BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 1 2011, pages 1–8
doi:10.1093/bioinformatics/btq600

Genome analysis

Advance Access publication October 29, 2010

A variable selection method for genome-wide association studies

Qianchuan He and Dan-Yu Lin*

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Genome-wide association studies (GWAS) involving half a million or more single nucleotide polymorphisms (SNPs) allow genetic dissection of complex diseases in a holistic manner. The common practice of analyzing one SNP at a time does not fully realize the potential of GWAS to identify multiple causal variants and to predict risk of disease. Existing methods for joint analysis of GWAS data tend to miss causal SNPs that are marginally uncorrelated with disease and have high false discovery rates (FDRs).

Results: We introduce GWASelect, a statistically powerful and computationally efficient variable selection method designed to tackle the unique challenges of GWAS data. This method searches iteratively over the potential SNPs conditional on previously selected SNPs and is thus capable of capturing causal SNPs that are marginally correlated with disease as well as those that are marginally uncorrelated with disease. A special resampling mechanism is built into the method to reduce false positive findings. Simulation studies demonstrate that the GWASelect performs well under a wide spectrum of linkage disequilibrium patterns and can be substantially more powerful than existing methods in capturing causal variants while having a lower FDR. In addition, the regression models based on the GWASelect tend to yield more accurate prediction of disease risk than existing methods. The advantages of the GWASelect are illustrated with the Wellcome Trust Case-Control Consortium (WTCCC) data.

Availability: The software implementing GWASelect is available at http://www.bios.unc.edu/∼lin.

Access to WTCCC data: http://www.wtccc.org.uk/

Contact: lin@bios.unc.edu

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

Received on May 25, 2010; revised on September 22, 2010; accepted on October 20, 2010

1 INTRODUCTION

Genome-wide association studies (GWAS) have become increasingly popular for studying complex human diseases. Within the last several years, the number of single nucleotide polymorphisms (SNPs) per DNA array has grown from 10 000 to 1 million (Altshuler *et al.*, 2008). Despite the very large number of SNPs that are genotyped in a study, GWAS data are commonly analyzed one SNP at a time. Indeed, the Armitage trend test (ATT) is used almost exclusively.

There are at least two strong reasons for considering all the SNPs or at least a large subset of them simultaneously. First, the marginal effects of SNPs (i.e. the effect of each SNP on disease when it is considered alone) may be quite different from their joint effects: (i) a SNP that is not related to disease but is correlated with a causal SNP will be marginally associated with disease; (ii) some SNPs may have weak marginal effects but strong joint effects. Conditional on causal SNPs that are already in the model, false positive signals tend to be weakened while marginally uncorrelated causal SNPs have a better chance of being selected. Second, the predictive power of a single SNP tends to be very low. The accuracy of prediction can be improved substantially by utilizing a large number of relevant SNPs.

It is extremely challenging to decide which set of SNPs should be included in the joint analysis because the number of SNPs in a GWAS is much larger than the sample size. This is commonly referred to as the 'small n, large p' problem. A major difficulty in this problem is that the number and extent of spurious associations between predictors and response increase rapidly with increasing p. Weak effects of causal variants and strong linkage disequilibrium (LD) among SNPs present additional challenges.

There is a large body of literature on variable selection methods, including bridge regression (Frank and Friedman, 1993), least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), smoothly clipped absolute deviation (SCAD; Fan and Li, 2001), elastic net (Zou and Hastie, 2005) and adaptive lasso (Zou, 2006). However, these methods were designed for a moderate number of predictors (i.e. tens or hundreds). For ultra-high p, these methods may be computationally infeasible and statistically inaccurate.

Recently, Fan and Lv (2008) developed the so-called sure independence screening (SIS) strategy for high-dimensional statistical modeling. The idea is to first reduce the dimension from a very large scale to a moderate scale that is below sample size by univariate correlation learning, and then select important predictors by a moderate-scale variable selection method, such as the LASSO or SCAD. In a similar spirit, Wu *et al.* (2009) reduced the dimension of SNPs in a GWAS to several hundreds using a simple score criterion and applied the LASSO to the reduced set of SNPs. A drawback of this approach is that important features that are marginally uncorrelated with response are bound to be missed because the univariate screening step is based entirely on marginal correlations. Fan and Lv (2008) suggested the iterative sure independence screening (ISIS) procedure, which iterates the SIS procedure conditional on the previously selected features so as to capture important features that are marginally uncorrelated with response. Fan and Lv's work is confined to linear regression of a

*To whom correspondence should be addressed.

continuous response, and the number of features they considered is merely thousands.

In this article, we extend Fan and Lv's ISIS idea to logistic regression of case–control GWAS data. This extension is challenging for several reasons. First, the ISIS performs linear regression analysis of residuals, but residuals cannot be used as response variables in logistic regression. Second, prediction errors tend to be much higher for binary outcomes than continuous outcomes. Third, the number of SNPs can be extremely large, typically more than half a million. Fourth, the effects of causal SNPs on complex diseases tend to be small to modest, so the signal-to-noise ratio in GWAS data is low. Fifth, the LD among SNPs is extensive and can be extremely high in certain regions.

A separate challenge is that the false discovery rate (FDR) associated with the ISIS, and indeed with any existing variable selection method, tends to be high. Recently, Meinshausen and B*ü*hlmann (2010) proposed the stability selection strategy to reduce the FDR. The idea is to repeatedly subsample the original data and perform variable selection on each subsample. The features selected frequently among the subsamples tend to be truly associated with outcome and thus should be included in the final model. In this article, we integrate stability selection into our ISIS procedure to develop a new approach, GWASelect, for genome-wide variable selection.

We describe our approach in the next section. In Section 3, we demonstrate through simulation studies that GWASelect has robust performance under a variety of LD structures and can substantially increase the power and reduce the FDR compared with existing methods. In addition, the regression models generated by GWASelect significantly improve prediction accuracy. In Section 4, we apply GWASelect to the GWAS data from the Wellcome Trust Case-Control Consortium (WTCCC, 2007) and show that it yields several novel discoveries and improves prediction accuracy.

