University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2015 *Paper* 114

Variable Selection with False Discovery Control

Kevin He, *University of Michigan - Ann Arbor*
Yanming Li, *University of Michigan School of Public Health*
Ji Zhu, *University of Michigan School of Public Health*
Hongliang Liu, *Duke University*
Jeffrey E. Lee, *University of Texas M.D. Anderson Cancer Center*
Christopher I. Amos, *Dartmouth College*
Terry Hyslop, *Duke University*
Jiashun Jin, *Carnegie Mellon University*
Qinyi Wei, *Duke University*
Yi Li, *University of Michigan School of Public Health*

Variable Selection with False Discovery Control

Kevin He, Yanming Li, Ji Zhu, Hongliang Liu, Jeffrey E. Lee, Christopher I. Amos, Terry Hyslop, Jiashun Jin, Qinyi Wei, and Yi Li

Abstract

Technological advances that allow routine identification of high-dimensional risk factors have led to high demand for statistical techniques that enable full utilization of these rich sources of information for genome-wide association studies (GWAS). Variable selection for censored outcome data as well as control of false discoveries (i.e. inclusion of irrelevant variables) in the presence of high-dimensional predictors present serious challenges. In the context of survival analysis with high-dimensional covariates, this paper develops a computationally feasible method for building general risk prediction models, while controlling false discoveries. We have proposed a high-dimensional variable selection method by incorporating stability selection to control false discovery. Comparisons between the proposed method and the commonly used univariate and Lasso approaches for variable selection reveal that the proposed method yields fewer false discoveries. The proposed method is applied to study the associations of 2,339 common single-nucleotide polymorphisms (SNPs) with overall survival among cutaneous melanoma (CM) patients. The results have confirmed that BRCA2 pathway SNPs are likely to be associated with overall survival, as reported by previous literature. Moreover, we have identified several new Fanconi anemia (FA) pathway SNPs that are likely to modulate survival of CM patients.

Variable selection with false discovery control

Kevin He [1],Yanming Li [1], Ji Zhu [2], Hongliang Liu [3], Jeffrey E. Lee [4], Christopher I. Amos [5], Terry Hyslop [6], Jiashun Jin [7], Qinyi Wei [3] and Yi Li [1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.

[2]Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA.

[3]Department of Medicine, Duke University School of Medicine and Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina, 27710, USA.

[4]Department of Surgical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, 77030, USA.

[5]Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Hanover, NH, 03750, USA.

[6]Department of Biostatistics and Bioinformatics, Duke University; and Duke Clinical Research Institute, Durham, North Carolina, 27710, USA.

[7]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Variable selection with false discovery control

Kevin He,Yanming Li, Ji Zhu, Hongliang Liu, Jeffrey E. Lee, Christopher I. Amos, Terry Hyslop, Jiashun Jin, Qinyi Wei and Yi Li

Abstract

Technological advances that allow routine identification of high-dimensional risk factors have led to high demand for statistical techniques that enable full utilization of these rich sources of information for genome-wide association studies (GWAS). Variable selection for censored outcome data as well as control of false discoveries (i.e. inclusion of irrelevant variables) in the presence of high-dimensional predictors present serious challenges. In the context of survival analysis with high-dimensional covariates, this paper develops a computationally feasible method for building general risk prediction models, while controlling false discoveries. We have proposed a high-dimensional variable selection method by incorporating stability selection to control false discovery. Comparisons between the proposed method and the commonly used univariate and Lasso approaches for variable selection reveal that the proposed method yields fewer false discoveries. The proposed method is applied to study the associations of 2,339 common single-nucleotide polymorphisms (SNPs) with overall survival among cutaneous melanoma (CM) patients. The results have confirmed that BRCA2 pathway SNPs are likely to be associated with overall survival, as reported by previous literature. Moreover, we have identified several new Fanconi anemia (FA) pathway SNPs that are likely to modulate survival of CM patients.

