best, SGL would perform better than lasso, and GL would show the worst performance overall.

Experiments on Real Data

To verify the ability of WSGL to detect candidate SNPs, we compared the performance of the four models using *Arabidopsis* flowering time data with known genetic associations. The dataset included 10 different phenotypes, FT10, FT16, FT22, LD, LDV, SD, SDV, LN10, LN16, and LN22, and the descriptions of the 10 phenotypes are shown in **Table 1**. We analyzed the associated number of genes covered by 100 SNPs with top probabilities of being target loci.

As shown in **Table 2**, WSGL could link more candidate genes with phenotypes FT10, FT16, FT22, LD, SD, and SDV. In particular, WSGL demonstrated excellent performance for FT10, not only by selecting less groups but also by including less SNPs within each group, and the ratio of candidate genes was 23.08%. By contrast, the ratios of candidate genes were 4.65%, 9.09%, and 5.13% for lasso, GL, and SGL, respectively. For phenotypes FT16, FT22, LD, SD, and SDV, WSGL still achieved the best detection performance. However, unexpectedly, the GL model obtained better results for the first four phenotypes. We consider that this may be due to the specific distribution of loci in the dataset. In cases for which most or all of

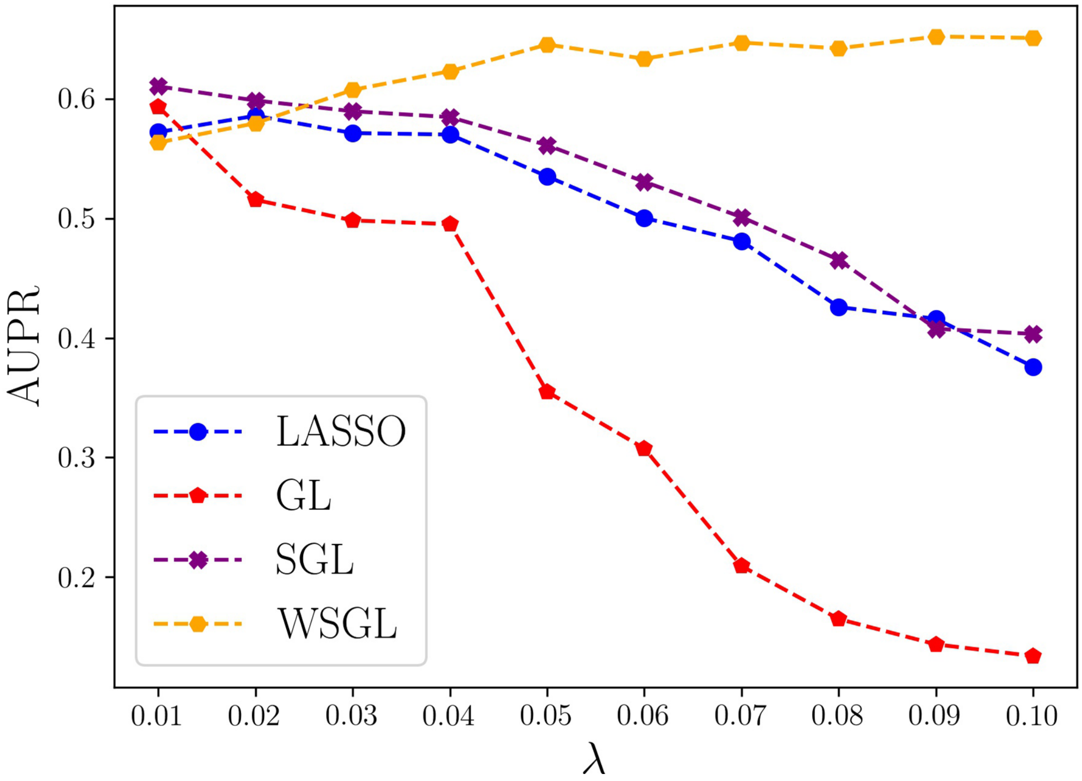

FIGURE 2 | Precision-recall (PR) curves of WSGL and the other methods for varying λ.

TABLE 1 | Description of the 10 flowering related phenotypes in *A.thaliana* in real data application.

Phenotype	Accessions	Phenotype description	Growths conditions	Phenotype scoring
LD	167	Days to flowering time (FT) under Long Day (LD) and Short Days (SD) +/– vernalization	18°C 16-h daylight	Number of days following stratification to opening of the first flower. The experiment was stopped at 200d, and accessions that had not flowered at the point were assigned a value of 200.
LDV	168		18°C 16-h daylight, vernalized (5wks 4)	
SD	162		18°C 16-h daylight	
SDV	159		18°C 16-h daylight, vernalized (5wks 4)	
FT10	194		10°C 16-h daylight	
FT16	193		16°C 17-h daylight	Plants were checked bi-weekly for presence of first buds, and the average flowering time and average leaf number of four plants of the same accession at each temperature were collected.
FT22	193	Flowering time (FT) and leaf number at flowering time (LN)	22 °C 18-h daylight	
LN10	177		10 °C 19-h daylight	
LN16	176		16°C 20-h daylight	
LN22	176		22°C 21-h daylight	

TABLE 2 | Summary of four methods associations found in real data.