2 METHODS

Our ISIS method consists of one marginal SIS and two rounds of conditional SIS. We first describe the marginal SIS procedure. The data contain n subjects and p SNPs. The genotypes of each SNP are standardized by its sample SD. The SIS theory suggests to reduce the original set of features to a small subset whose dimension is in the order of $n/\log n$. Since binary outcomes generally contain less information than continuous outcomes, we shrink the dimension of SNPs from p to $n/(4\log n)$. The SIS theory also suggests to use a large proportion of the subset for the marginal SIS; therefore, we choose to use t SNPs, where t is the integer part of $0.9n/(4\log n)$. That is, we perform the ATT (under the additive model) on each SNP and select the t most significant SNPs to form a set $\mathcal{S}_1$. Then we apply the LASSO to $\mathcal{S}_1$ as follows.

For $i=1,\dots,n$, let Y_i denote the disease status (1 = case, 0 = control), and X_i denote the $(t+1)$-vector consisting of 1 and the genotypes of the t SNPs in $\mathcal{S}_1$. The genotype of each SNP is represented by the number of minor alleles. It is natural to assume the logistic regression model

$$\Pr(Y_i=1|X_i)=\frac{\exp(\beta^{\mathrm{T}}X_i)}{1+\exp(\beta^{\mathrm{T}}X_i)},$$

where $\beta=(\beta_0,\beta_1,\dots,\beta_t)^{\mathrm{T}}$ denotes the vector of unknown regression coefficients. The penalized log-likelihood function takes the form

$$\widetilde{l}(\beta)=\sum_{i=1}^{n}\left[Y_i\beta^{\mathrm{T}}X_i-\log\{1+\exp(\beta^{\mathrm{T}}X_i)\}\right]-\lambda\sum_{j=1}^{t}|\beta_j|,$$

where λ is the tuning parameter.

We adopt the cyclic coordinate decent (CCD) algorithm (Friedman *et al.*, 2010; Genkin *et al.*, 2007), which is tantamount to maximizing $\widetilde{l}(\beta)$ in a component-wise manner. Cross-validation can be used to determine the tuning parameter (and consequently the model size), but for now, we set the model size to a user-specified number, say d. (We will show later how to determine the model size adaptively.) That is, we run the LASSO on a dense grid of λ until it generates a model containing d predictors. If the exact number of d cannot be achieved, we choose the model whose size is right below d. This model is labeled $\mathcal{M}_1$.

To reduce potential collinearity, we prune $\mathcal{M}_1$ using pairwise correlations. Our analysis revealed that 99.9% of the pairwise correlations among the Illumina300K SNPs have absolute values less than 0.8 (corresponding to r^2 of 0.64). Thus, we set the pruning threshold for r^2 to 0.64 so as to minimize the loss of information due to pruning. The pruned model is labeled $\mathcal{M}_1^*$. This marks the end of the marginal SIS.

Assuming that $\mathcal{M}_1^*$ contains t_1 SNPs, we label the set of the remaining $(p-t_1)$ SNPs as $\overline{\mathcal{M}_1^*}$. We use the conditional SIS described below to capture important SNPs in $\overline{\mathcal{M}_1^*}$ that are marginally uncorrelated with disease. The first step is to screen all the SNPs in $\overline{\mathcal{M}_1^*}$ to identify a small set of candidate SNPs that are correlated with Y conditional on $\mathcal{M}_1^*$. This step is computationally challenging because the cardinality of $\overline{\mathcal{M}_1^*}$ is close to p, which can be 1 million. We develop the following conditional score test to accomplish this task in a very efficient manner.

For the i-th subject, let W_i be the (t_1+1)-vector consisting of 1 and the genotypes of the t_1 SNPs in $\mathcal{M}_1^*$. Let Z_j be the j-th SNP in $\overline{\mathcal{M}_1^*}$, and Z_{ji} be the value of Z_j on the i-th subject, where $j=1,\dots,p-t_1$. We assume the logistic regression model:

$$\Pr(Y_i=1|Z_{ji},W_i)=\frac{\exp(\gamma Z_{ji}+\eta^{\mathrm{T}}W_i)}{1+\exp(\gamma Z_{ji}+\eta^{\mathrm{T}}W_i)},$$

where γ and η are unknown regression coefficients. We are interested in testing the null hypothesis $H_0:\gamma=0$. It is computationally intensive to fit the above model for each of the $(p-t_1)$ SNPs. To bypass this difficulty, we perform the conditional score test. Specifically, we calculate $S=U/V^{1/2}$, where

$$U=\sum_{i=1}^{n}\left\{Y_i-\frac{\exp(\widehat{\eta}^{\mathrm{T}}W_i)}{1+\exp(\widehat{\eta}^{\mathrm{T}}W_i)}\right\}Z_{ji},$$

$$V=I_{\gamma\gamma}-I_{\gamma\eta}I_{\eta\eta}^{-1}I_{\gamma\eta}^{\mathrm{T}},$$

$$I_{\gamma\gamma}=\sum_{i=1}^{n}\frac{\exp(\widehat{\eta}^{\mathrm{T}}W_i)}{\{1+\exp(\widehat{\eta}^{\mathrm{T}}W_i)\}^2}Z_{ji}^2,$$

$$I_{\gamma\eta}=\sum_{i=1}^{n}\frac{\exp(\widehat{\eta}^{\mathrm{T}}W_i)}{\{1+\exp(\widehat{\eta}^{\mathrm{T}}W_i)\}^2}Z_{ji}W_i^{\mathrm{T}},$$

$$I_{\eta\eta}=\sum_{i=1}^{n}\frac{\exp(\widehat{\eta}^{\mathrm{T}}W_i)}{\{1+\exp(\widehat{\eta}^{\mathrm{T}}W_i)\}^2}W_iW_i^{\mathrm{T}},$$