1 Introduction

Rapid advances in technology that have generated vast amounts of data from genetic or genome studies have led to a high demand for developing powerful statistical learning meth-

ods for extracting information effectively. For instance, understanding clinical and pathophysiologic heterogeneities among subjects at risk and designing effective treatment for appropriate subgroups is one of the most active areas in genetic studies. Wide heterogeneities present in patients' response to treatments or therapies. Understanding such heterogeneities is crucial in personalized medicine, and discovery of genetic variants offers a feasible approach. However, serious statistical challenges arise when identifying real predictors among hundreds of thousands of candidates, and an urgent need has emerged for the development of effective algorithms for model building and variable selection.

The last three decades have given rise to many new statistical learning methods, including CART (Breiman et al., 1984), random forest (Breiman, 2001), neural networks (Bishop, 1995), SVMs (Boser et al., 1992) and high dimensional regression (Tibshirani, 1996; Tibshirani, 1997; Fan and Li, 2001; Fan and Li, 2002; Gui and Li 2005). Boosting has emerged as a powerful framework for statistical learning. It was originally introduced in the field of machine learning for classifying binary outcomes (Freund and Schapire, 1996), and later its connection with statistical estimation was established by Friedman et al. (2000). Friedman (2001) proposed a gradient boosting framework for regression settings. Bühlmann and Yu (2003) proposed a componentwise boosting procedure based on cubic smoothing splines for L2 loss functions. Bühlmann (2006) demonstrated that the boosting procedure works well in high-dimensional settings. For censored outcome data, Ridgeway (1999) applied boosting to fit proportional hazards models, and Li and Luan (2005) developed a boosting procedure for modeling potentially non-linear functional forms in proportional hazards models.

Despite the popularity of aforementioned methods, issues such as false discovery (e.g. seletion of irrelevant SNPs) and difficulty in identifying weak signals present further barriers. Simultaneous inference procedure, including the Bonferroni correction, has been widely used in large-scale testing literature. However, in many high-dimensional settings, such as

in genetic studies, variable selection is serving as a screening tool to identify a set of genetic variants for further investigation. Hence, a small number of false discoveries would be tolerable and simultaneous inference would be too conservative. In contrast, the false discovery rate (FDR), defined as the expected proportion of false positives among significant tests (Benjamini and Hochberg, 1995), is a more relevant metric for false discovery control under the framework of variable selection. However, few existing variable selection algorithms control false discoveries. This has brought an urgent need of developing computationally feasible methods that tackle both variable selection and false discovery control.

We propose a novel high-dimensional variable selection method for GWAS by improving the existing variable selection methods in several aspects. First, we have developed a computationally feasible variable selection approach for high-dimensional survival analysis. Second, we have designed a random sampling scheme to improve the control of the false discovery rate. Finally, the proposed framework is flexible to accommodate complex data structures.

The rest of the article is organized as follows. In section 2 we introduce notation and briefly review the L_1 penalized estimation and gradient boosting method that are of direct relevance to our proposal. In section 3 we develop the proposed approach, and in section 4 we evaluate the practical utility of the proposal via intensive simulation studies. In section 5 we apply the proposal to analyze a genome-wide association study of cutaneous melanoma. We conclude the paper with a brief discussion in section 6.

2 Model

2.1 Notation

Let D_i denote the time from onset of cutaneous melanoma to death and C_i be the potential censoring time for patient i, $i = 1, \ldots, n$. The observed survival time is $T_i = \min\{D_i, C_i\}$,

and the death indicator is given by $\delta_i = I(D_i \leq C_i)$. Let $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T$ be a p-dimensional covariate vector (contains all the SNP information) for the ith patient. We assume that, conditional on $\mathbf{X}_i$, D_i is independently censored by C_i. To model the death hazard, consider

$$\lambda_i(t|\mathbf{X}_i) = \lim_{dt \to 0} \frac{1}{dt} Pr(t \leq D_i < t + dt | D_i \geq t, \mathbf{X}_i) = \lambda_0(t) \exp(\mathbf{X}_i^T \boldsymbol{\beta}),$$

where $\lambda_0(t)$ is the baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is a vector of parameters. The corresponding log-partial likelihood is given by

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{X}_i^T \boldsymbol{\beta} - \log \left\{ \sum_{\ell \in R_i} \exp\left(\mathbf{X}_\ell^T \boldsymbol{\beta}\right) \right\} \right],$$

where $R_i = \{\ell : T_\ell \geq T_i\}$ is the at-risk set. The goal of variable selection is to identify $S_0 = \{j : \beta_j \neq 0\}$, which contains all the variables that are associated with the risk of death.