Phenotype	Method	Number of genes covered by top 100 SNPs	Number of genes in the 19 genes	Ratio of candidate genes
FT10	Lasso	86	4	4.65%
	GL	66	6	9.09%
	SGL	78	4	5.13%
	WSGL	26	6	23.08%
FT16	Lasso	76	8	10.53%
	GL	62	8	12.9%
	SGL	64	7	10.94%
	WSGL	67	10	14.93%
FT22	Lasso	78	7	8.79%
	GL	72	7	9.72%
	SGL	77	6	7.79%
	WSGL	71	9	12.68%
LD	Lasso	81	9	11.11%
	GL	67	9	13.43%
	SGL	73	11	15.07%
	WSGL	74	12	16.22%
LDV	Lasso	6	6	–
	GL	6	6	–
	SGL	6	6	–
	WSGL	6	6	–
SD	Lasso	78	5	6.41%
	GL	70	5	7.14%
	SGL	79	6	7.59%
	WSGL	77	6	7.79%
SDV	Lasso	84	1	1.19%
	GL	66	1	1.52%
	SGL	78	2	2.56%
	WSGL	72	2	2.78%
LN10	Lasso	6	6	–
	GL	6	6	–
	SGL	6	6	–
	WSGL	6	6	–
LN16	Lasso	6	6	–
	GL	6	6	–
	SGL	6	6	–
	WSGL	6	6	–
LN22	Lasso	6	6	–
	GL	6	6	–
	SGL	6	6	–
	WSGL	6	6	–

the candidate objects are located in only one group, GL will apparently show a good result. By contrast, all four methods could link all six genes with LDV, LN10, LN16, and LN22. This surprising result may reflect the strong association between the selected SNPs and these phenotypes, which is highly discriminable. Nevertheless, this assessment demonstrated that our new weighted method achieves the best performance overall, highlighting the importance of considering prior biological information for selection of candidate SNPs.

CONCLUSION

We proposed a method named weighted sparse group lasso (WSGL) to improve the detection of genetic variants. WSGL incorporates the ℓ_1 penalty, ℓ_2 penalty, and prior biological knowledge into a single linear regression model, and then uses SGL to either select or clear out all SNPs in a group potentially associated with a phenotype of interest. To screen candidate low-frequency variants, we introduced the MAF as the weight to re-scale each element for calculating ℓ_1 loss. In addition, WSGL can detect meaningful associations with more accuracy compared to available methods, which conforms with the general assumption that complex traits are affected by a few SNPs in a few genes. Experiments with both simulation and real data of SNPs related to the flowering time of *A. thaliana* demonstrated the effectiveness of our approach.

DATA AVAILABILITY STATEMENT

We used *Arabidopsis thaliana* data from Atwell et al. (2010), downloaded from https://github.com/Gregor-Mendel-Institute/atpolydb for the simulation.

AUTHOR CONTRIBUTIONS

Conceptualization: KC. Formal analysis: KC. Funding acquisition: MG, CW, and XL. Methodology: KC. Validation: KC and XC. Writing—original draft: KC and XC. Writing—review and editing, KC, MG, CW, and XL.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61571163, 61532014, 61671189, 61872114, and 61871020) and the National Key Research and Development Plan of China (Grant No. 2016YFC0901902).

REFERENCES

Alzubi, R., Ramzan, N., Alzoubi, H., and Amira, A. (2017). A hybrid feature selection method for complex diseases snps. *IEEE Access* 6, 1292–, 1301. doi: 10.1109/ACCESS.2017.2778268

An, L., Adeli, E., Liu, M., Zhang, J., Lee, S.-W., and Shen, D. (2017). A hierarchical feature and sample selection framework and its application for alzheimer's disease diagnosis. *Sci. Rep.* 7, 45269. doi: 10.1038/srep45269

Anekboon, K., Lursinsap, C., Phimoltares, S., Fucharoen, S., and Tongsima, S. (2014). Extracting predictive snps in crohn's disease using a vacillating genetic algorithm and a neural classifier in case–control association studies. *Comput. Biol. Med.* 44, 57–65. doi: 10.1016/j.compbiomed.2013.09.017

Arbet, J., McGue, M., Chatterjee, S., and Basu, S. (2017). Resampling-based tests for lasso in genome-wide association studies. *BMC Genet.* 18, 70. doi: 10.1186/s12863-017-0533-3

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature* 465, 627. doi: 10.1038/nature08800

Batnyam, N., Gantulga, A., and Oh, S. (2013). "An efficient classification for single nucleotide polymorphism (snp) dataset," in *Computer and Information Science* (Berlin, Germany: Springer), 171–185. doi: 10.1007/978-3-319-00804-2_13

Cao, S., Qin, H., Deng, H.-W., and Wang, Y.-P. (2014). A unified sparse representation for sequence variant identification for complex traits. *Genet. Epidemiol.* 38, 671–679. doi: 10.1002/gepi.21849