and $\widehat{\eta}$ is the maximum likelihood estimator of η under H_0. Note that $\widehat{\eta}$ and $I_{\eta\eta}$ do not involve any data in $\overline{\mathcal{M}_1^*}$ and thus need to be calculated only once at the outset of the conditional SIS. Given $\widehat{\eta}$ and $I_{\eta\eta}^{-1}$, we calculate the test statistic S for each of the $(p-t_1)$ SNPs in $\overline{\mathcal{M}_1^*}$. In vein with the SIS theory, we choose the most significant q SNPs, where q is the integer part of $0.05n/(4\log n)$, and call this set of SNPs $\mathcal{S}_2$. (We use 0.05 since $0.9+0.05+0.05=1$, where 0.9 pertains to the marginal SIS, and $(0.05+0.05)$ to the two rounds of conditional SIS.)

The first step of the conditional SIS is aimed at identifying important SNPs that are marginally uncorrelated (but conditionally correlated) with disease while weakening the priority of those unimportant SNPs that are highly associated with disease through their correlations with the SNPs in $\mathcal{M}_1^*$. In the second step, we combine $\mathcal{S}_2$ with $\mathcal{M}_1^*$ and run the LASSO to select a model $\mathcal{M}_2$ with d SNPs. During this process, new SNPs may be selected, and previously selected SNPs have a chance to be removed from

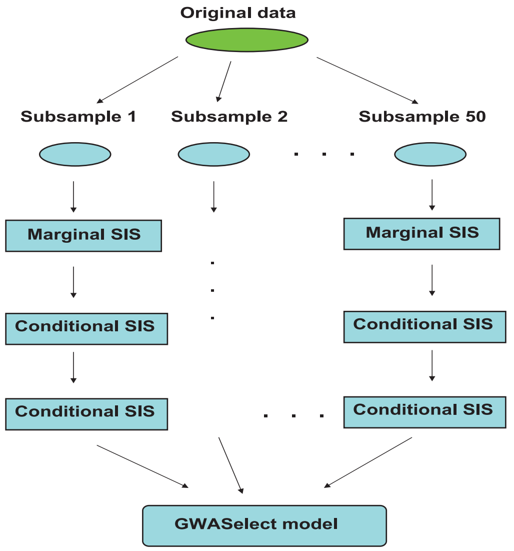

Fig. 1. Flowchart of the proposed GWASelect method.

the model. We prune $\mathcal{M}_2$ to form a new model $\mathcal{M}_2^*$. This completes the conditional SIS.

To increase the opportunities of capturing important SNPs, we repeat the conditional SIS once and call the final model $\mathcal{M}_3^*$. We refer to $\mathcal{M}_3^*$ as the ISIS model.

To reduce the FDR, we combine the extended ISIS procedure with the stability selection strategy (Meinshausen and Bühlmann, 2010) to create the GWASelect method, as illustrated in Figure 1. Specifically, we randomly obtain half of the cases and half of the controls from the GWAS data to form a subsample and then run the ISIS on this subsample. The resulting model is named $\mathcal{T}_1$. Repeating this subsampling concatenated with the ISIS 50 times, we obtain $\mathcal{T}_1,\dots,\mathcal{T}_{50}$. Let $\mathcal{T}=\cup_{j=1}^{50}\mathcal{T}_j$, and denote $\mathcal{T}=\{v_1,\dots,v_L\}$. We then calculate the selection probabilities for the L SNPs in $\mathcal{T}$

$$\pi_l=\sum_{j=1}^{50} I(v_l\in\mathcal{T}_j)/50,\quad l=1,\dots,L,$$

where $I(\cdot)$ is the indicator function. We choose the d SNPs with the highest selection probabilities from $\mathcal{T}$ to form the GWASelect model.

It is sometimes desirable to determine the model size adaptively from the data. To this end, we develop dynamic-GWASelect (d-GWASelect), which contains two modifications to the GWASelect. The first modification is that cross-validation is used to determine the tuning parameter for the LASSO embedded in the ISIS. Specifically, we divide the data randomly into five equal parts, with the k-th ($k=1,\dots,5$) part being the testing data and the remaining four parts being the training data. For a given tuning parameter λ, we apply the LASSO to the training data and select the SNPs that have non-zero regression coefficients. We calculate the liability score (i.e. the linear predictor) for each testing subject. Let $\mathcal{J}_1$ denote the set of subjects with the highest $\delta\times 100\%$ liability scores, and $\mathcal{J}_2$ the set with the lowest $\delta\times 100\%$, where δ is a user-specified number between 0 and 0.5. We then calculate the δ-error rate, defined as $(\sum_{i\in\mathcal{J}_1}|Y_i-1|+\sum_{i\in\mathcal{J}_2}|Y_i-0|)/(2\delta\widetilde{n})$, where $\widetilde{n}$ is the number of subjects in the testing data. We choose the value of λ that minimizes the δ-error rate averaged over the five testing datasets for $\delta=0.1$.

The second modification is that, instead of fixing the model size at d, we specify a selection threshold ξ and select all SNPs with selection probabilities $\geq\xi$. As shown in the next section, the influence of ξ on the final model is typically small.

3 SIMULATION STUDIES

Each simulated dataset contained 2000 cases and 2000 controls. For each subject, we simulated 20 chromosomes, each containing 3000 SNPs. The disease status was generated from the logistic regression model containing 10 causal variants, $G_1,\dots,G_{10}$, with the vector of log odds ratios β^*.

We considered three simulation schemes for the causal SNPs. In the first scheme, we simulated 10 independent causal SNPs that are located on 10 different chromosomes, with minor allele frequencies (MAFs) of 0.3. We set $\beta^*=(-0.35,-0.35,0.35,0.35,0.35,0.35,0.35,-0.35,-0.35,-0.35)^{\mathrm{T}}$.