2.2 L_1 Penalized Estimation

Tibshirani (1997) proposed a Lasso procedure in the Cox model, e.g., estimate $\boldsymbol{\beta}$ via the penalized partial likelihood optimization

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \{ l_n(\boldsymbol{\beta}) - \lambda \|\boldsymbol{\beta}\|_1 \}, \tag{1}$$

where $\| \cdot \|_1$ is the L_1 norm. To solve (1), Tibshirani (1997) considered a penalized reweighted least squares approach. Let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ be the $p \times n$ covariate matrix and define $\boldsymbol{\eta} = \mathbf{X^T}\boldsymbol{\beta}$. Let $l_n'(\boldsymbol{\eta})$ and $l_n''(\boldsymbol{\eta})$ be the gradient and Hessian of the log-partial likelihood with respect to $\boldsymbol{\eta}$ respectively. Given the current estimator $\widehat{\boldsymbol{\eta}} = \mathbf{X^T}\widehat{\boldsymbol{\beta}}$, a two-term Taylor expansion of the log-partial likelihood leads to

$$l_n(\boldsymbol{\beta}) \approx \frac{1}{2} (z(\widehat{\boldsymbol{\eta}}) - \mathbf{X^T}\boldsymbol{\beta})^{\mathbf{T}} \mathbf{l_n''}(\widehat{\boldsymbol{\eta}}) (\mathbf{z}(\widehat{\boldsymbol{\eta}}) - \mathbf{X^T}\boldsymbol{\beta}),$$

where $z(\widehat{\boldsymbol{\eta}}) = \widehat{\boldsymbol{\eta}} - l_n''(\widehat{\boldsymbol{\eta}})^{-1} l_n'(\widehat{\boldsymbol{\eta}})$. Similar to the problem of conditional likelihood (Hastie and Tibshirani 1990), the matrix $l_n''(\widehat{\boldsymbol{\eta}})$ is non-diagonal, and solving (1) may require $O(n^3)$

computations. To avoid this difficulty, Tibshirani (1997) used some heuristic arguments to approximate the Hessian matrix with a diagonal one, e.g., treated off-diagonal elements as zero. An iteratively procedure is then conducted based on the penalized reweighed least squares

$$\frac{1}{n}\sum_{i=1}^{n} w(\widehat{\boldsymbol{\eta}})_i(\mathbf{z}(\widehat{\boldsymbol{\eta}})_\mathbf{i} - \mathbf{X_i^T}\boldsymbol{\beta})^\mathbf{2} + \lambda\|\boldsymbol{\beta}\|_\mathbf{1}, \tag{2}$$

where the weight $w(\widehat{\boldsymbol{\eta}})_i$ for subject i is the ith diagonal entry of $l_n''(\widehat{\boldsymbol{\eta}})$. However, it is unclear whether the diagonal approximation always converges to the right solution and further evaluation may be needed.

To obtain a more accurate estimation, Gui and Li (2005) used a Cholesky decomposition to obtain $A = (l_n''(\widehat{\boldsymbol{\eta}}))^{1/2}$ such that $\mathbf{A^T A} = \mathbf{l_n''}(\widehat{\boldsymbol{\eta}})$. The iterative procedure in (2) is then revised as

$$\frac{1}{n}\sum_{i=1}^{N}(\mathbf{z}^*(\widehat{\boldsymbol{\eta}})_\mathbf{i} - \mathbf{X_i^{*T}}\boldsymbol{\beta})^\mathbf{2} + \lambda\|\boldsymbol{\beta}\|_\mathbf{1},$$

where $\mathbf{z}^*(\widehat{\boldsymbol{\eta}}) = \mathbf{A}\mathbf{z}(\widehat{\boldsymbol{\eta}})$ and $\mathbf{X}^* = \mathbf{A}\mathbf{X}$. Alternatively, Geoman (2010) combined gradient descent with Newton's method and implemented his algorithm in an R package *penalized*. Both of these algorithms perform well in settings with a moderately large number of predictors. However, for GWAS studies that often present a very large number of predictors, these algorithms are not feasible.