Cherlin, S., Howey, R. A., and Cordell, H. J. (2018).Using penalized regression to predict phenotype from snp data, in: *BMC Proc. (BioMed Central)* 12 223–228. doi: 10.1186/s12919-018-0149-2

Du, L., Liu, K., Zhang, T., Yao, X., Yan, J., Risacher, S. L., et al. (2018). A novel scca approach via truncated l1-norm and truncated group lasso for brain imaging genetics. *Bioinformatics* 34, 278–285. doi: 10.1093/bioinformatics/btx594

Dudbridge, F. (2016). Polygenic epidemiology. *Genet. Epidemiol.* 40, 268–272. doi: 10.1002/gepi.21966

Evans, D. T. (2010). *A SNP microarray analysis pipeline using machine learning techniques. Ph.D. thesis* (Athens, OH, USA: Ohio University).

Gossmann, A., Cao, S., Brzyski, D., Zhao, L.-J., Deng, H.-W., and Wang, Y.-P. (2017). A sparse regression method for group-wise feature selection with false discovery rate control. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 15, 1066–1078. doi: 10.1109/TCBB.2017.2780106

Guo, Y., Wu, C., Guo, M., Zou, Q., Liu, X., and Keinan, A. (2019). Combining sparse group lasso and linear mixed model improves power for finding genetic variants underlying quantitative traits. *Front. Genet.* 10, 271. doi: 10.3389/fgene.2019.00271

Hall, M. A., and Smith, L. A. (1999). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper, in: *FLAIRS conference.*, Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference; 1999 March 1-5. (Orlando, Florida, USA: DBLP) 1999 235–239.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., et al. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science* 308, 385–389. doi: 10.1126/science.1109557

Lee, J., Batnyam, N., and Oh, S. (2013). Rfs: Efficient feature selection method based on r-value. *Comput. Biol. Med.* 43, 91–99. doi: 10.1016/j.compbiomed.2012.11.010

Li, J., Wang, Z., Li, R., and Wu, R. (2015). Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. Stat.* 9, 640. doi: 10.1214/15-AOAS808

Li, J., Dong, W., and Meng, D. (2017). Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 15, 2028–2038. doi: 10.1109/TCBB.2017.2761871

Liang, J., Yang, S., and Winstanley, A. (2008). Invariant optimal feature selection: A distance discriminant and feature ranking based solution. *Pattern Recognition* 41, 1429–1439. doi: 10.1016/j.patcog.2007.10.018

Lim, M., and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graphical Stat* 24, 627–654. doi: 10.1080/10618600.2014.938812

Liu, H., and Setiono, R. (1996). "A probabilistic approach to feature selection-a filter solution," in *ICML (Citeseer)*, vol. 96. , 319–327.

Rao, N., Nowak, R., Cox, C., and Rogers, T. (2015). Classification with the sparse group lasso. *IEEE Trans. Signal Process.* 64, 448–463. doi: 10.1109/TSP.2015.2488586

Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., et al. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* 32, 135. doi: 10.1038/ng947

Robnik-Šikonja, M., and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.* 53, 23–69. doi: 10.1023/A:1025667309714

Samal, S., Radulescu, O., Weber, A., and Fröhlich, H. (2017). Linking metabolic network features to phenotypes using sparse group lasso. *Bioinf. (Oxf. Engl.)* 33, 3445–3453. doi: 10.1093/bioinformatics/btx427

Seo, M., and Oh, S. (2012). Cbfs: High performance feature selection algorithm based on feature clearness. *PloS One* 7, e40419. doi: 10.1371/journal.pone.0040419

Setiawan, D., Kusuma, W. A., and Wigena, A. H. (2018). Snp selection using variable ranking and sequential forward floating selection with two optimality criteria. *J. Eng. Sci. Technol. Rev.* 11. doi: 10.25103/jestr.115.09

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graphical Stat* 22, 231–245. doi: 10.1080/10618600.2012.681250

Tamba, C. L., Ni, Y.-L., and Zhang, Y.-M. (2017). Iterative sure independence screening em-bayesian lasso algorithm for multi-locus genome-wide association studies. *PloS Comput. Biol.* 13, e1005357. doi: 10.1371/journal.pcbi.1005357

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001). Patterns of dna sequence polymorphism along chromosome 1 of maize (zea mays ssp. mays l.). *Proc. Natl. Acad. Sci.* 98, 9161–9166. doi: 10.1073/pnas.151244298

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B. (Methodol.)* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tsamardinos, I., Borboudakis, G., Katsogridakis, P., Pratikakis, P., and Christophides, V. (2019). A greedy feature selection algorithm for big data of high dimensionality. *Mach. Learn.* 108, 149–202. doi: 10.1007/s10994-018-5748-7

Waddell, M., Page, D., and Shaughnessy, J.Jr. (2005). "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma," in *Proceedings of the 5th International Workshop on Bioinformatics* (ACM), 21–28.

Wang, H., Lengerich, B. J., Aragam, B., and Xing, E. P. (2019). Precision lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics* 35, 1181–1187. doi: 10.1093/bioinformatics/bty750

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. (Stat. Methodol.)* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.