In the second scheme, we let $\{G_1,\dots,G_{10}\}$ reside on one chromosome and have a special correlation structure such that the correlation between any two causal variants is nearly 0.6. We set $\beta^*=(0.3,0.3,0.3,0.3,0.3,0.3,0.3,0.3,0.3,0.3)^{\mathrm{T}}$.

In the third scheme, multiple causal SNPs were generated to be marginally uncorrelated with Y. We let the first causal SNP be independent of the other nine causal SNPs. The latter were simulated to form three clusters, $\{G_2,G_3,G_4\}$, $\{G_5,G_6,G_7\}$ and $\{G_8,G_9,G_{10}\}$, each cluster residing on one chromosome. The three clusters are independent of each other, but within each cluster, SNPs have a compound symmetry correlation structure with correlation 0.5. We set $\beta^*=(0.5,-0.5,-0.5,0.5,0.5,-0.5,0.5,-0.5,0.5,-0.5)^{\mathrm{T}}$. Under this scheme, $\mathrm{corr}(Y,G_4)$, $\mathrm{corr}(Y,G_6)$ and $\mathrm{corr}(Y,G_9)$ are equal to 0.

For all three schemes, the positions of the causal variants on the chromosomes were randomly chosen. Thus, our simulation results would not be affected by any local LD patterns.

It is not trivial to simulate non-causal SNPs as they are desired to mimic the actual LD structure of human population. There exist several genome simulators based on the coalescent approach (Hudson, 2002), for which the users have to arbitrarily specify a number of parameters. As an alternative, the GWAsimulator of Li and Li (2008) employs a moving-window mechanism and can simulate genotypes based on the Illumina HumanHap300 chip data. We adopted the latter approach. Because the average distance between two SNPs for the Illumina HumanHap300 chip data is roughly 10 kb, the total length we simulated is approximately 600 Mb, which accounts for 1/5 of the whole genome. The LD was well preserved and no trimming was done for the simulated data.

Hoggart *et al.* (2008) explored variable selection from a Bayesian point of view by imposing the Laplace prior or the normal exponential gamma prior on each SNP. The former prior yields the LASSO procedure, while the latter generates a more sparse model and is called hyper-LASSO (HLASSO). We included the HLASSO in our simulation studies. Thus, we analyzed the simulated data by five methods: (i) the ATT method, for which the threshold for declaring significance was set to 0.05/60000 (i.e. Bonferroni correction); (ii) the method by Wu *et al.* (2009); (iii) the (extended) ISIS method; (iv) the HLASSO; and (v) the GWASelect method. We set the model sizes to 15 for all methods (except the ATT) because most biology labs are likely to restrict their resources to a small number of top SNPs.

There are different criteria to evaluate a variable selection method. We chose to use the true discovery rate (TDR) and false discovery

Table 1. True and false discoveries of variable selection methods when the model sizes are fixed at 15 (except for the ATT method)

	ATT	Wu *et al.*	ISIS	HLASSO	GWASelect
Scheme 1					
Model size	26	15	15	15	15
TPC[a]	9.59	9.92	9.98	9.99	9.93
FPC[b]	0.06	2.02	4.04	3.84	1.52
TDR (%)	95.9	99.2	99.8	99.9	99.3
FDR (%)	0.6	15.8	28.5	26.4	12.4
Scheme 2					
Model size	103	15	15	15	15
TPC[a]	9.90	9.68	7.99	9.27	9.07
FPC[b]	4.8	1.05	3.88	4.89	0.03
TDR (%)	99.0	96.8	79.9	92.7	90.7
FDR (%)	31.5	8.6	31.0	32.8	0.3
Scheme 3					
Model size	41	15	15	15	15
TPC[a]	7.07	6.97	8.80	9.99	9.29
FPC[b]	0.08	4.97	4.88	4.47	1.92
TDR (%)	70.7	69.7	88.0	99.9	92.9
FDR (%)	1.0	40.3	35.3	29.4	16.0
G4[c] (%)	0	1	89	100	96

[a] Number of true positive clusters.
[b] Number of false positive clusters.
[c] The rate of capturing the fourth causal SNP, which is marginally uncorrelated with disease under scheme 3.

rate (FDR; Benjamini and Hochberg, 1995) because the main goal of GWAS is to identify causal variants. For genetic studies, how to define the true discovery and false discovery is a delicate issue. This is because once a SNP is declared to be significant, all SNPs that are close to and in LD with that SNP will be followed up. We defined the true positive and false positive as follows. If a captured SNP was no more than 50 SNPs away from a true causal SNP and had $r^2 > 0.05$ with that same causal SNP, then we classified it as a true positive. [Our experiments revealed that replacing 50 with 20 yielded similar results; Hoggart *et al.* (2008) provided a rationale for choosing 0.05 for r^2.] If more than one SNP satisfied these conditions, we counted them only as one true positive cluster. The remaining captured SNPs were classified as false positives. If two false positive SNPs were no more than 10 SNPs apart (i.e. within 100 kb in distance), we counted them as only one false positive cluster. The calculations of the TDR and FDR were based on clusters, rather than on individual SNPs. For each simulation scheme, the number of replications was set to 200. The results are shown in Table 1.

Scheme 1 was designed to compare the five methods under a scenario where all causal variants are independent and their effects are moderate. Under this scheme, all five methods yield high TDRs (>95%), but the FDRs are highly variable. Despite a large model size, the ATT method has the lowest FDR. This seemingly paradoxical phenomenon is explained by the fact that most of the SNPs in the ATT model are highly clustered due to strong LD. The GWASelect model has an elevated FDR, but far lower than the ISIS and the HLASSO, and slightly lower than the Wu *et al.* model. This demonstrates that, by repeated subsampling and variable selection, GWASelect is able to remove many noise features from the model. The ATT method appears to be a good option when causal variants are independent with moderate effects, but if one wishes to achieve higher power without too many false discoveries, the GWASelect method would be a reasonable choice.