2.3 Gradient Boosting

Gradient boosting has emerged as a powerful tool for building predictive models; its application in the Cox proportional hazards models can be found in Ridgeway (1999) and Li and Luan (2005). The idea is to pursue iterative steepest ascent of the log likelihood function. At each step, given the current estimate of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}$, let $\widehat{\boldsymbol{\eta}} = \mathbf{X}^T\widehat{\boldsymbol{\beta}}$. The algorithm computes

the gradient of the log-partial likelihood with respect to η_i, the ith component of $\boldsymbol{\eta}$,

$$U_i = \frac{\partial}{\partial \eta_i} l_n(\boldsymbol{\eta}) \Big|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}} = \delta_i - \sum_{\ell=1}^{n} \frac{\delta_\ell I(T_i \geq T_\ell) \exp(\hat{\eta}_i)}{\sum_{k=1}^{n} I(T_k \geq T_\ell) \exp(\hat{\eta}_k)},$$

for $i = 1, \cdots, n$, and then fits this gradient (also called working response or pseudo response) to $\mathbf{X}$ by a so-called base procedure (e.g. least squares estimation). Specifically, to facilitate variable selection, a componentwise algorithm can be implemented by restricting the search direction to be componentwise (Bühlmann and Yu, 2003; Li and Luan 2005). For instance, fit componentwise model

$$\widetilde{\beta}_j = \underset{\beta_j}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} (U_i - X_{ij}\beta_j)^2,$$

for $j = 1, \ldots, p$. Compute

$$j^\star = \underset{1 \leq j \leq p}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} (U_i - X_{ij}\widetilde{\beta}_j)^2.$$

and update $\hat{\beta}_{j^\star} = \hat{\beta}_{j^\star} + v\ \widetilde{\beta}_{j^\star}$, where v is a positive small constant (say 0.01) controlling the learning rate (Friedman, 2001).

This approach is to detect a componentwise direction along which the partial likelihood would ascend most rapidly. At each boosting iteration only one component of $\boldsymbol{\beta}$ is selected and updated. The variable selection can be achieved if boosting stops at an optimal number of iterations. This optimal number works as the regularization parameter and it can be determined by cross-validation (Simon et al., 2011). However, as we will show in simulation, the cross-validated choice still includes certain amount of false positive selections. A computationally feasible method is needed to control false discoveries.

2.4 Control of the False Discovery Rate (FDR)

Benjamini and Hochberg's FDR-controlling procedure (Benjamini and Hochberg, 1995), or BH's procedure for short, is a recent innovation for controlling the FDR. Consider a setting

where we test a large number of tests simultaneously. Let R be the number of total discoveries (selection of SNPs) and let V be the number of false discoveries (selection of irrelevant SNPs). If we denote the False Discovery Proportion by

$$FDP = V/R,$$

then FDR is simply the expectation of FDP. In the simplest setting (i.e., p-values associated all component tests are independent), BH's procedure is able to control the FDR at any preselect level $0 < q < 1$ (called the FDR-control parameter).

In the past 20 years, BH's procedure has inspired a great deal of research: many variants of the procedure have been proposed, and many insights and connections have been discovered. For instance, Efron (2008, 2011) and Storey (2003) have pointed out an interesting connection between the BH's procedure and the popular Empirical Bayes method. In particular, they proposed a Bayesian version of the FDR which they call the *Local FDR* (Lfdr) and showed that two versions of FDR are intimately connected to each other. Another useful variant of BH's procedure is the Significance Analysis of Microarrays (SAM; Tusher et al. 2001), a method that was originally designed to identify genes in microarray experiments. While the success of the BH's procedure hinges on an accurate approximation of the p-values associated with individual tests, SAM is comparably more flexible for it is able to handle more general experimental layouts and summary statistics, where the p-values may be hard to obtain or to approximate. See Efron (2011) for a nice review on FDR-controlling methods, Lfdr, and SAM.

3 Proposed Methods

3.1 Componentwise Gradient Boosting Procedure

To introduce the proposed method, we first consider a variant of componentwise gradient boosting method that is computationally efficient in high-dimensional settings.