Table 2. Prediction accuracy of variable selection methods when the model sizes are fixed at 15 (except for the ATT method)

	ATT	Wu *et al.*	ISIS	HLASSO	GWASelect
Scheme 1					
p-diff[a]	0.023	0.023	0.028	0.028	0.021
liab-correl[b]	0.931	0.943	0.919	0.920	0.948
log-likelihood	−760.0	−759.3	−763.4	−763.1	−758.6
Scheme 2					
p-diff[a]	0.067	0.028	0.073	0.048	0.053
liab-correl[b]	0.912	0.986	0.912	0.961	0.955
log-likelihood	−976.0	−938.9	−983.8	−955.7	−957.7
Scheme 3					
p-diff[a]	0.045	0.050	0.037	0.028	0.027
liab-correl[b]	0.801	0.771	0.874	0.937	0.926
log-likelihood	−720.0	−725.9	−710.9	−701.9	−701.8

[a] The absolute difference between the model-predicted and true disease probabilities.
[b] Liability correlation.

In scheme 2, all 10 causal variants are correlated with each other, which makes variable selection more challenging. It can be shown that under this scheme, the marginal effects of the causal SNPs are much higher than their joint effects. For variable selection, this has the undesired effect of selecting unimportant SNPs that are in proximity of the causal SNPs. Reflecting this fact, the ATT, the ISIS, and the HLASSO all have FDRs above 30%. The GWASelect is able to keep the FDR at a low level and preserve most of the power because of stability selection. The Wu *et al.* method has high power and a relatively low FDR, suggesting that this method is particularly capable of distinguishing causal SNPs from unimportant SNPs that are in LD with them.

Scheme 3 represents a more complex correlation structure in which the three causal SNPs (i.e. the fourth, sixth and ninth SNPs) are marginally uncorrelated with Y. As expected, the methods that are strongly driven by marginal correlations, such as the ATT and the Wu *et al.* method, almost completely missed G_4, which drives down their power to 70%. Both the ISIS and the HLASSO methods achieved higher power, but at the price of high FDRs (around 30%). The GWASelect model offers a more balanced solution in terms of the TDR and FDR.

In summary, only the HLASSO and the GWASelect were able to keep their power above 90% under all three schemes, and the latter appears to have a much lower FDR. The other three methods either lack power under some schemes or entail high FDRs in others.

Next, we investigated the prediction accuracy of the five methods. For each scheme, we further simulated 2000 testing subjects under the prospective sampling. To avoid numerical instabilities, we pruned the obtained models and used the pruned models for prediction. We calculated the true liability score and the estimated liability score for each subject and used the correlation between the two scores as a measure of prediction accuracy. We also calculated the absolute difference between the model-predicted and true disease probabilities, termed as p-diff, to measure the prediction error. The results are shown in Table 2.

Table 3. True and false discoveries of variable selection methods with cross-validation incorporated (except for the ATT method)

	ATT	Wu *et al.*	ISIS	HLASSO	d-GWASelect
Scheme 1					
Model size	32	102	75	42	20
TPC[a]	9.95	10.00	10.00	10.00	10.00
FPC[b]	0.04	63.61	47.24	30.77	0.85
TDR (%)	99.5	100.0	100.0	100.0	100.0
FDR (%)	0.4	86.1	82.0	40.8	7.2
Scheme 2					
Model size	77	49	50	14	20
TPC[a]	9.73	9.88	9.40	7.96	9.09
FPC[b]	1.26	16.95	21.63	5.97	0.21
TDR (%)	97.3	98.8	94.0	79.6	90.9
FDR (%)	10.8	54.8	67.7	17.1	2.0
Scheme 3					
Model size	39	101	68	59	22
TPC[a]	7.01	7.13	9.99	9.87	9.85
FPC[b]	0.04	62.82	39.20	47.41	0.65
TDR (%)	70.1	71.3	99.9	98.7	98.5
FDR (%)	0.4	89.6	78.8	57.3	5.7
G4[c] (%)	0	1	100	89	87

[a] Number of true positive clusters.
[b] Number of false positive clusters.
[c] The rate of capturing the fourth causal SNP, which is marginally uncorrelated with the disease outcome under scheme 3.

Table 4. Prediction accuracy of variable selection methods with cross-validation incorporated (except for the ATT method)

	ATT	Wu *et al.*	ISIS	HLASSO	d-GWASelect
Scheme 1					
p-diff[a]	0.018	0.076	0.073	0.046	0.016
liab-correl[b]	0.951	0.671	0.702	0.849	0.968
log-likelihood	−661.4	−745.9	−739.3	−719.6	−659.7
Scheme 2					
p-diff[a]	0.033	0.046	0.063	0.027	0.027
liab-correl[b]	0.912	0.873	0.802	0.942	0.946
log-likelihood	−754.3	−771.4	−795.4	−756.8	−749.4
Scheme 3					
p-diff[a]	0.039	0.082	0.061	0.062	0.017
liab-correl[b]	0.786	0.519	0.751	0.787	0.964
log-likelihood	−645.1	−725.5	−681.7	−713.7	−624.8

[a] The absolute difference between the model-predicted and true disease probabilities.
[b] Liability correlation.

The Wu *et al.* method excels under scheme 2, consistent with its high TDR and low FDR under this scheme. However, both the Wu *et al.* and the ATT are less accurate than the other methods under scheme 3 because they missed those marginally uncorrelated SNPs. The HLASSO performs well under schemes 2 and 3, suggesting that high prediction power can be achieved even if some noise features are included in the model. Only the HLASSO and the GWASelect have prediction accuracy above 0.9 under all three schemes.

To assess data-adaptive choice of model size, we repeated the above simulation studies but now incorporated a 5-fold cross-validation into all the methods (except the ATT) by using the 10% error rate as the evaluation criterion (see Section 2). For d-GWASelect, we set the selection threshold ξ to 0.3. All effect sizes were set to be moderate. For both schemes 1 and 3, $\beta^* = (0.4, -0.4, -0.4, 0.4, 0.5, -0.5, 0.5, -0.6, 0.6, -0.6)^{\mathrm{T}}$. For scheme 2, $\beta^* = (0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2)^{\mathrm{T}}$. The results are shown in Tables 3 and 4.

The d-GWASelect remains a robust variable selection method under all three schemes and indeed appears to have a better performance than its counterpart with a fixed model size. The Wu *et al.*, the ISIS and the HLASSO now entail extremely high FDRs and poor prediction accuracy. The reason is that cross-validation often favors a large model size for logistic regression, especially when the signal–noise ratio is low. The d-GWASelect method, however, has a well-controlled model size because stability selection sifts away many noise features. Overall, the d-GWASelect enjoys low FDR, high TDR and excellent prediction performance. Replacing the selection threshold with 0.4 yielded highly similar results (data not shown). We also explored the use of deviance (instead of the 10% error rate) as the evaluation criterion for cross-validation, and the d-GWASelect remains more favorable than the other methods (Supplementary Tables S1 and S2).

4 ANALYSIS OF WTCCC DATA

The WTCCC study examined approximately 2000 subjects for each of seven common diseases and a shared set of approximately 3000 controls. Each subject was genotyped on the Affymetrix GeneChip 500K Mapping Array Set. We provide in this section a detailed analysis of the data on Type II diabetes [T2D (MIM 125853, http://www.ncbi.nlm.nih.gov/omim)] and Type I diabetes [T1D (MIM 222100)]; the analysis for the other five diseases is presented in Supplementary Tables S3–S7.

We excluded a small number of subjects according to the sample exclusion lists provided by the WTCCC. In addition, we excluded a SNP if (i) it is on the SNP exclusion list provided by the WTCCC; (ii) it has a poor cluster plot as defined by the WTCCC; (iii) its MAF <0.01 in both cases and controls; or (iv) it has extreme departure from Hardy–Weinberg equilibrium ($P < 10^{-4}$). Approximately 390 000 SNPs were used in the analysis, and there were 2938 controls, 1924 T2D cases and 1963 T1D cases.

Figure 2 indicates the SNPs selected by the ATT, Wu *et al.*, HLASSO and GWASelect for T2D; the details are shown in Table 5 and Supplementary Table S9. Under the ATT method, 15 SNPs reach the genome-wide significance of $P < 10^{-7}$. The most significant one is rs4506565 (P-value $= 7.5 \times 10^{-13}$), which is located in gene *TCF7L2*. The other 14 SNPs are clustered within either *TCF7L2* or *FTO*. These results are consistent with the WTCCC's findings. The HLASSO model is essentially identical to the ATT model, albeit with a smaller model size.

For the Wu *et al.* and GWASelect methods, we set the model sizes to 20. Both methods successfully detected *TCF7L2* and *FTO*. They also identified a locus that spans *TSPAN8/LGR5*, which was one of the most significant loci reported in a recent meta-analysis of 10 128 subjects (Zeggini *et al.*, 2008). This finding demonstrates empirically that regression-based variable selection methods can be more powerful than the ATT method.

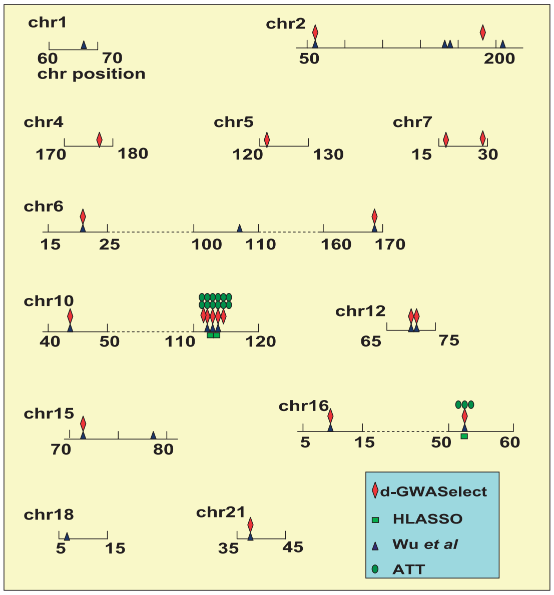

Fig. 2. The T2D models selected by four different methods.

It is interesting to compare the GWASelect and Wu *et al.* models. Five SNPs, rs11688935, rs6846031, rs6872465, rs2389591 and rs10435018, show up only in the GWASelect model. Among these SNPs, rs6846031 was selected partly due to its conditional correlation with T2D, underscoring the importance of conditional screening in variable selection. This finding also indicates that genetic factors underlying T2D are not simply in parallel with each other, but rather form a complex structure that needs to be carefully dissected.

Several SNPs in our GWASelect model have not been reported in the literature on T2D. Some of them are plausibly related to T2D. For example, GULP1 is an adaptor protein that binds and directs the trafficking of LRP1 (Su *et al.*, 2002), a protein that has been shown to play a critical role in adipocyte energy homeostasis and insulin sensitivity (Hofmann *et al.*, 2007). Thus, genetic variants in *GULP1* may potentially influence the amount of LRP1 in adipocyte cells and thereby modulate a person's risk to T2D. As another example, the CREB5 was recently found to be downregulated along with other members of the insulin signaling cascade when stimulated by a ligand of PPARγ, which is known to be associated with T2D (Herrmann *et al.*, 2009). This suggests that CREB5 is closely related to PPARγ and the insulin pathway. The other SNPs do not have known connections with T2D, but further investigation of those loci may reveal novel mechanisms or pathways related to T2D.

For prediction of T2D, the δ-error rates (with $\delta = 0.1$) of all four models are over 40%, suggesting that T2D is greatly influenced by other types of genetic variations and environmental factors. Since it is not very meaningful to compare prediction errors at such a high level, we turned our attention to the T1D data because it is well-known that T1D is genetically more homogeneous than T2D.