Algorithm 1 (Componentwise Gradient Boosting)

Initialize $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$. For $m = 1, \cdots, M_{stop}$, iterate the following steps:

(a) For $j = 1, \ldots, p$, compute the componentwise gradient

$$G_j = \frac{\partial}{\partial \beta_j} l_n(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(m-1)}}.$$

(b) Compute $j^\star = \text{argmax}_{1\le j\le p} |G_j|$.

(c) Update $\hat{\beta}_{j^\star}^{(m)} = \hat{\beta}_{j^\star}^{(m-1)} + v\ \widetilde{\beta}_{j^\star}$, where $\widetilde{\beta}_{j^\star}$ can be estimated by one-step Newton's update

$$\widetilde{\beta}_{j^\star} = \left\{ \frac{\partial^2}{\partial \beta_{j^\star}^2} l_n(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(m-1)}} \right\}^{-1} \frac{\partial}{\partial \beta_{j^\star}} l_n(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(m-1)}}.$$

(d) Iterate until $m = M_{stop}$ for some stopping iteration M_{stop}.

Under the chain rule of differentiation, Algorithm 1 is equivalent to the traditional boosting procedure we described in Section 2.3, which first computes the working response, U_i, and then fits the working response to each covariate by least squares. In contrast, Algorithm 1 is based on gradient with respect to $\boldsymbol{\beta}$ and it avoids the calculation of working response. Such a componentwise update is connected with a minimization-maximization (MM) algorithm (Hunter and Lange, 2004; Lange 2012). For instance, in a minorization step, given the mth step estimate $\widehat{\boldsymbol{\beta}}^{(m-1)}$, an application of Jensen's inequality leads to the following

minority surrogate function

$$\begin{aligned} l_n(\boldsymbol{\beta}) \;\geq\; & \sum_{j=1}^{p}\sum_{i=1}^{n}\alpha_j\delta_i\left[\frac{X_{ij}}{\alpha_j}(\beta_j-\hat{\beta}_j^{(m-1)})+\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}^{(m-1)}\right. \\ & \left.-\log\left\{\sum_{\ell\in R_i}\exp\left(\frac{X_{\ell j}}{\alpha_j}(\beta_j-\hat{\beta}_j^{(m-1)})+\mathbf{X}_\ell^T\widehat{\boldsymbol{\beta}}^{(m-1)}\right)\right\}\right] \\ & = g(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(m-1)})=\sum_{j=1}^{p}g(\beta_j|\widehat{\boldsymbol{\beta}}^{(m-1)}), \end{aligned}$$

where $g(\beta_j|\widehat{\boldsymbol{\beta}}^{(m-1)})$ is defined implicitly, all $\alpha_j \geq 0$, $\sum_j \alpha_j = 1$ and $\alpha_j > 0$ whenever $X_{ij} \neq 0$. In the maximization step, we maximize (or monotonically increase) the selected component of the surrogate function to produce the next iteration estimators, e.g., consider $g(\beta_{j^\star}|\widehat{\boldsymbol{\beta}}^{(m-1)})$ and update $\beta_{j^\star}$. Then the boosting algorithm monotonically increase the original log-partial likelihood by increasing the surrogate functions. Note that as long as the ascent property is achieved, the choice of α_j is not crucial, e.g., it can be considered as part of a control for step size. Moreover, as one only needs to increase the surrogate function instead of maximizing it, one-step Newton iterations (with step-size control) shall provide sufficient and rapid updates at each boosting step. Instead of using $\widetilde{\beta}_{j^\star}$, an alternative approach is to use the normalized updates with norm normalized to be 1, e.g., $\hat{\beta}_{j^\star}^{(m)} = \hat{\beta}_{j^\star}^{(m-1)} + v \times sign(G_j)$. Its main disadvantage is that its performance is sensitive to the choice of learning rate. Although $sign(G_j)$ provides an ascent direction, a sufficiently small step length may be needed. Empirically we found that the procedure with fitted $\widetilde{\beta}_{j^\star}$ provides better performance.

It is known that finding the proper regularization parameter is very difficult for the Lasso procedure, especially for survival settings for which piece-wise linear solution path (LARS; Efron et al., 2004) is not available and a grid search (Simon et al. 2011) is required. In contrast, in boosting procedure, the number of iteration works as tuning parameter and the optimal choice is less critical as boosting is more robust to overfitting (Hastie et al., 2009).