For the T1D data, we used cross-validation to choose the tuning parameter for the d-GWASelect method and set the selection threshold ξ to 0.20. For the Wu *et al.* method, we set the model size to 15. The results are shown in Figure 3, Table 6 and Supplementary Table S8. The d-GWASelect model contains 14 SNPs, among which

Table 5. List of SNPs selected by the GWASelect for T2D in the WTCCC study

SNP[a]	Chromosome	Gene[b]
rs11688935	2	*GULP1*
rs903228	2	*ASB3/LOC129656*
rs6846031	4	*VEGFC/NEIL3*
rs6872465	5	*PRDM6*
rs10806665	6	*THBS2/SMOC2*
rs9465871	6	*CDKAL1*
rs2389591	7	*TMEM195/LOC729920*
rs10435018	7	*CREB5*
rs7917983	10	*TCF7L2*
rs7901695	10	*TCF7L2*
rs4506565	10	*TCF7L2*
rs4132670	10	*TCF7L2*
rs7077039	10	*TCF7L2*
rs9326506	10	*ZNF239*
rs1495377	12	*TSPAN8/LGR5*
rs7961581	12	*TSPAN8/LGR5*
rs2930291	15	*CCDC33*
rs8050136	16	*FTO*
rs2099106	16	*C16orf72/GRIN2A*
rs6517434	21	*KCNJ6*

[a] rs number identified from dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/).
[b] Gene symbol from Entrez Gene (http://www.ncbi.nlm.nih.gov/gene/).

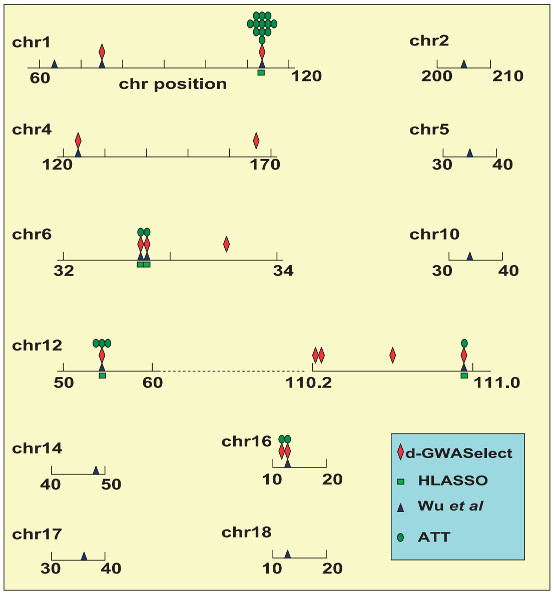

Fig. 3. The T1D models selected by four different methods.

ADAM29, *SYNGAP1*, *CUX2* and *ALDH2* do not appear in any of the other three models. The gene *SYNGAP1* was observed to have strong conditional correlation with T1D, demonstrating again that selection solely based on marginal correlation is insufficient. Searching the T1DBase (http://www.t1dbase.org) revealed that all four genes have expressions in pancreas, although none has been previously considered as strong candidates for T1D. Interestingly, the *CUX2* has been shown to directly regulate the expression of

Table 6. List of SNPs selected by the d-GWASelect method for T1D in the WTCCC study

SNP[a]	Chromosome	Gene[b]
rs6679677	1	*RSBN1/PTPN22*
rs41515647	1	*ST6GALNAC5*
rs17388568	4	*ADAD1*
rs330483	4	*ADAM29*
rs9273363	6	*HLA-DQB1*
rs9272346	6	*HLA-DQA1*
rs411136	6	*SYNGAP1*
rs1265566	12	*CUX2*
rs7398833	12	*CUX2*
rs10744777	12	*ALDH2*
rs17696736	12	*C12orf30*
rs11171739	12	*ERBB3*
rs12708716	16	*CLEC16A*
rs12924729	16	*CLEC16A*

[a] rs number identified from dbSNP.
[b] Gene symbol from Entrez Gene.

Table 7. Prediction errors of four variable selection methods for the WTCCC T1D data

Model	Effective size	δ-error rate			Log-likelihood
		0.1	0.15	0.25	
ATT	5	0.110	0.139	0.181	−2116.9
Wu *et al.*	14	0.119	0.139	0.179	−2075.1
	21	0.135	0.157	0.196	−2059.8
HLASSO	4	0.116	0.141	0.176	−2113.6
	21	0.126	0.151	0.191	−2073.5
d-GWASelect	21	0.107	0.131	0.178	−2058.6

NeuroD (Lulianella *et al.*, 2008), a gene that can cause T1D if mutated.

Finally, we compared the prediction accuracy of the four methods. We randomly divided the data into three parts, two as the training data and one as the testing data. Since the training dataset contains only 2/3 of the original data, we reduced ξ from 0.20 to 0.10 to ensure that a similar number of loci are included in the d-GWASelect model. Since the true liability scores and disease probabilities are unknown in real data, we measured the prediction errors by the δ-error rates for $\delta = 0.1$, 0.15 and 0.25 (see Section 2 for detail). Considering that pruning was done before each model was used for prediction, we report the actual (i.e. effective) number of SNPs used by each model for prediction. Under default settings, the effective model sizes of the Wu *et al.*, the HLASSO and the d-GWASelect are 14, 4 and 21, respectively. Since the former two models are much smaller, we also evaluated their prediction accuracy with 21 effective SNPs. (We were not able to evaluate the ATT with 21 effective SNPs due to numerical instabilities.) The results are reported in Table 7. Clearly, the d-GWASelect performs the best or nearly the best for all three δ-error rates. We have also calculated the area under the ROC curve for the four methods, and GWASelect achieves the highest value (Supplementary Table S11).

5 DISCUSSION

We have developed a new tool, GWASelect, for variable selection at the genome-wide level. This regression-based method has the ability to capture both marginally correlated and marginally uncorrelated causal SNPs and has low FDR. The advantages over the existing methods have been demonstrated through simulated and real data. Our method has two versions. The first version requires the specification of the model size d, for which we suggest to choose a number that is consistent with the current biological knowledge of the studied disease. The second version (d-GWASelect) does not require the specification of the model size, and this is the version we recommend for general use.

The correlation structures for causal variants used in our simulation studies have biological relevance. Scheme 2 mimics a scenario in which the causal variants form a gene cluster that contributes synergistically to the disease outcome, while scheme 3 reflects a scenario in which several biological pathways (or networks) affect the disease development.

We did not include least angle regression (LARS) in our studies because it has been shown to have highly similar performance to LASSO (Hastie *et al.*, 2009). Indeed, LASSO can be implemented by LARS with a small modification. Wu and Lange (2008) demonstrated that CCD is ‘considerably faster and more robust than LARS’ and is ‘more successful than LARS in model selection’.

The HLASSO adopts a concave penalty function, but the CCD algorithm may not converge for non-convex penalty functions (Friedman *et al.*, 2010; Wu *et al.*, 2009). A valid algorithm to implement concave penalty functions is local linear approximation (Zou and Li, 2008), which amounts to multiple rounds of CCD and would make the HLASSO computation prohibitively expensive. For the WTCCC T1D data, running the CCD version of the HLASSO with 10 iterations on an Intel Quadcore Nehalem processor (2.4 GHz, 16 GB memory) requires 67.5 to 175 h, depending on the value of the tuning parameter. In contrast, we have been running the GWASelect in a parallel computing environment, and the same analysis can be completed within several hours on 16 processors.

In an independent effort, Fan *et al.* (2009) developed an ISIS method for generalized linear models in the context of microarray data analysis. In their method, the conditional screening procedure requires fitting a separate regression model for each feature, which would create heavy computational burden for GWAS data. In addition, their method tends to have high FDR. They observed that cross-validation tends to yield large models for logistic regression, resonating our findings.

We can extend our methods to select interactions. Instead of considering all possible interaction terms, we may incorporate known biological network information (Franke *et al.*, 2006) into our selection procedure. Another approach is to first extend the existing genetic network identification tools, such as the liquid association (Li, 2002) and bounded mode stochastic search (Dobra, 2007), to infer SNP interactions and then incorporate such information into our GWASelect procedure. Recently, Han *et al.* (2010) proposed a Markov blanket-based method to evaluate epistatic interactions for GWAS data. It will be interesting to compare to that method when we extend our work to interaction effects.

How to obtain P-values for high-dimensional variable selection is an active research area. The stochastic error introduced by the selection process makes it very difficult to assign P-values to the selected features. Meinshausen *et al.* (2009) offered one possible solution by 'aggregating' P-values from stability selection, but our experiments indicated that this procedure is too conservative for SNP data, likely due to the ultra-high dimension and strong LD.

The prediction of genetic risk using GWAS data has drawn considerable attention in recent years. Wray *et al.* (2007) pioneered this area of research. Their approach selected genetic predictors by a univariate screening method. As shown in this article, our GWASelect method tends to provide more accurate prediction than univariate screening when the SNPs are in strong LD. Wei *et al.* (2009) explored genetic risk prediction through a Support Vector Machine algorithm. It is difficult to compare our results directly with theirs because (i) their analysis involved two other datasets besides the WTCCC T1D data; (ii) our testing samples are far smaller than theirs; and (iii) interaction effects are not considered in our current work.

Funding: National Cancer Institute grant (1-P01-CA142538-01).

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al*. (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Dobra,A. (2007) Variable selection and dependency networks for genomewide data. *Biostatistics*, **8**, 1–18.

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

Fan,J and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, **70**, 849–911.

Fan,J. *et al*. (2009) Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 2013–2038.

Frank,I.E. and Friedman,J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–135.

Franke,L. *et al*. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

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

Genkin,A. *et al*. (2007) Large-scale Bayesian logistic regression for text categorization. *Technometrics*, **49**, 291–304.

Han,B. *et al*. (2010) A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinform.*, **11** (Suppl. 3), S5.

Hastie,T. *et al*. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer, New York, pp. 75–76.

Herrmann,J. *et al*. (2009) Isomer-specific effects of CLA on gene expression in human adipose tissue depending on PPARγ P12A polymorphism: a double blind, randomized, controlled cross-over study. *Lipids Health Dis.*, **8**, 35.

Hoggart,C.J. *et al*. (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Hofmann,S.M. *et al*. (2007) Adipocyte LDL receptor-related protein1 expression modulates postprandial lipid transport and glucose homeostasis in mice. *J. Clin. Invest.*, **117**, 3271–3282.

Hudson,R.R. (2002) Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, **18**, 337–378.

Li,C. and Li,M. (2008) GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics*, **24**, 140–142.

Li,K.-C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

Lulianella,A. *et al.* (2008) Cux2 (Cutl2) integrates neural progenitor development with cell-cycle progression during spinal cord neurogenesis. *Development*, **135**, 729–741.

Meinshausen,N. *et al*. (2009) P-values for high-dimensional regression. *J. Am. Stat. Assoc.*, **104**, 1671–1681.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B*, **72**, 417–448.

Su,H.P. *et al*. (2002) Interaction of CED-6/GULP, an adapter protein involved in engulfment of apoptotic cells with CED-1 and CD91/low density lipoprotein receptor-related protein (LRP). *J. Biol. Chem.*, **281**, 12081–12092.

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

Wei,R. *et al*. (2009) From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLos Genet.*, **5**, e1000678.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.

Wray,N.R. *et al*. (2007) Prediction of individual risk to disease from genome-wide association studies. *Genome Res.*, **17**, 1520–1528.

Wu,T.T. and Lange,K. (2008) Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, **2**, 224–244.

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

Zeggini,E. *et al*. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.*, **40**, 638–645.

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

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

Zou,H. and Li,R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.*, **36**, 1509–1533.