3.2 Boosting with Stability Selection for False Discovery Control

Stability Selection was recently introduced by Meinshausen and Bühlmann (2010) as a general technique designed to improve the performance of a variable selection algorithm. The idea is to identify variables that are included in the model with high probabilities when a variable selection procedure is performed on randomly sampled of the observations. For completeness of exposure, we summarize the procedure of stability selection as follows. Let $\mathbf{I}$ be a random subsample of $\{1, \cdots, n\}$ of size $\lfloor n/2 \rfloor$, draw without replacement. Here $\lfloor n/2 \rfloor$ is defined as the largest integer not greater than $n/2$. For variable $j \in \{1, \cdots, p\}$, the random sampling probability that the jth variable is selected by the stability selection is

$$\widehat{\Pi}_j = Pr^*[j \in \hat{S}(\mathbf{I})],$$

where $\hat{S}(\mathbf{I}) = \{j : \hat{\beta}_j^{(\mathbf{I})} \neq 0\}$ denotes the variable selected by the variable selection procedure based on the subsample $\mathbf{I}$, and the empirical probability Pr^* is with respect to the random sampling. For a threshold $\Pi_{thres} \in (0, 1)$, the set of variables selected by stability selection is then defined as

$$\hat{S}_{stable} = \{j : \widehat{\Pi}_j \geq \Pi_{thres}\}.$$

A particularly attractive feature of stability selection is that its relatively insensitive to the tuning parameter (e.g., M_{stop} for boosting) and hence cross-validation can be avoided. However, a new regularization parameter needs to be determined is the threshold Π_{thres}. To address this question, an error control was provided by an upper bound on the expected number of falsely selected variables (Meinshausen and Bühlmann, 2010; Theorem 1). More formally, let $E|\hat{S}(I)|$ be the expected number of selected variables and define V to be the number of falsely selected variables. Assume an exchangeable condition, then the expected

number V of falsely selected variables is bounded for $\Pi_{thres} \in (0.5, 1)$ by

$$E[V] \leq \frac{1}{2\Pi_{thres} - 1} \frac{(E|\hat{S}(I)|)^2}{p}.$$

Based on such a bound, the tuning parameter Π_{thres} can be chosen such that $E[V]$ is controlled at the desired level, e.g., for $E[V] < 1$, if $E[V] < p^{\frac{1}{2}}$,

$$\Pi_{thres} = \left(1 + \frac{(E[V])^2}{p}\right)/2. \tag{3}$$

The property of the above procedure relies on restricted assumptions such as exchangeability condition (e.g., the joint distribution of outcomes and covariates is invariant under permutations of noninformative variables), which, as noted by Meinshausen and van de Geer (2011), are not likely to hold for real data. In GWAS with extensive correlation structure among SNP markers, the exchangeability condition fails and using threshold in (3) has been shown to suffer a loss of power (Alexander and Lange, 2011). Moreover, in computing the threshold in (3), we face a tradeoff. Commonly used variable selection procedures will select certain amount of false positives. On one hand, we want $E[V]$ to be large to select the true informative predictors, but on the other hand, a large $E[V]$ also can render Π_{thres} large (which leads to too conservative threshold). If $E[V] > p^{\frac{1}{2}}$, we cannot control the error $E[V]$ with the formula in (3).

To improve the performance of stability selection and determine a data-driven threshold for the selection frequency, we adopt the idea of SAM (Tusher et al. 2001) and propose a random permutation based stability selection boosting procedure.

Algorithm 2 (Boosting with Stability Selection and Permutation)

(a) For $s = 1, \cdots, 100$, we draw random subsample of the data of size $\lfloor n/2 \rfloor$. On the sth subsample, implement the proposed boosting approach (e.g., Algorithm 1). Record the set of selected predictors at the sth subsampling, $\hat{S}^{(s)} = \{j : \hat{\beta}_j^{(s)} \neq 0\}$, and compute

$\widehat{\Pi}_j = \frac{1}{S}\sum_{s=1}^{S} I(j \in \hat{S}^{(s)})$, where $I(A)$ is an indicator function taking the value 1 when condition A holds and 0 otherwise.

(b) For $b = 1, \cdots, B$, randomly permute the outcomes so that the relation between covariates and outcomes is decoupled. Repeat the stability-based boosting described in step (a) on the permuted sample and record the set of selected predictors $\tilde{S}^{(b)}$, and compute $\widetilde{\Pi}_j^b = \frac{1}{S}\sum_{s=1}^{S} I(j \in \tilde{S}^{(b)})$.

(c) Order the values of $\widehat{\Pi}_j$ for $1, \cdots, p$, and let $\widehat{\Pi}_{(j)}$ be the jth largest value. Likewise let $\widetilde{\Pi}_{(j)}^{(b)}$ be the jth largest value of $\widetilde{\mathbf{\Pi}}^{(b)} = (\widetilde{\Pi}_1^{(b)}, \cdots, \widetilde{\Pi}_p^{(b)})$.

(d) Define $\widetilde{\Pi}_{(j)} = \sum_{b=1}^{B} \widetilde{\Pi}_{(j)}^{(b)}/B$.

(e) Define the estimated empirical Bayes false discovery rate (Efron 2011) corresponding to the jth largest $\widehat{\Pi}_{(j)}$ as

$$\overline{Fdr}_{(j)} = \min\left\{\frac{1}{B}\frac{\sum_{b=1}^{B}\sum_{j=1}^{p} I(\widetilde{\Pi}_j^{(b)} \geq \widehat{\Pi}_{(j)})}{\sum_{j=1}^{p} I(\widehat{\Pi}_j \geq \widehat{\Pi}_{(j)})}, 1\right\}.$$

(f) For a pre-specified value $q \in (0, 1)$, calculate a data-driven threshold

$$\widehat{\Pi}_{thres}(q) = \min\{\widehat{\Pi}_{(j)} : \overline{Fdr}_{(j)} \leq q\}.$$

Then this $\widehat{\Pi}_{thres}(q)$ can be used to determine the selected variables. If $q = 0.2$ and 5 variables are selected with selection frequency greater than $\widehat{\Pi}_{thres}(0.2)$, then 1 of these 5 variables would be expected to be false positive.

4 Simulations

Finite-sample properties of the proposed method were evaluated through a series of simulation studies. Death times were generated from the exponential model, $\lambda(t|\mathbf{X}_i) = 0.5\exp(\mathbf{X}_i^T\boldsymbol{\beta})$

for $i = 1, \ldots, n$, where $\mathbf{X}_i = (X_{i1}, \cdots, X_{i2000})^T$ came from multivariate normal distributions. These 2,000 predictors were in 10 blocks with equal numbers of predictors within each block. We considered three simulation schemes with within-block correlation coefficients varying between 0.2, 0.5 and 0.8. For all three schemes, the between-block correlation coefficients were 0 (i.e., independent between blocks). We chose 10 true signals; one from each block, with true β in $\pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 2.5$. All other covariate effects are zero. Censoring times were generated from uniform distributions, with the percentage of censored subjects then being approximately 20-30%. Each data configuration was replicated 100 times.

We compared the proposed methods, Lasso for proportional hazard models (Simon et al., 2011), univariate approaches with either Bonferroni correction (termed Univariate Bonferroni in Table 1) or Benjamini and Hochberg (1995)'s procedure for FDR control (below a threshold 0.2; termed Univariate FDR in Table 1). For the boosting approach without stability control (Algorithm 1), 10-fold cross-validation was implemented to determine the optimal stopping iteration. For the boosting approach with stability selection (Algorithm 2), we repeatedly drew 100 random subsamples of the data of size $\lfloor n/2 \rfloor$. The maximum selection frequency on one permutation data was used as threshold for variable selection. Table 1 shows that the boosting without stability selection (termed Boosting in Table 1) outperform the univariate approaches in the average number of false positives (FP), average false discovery proportion (Fdp), average number of false negative (FN) and the empirical probabilities to identify the true signal (Power). Though the Lasso has comparable performances in terms of FN and Power, the FPs of the boosting methods are substantially fewer than the Lasso. Finally, the proposed boosting method with stability selection and permutation (termed S-Boosting in Table 1) further reduce the FPs.

Summary of Simulation Results

Correlation	Methods	FP	Fdp	FN	Power
0	Univariate Bonferroni	0.01	0	2.28	0.77
	Univariate FDR	1.94	0.18	1.49	0.85
	Lasso	185.22	0.95	0	1
	Boosting	15.76	0.61	0	1
	S-Boosting	0.01	0	0	1
0.5	Univariate Bonferroni	85.29	0.92	2.32	0.77
	Univariate FDR	172.32	0.95	0.81	0.92
	Lasso	186.17	0.95	0	1
	Boosting	22.31	0.69	0	1
	S-Boosting	0.03	0	0	1
0.8	Univariate Bonferroni	131.42	0.94	2.17	0.78
	Univariate FDR	207.52	0.96	0.68	0.93
	Lasso	185.14	0.95	0	1
	Boosting	29.25	0.75	0	1
	S-Boosting	0.12	0.01	0	1

FP: the average number of false positives; Fdp: false discovery proportion; FN: average number of false negative; Power: the empirical probabilities to identify the true signal

5 Application of Cutaneous Melanoma Data

Cutaneous melanoma (CM) is one of the most aggressive skin cancers, causing the greatest number of skin cancer related deaths worldwide. Among the CM patients, wide heterogeneities are present. The commonly used clinicopathological variables, such as tumor stage and Breslow thickness (Balch et al, 2009), may have insufficient discriminative ability (Schramm and Mann, 2011). Discovery of genetic variants would offer a feasible approach to understanding mechanisms that may affect clinical outcomes and the sensitivity of individual cancer to therapy (Liu et al., 2012; Li et al., 2013; Rendleman et al., 2013). We applied our proposed procedures to a genome-wide association study reported by Yin et al. (2014) to analyze the association of 2,339 common single-nucleotide polymorphisms (SNPs) with overall survival in CM patients. Our goal was to identify SNPS that are relevant to overall

survival among the patients.

The dataset contains a total of 858 CM patients, with 133 deaths observed during the follow-up, where the median follow-up time was 81.1 months. The overall survival time was calculated from the date of diagnosis to the date of death or the date of the last follow-up. Genotyped or imputed common SNPs (minor allele frequency ≥ 0.05, genotyping rate $\geq 95\%$, Hardy-Weinberg equilibrium P-value ≥ 0.00001, and imputation $r^2 \geq 0.8$) within these genes or their $\pm$20-kb flanking regions were selected for association analysis (Yin et al. 2014). As a result, 321 genotyped SNPs and 2,018 imputed SNPs in the FA pathway were selected for further analysis. Other covariates to adjust for included age at diagnosis, Clark level, tumor stage, Breslow thickness, sentinel lymph node biopsy, and the mitotic rate.

The proposed boosting procedure with stability selection was implemented to select informative SNPs (coded as 0, 1; without or with minor alleles). The importance of predictors is evaluated by the proportion of times that the predictor is selected in the model among the 100 subsamples. We also compared the proposed methods with the Lasso, the boosting procedure without stability selection and univariate approaches. The results are summarized in Table 2. The Lasso procedure selected 25 SNPs. Among them, 12 SNPs with absolute coefficients larger than 0.01 are listed in Table 2. None of these predictors pass the univariate approaches with Bonferroni correction or Benjamini and Hochberg (1995)'s procedure for FDR control (with a threshold 0.2). As we found in section 4, these results argue that the univariate approaches may have more false negatives than other methods. In contrast, the boosting procedure selected 7 predictors, which were a subset of top 12 SNPs selected by the Lasso. To further control the false selections, the estimated false discovery rate, $\overline{Fdr}$, were also calculated to determine a data-driven threshold for the selection frequency such that $\overline{Fdr} \leq 0.2$. Three of the SNPs selected by both Lasso and boosting pass the threshold $\widehat{\Pi}_{thres}(0.2) = 72\%$. The remaining variables find insignificant support from stability

selection.

Figure 1 shows the stability path (selection frequencies across boosting iterations). The variables with selection frequencies larger than the threshold (estimated empirical Bayes false discovery rate $\overline{Fdr} \leq 0.2$; based 500 permuted samples) are plotted as solid lines, while the path of the remaining variables are shown as broken lines. The top 3 variables stand out clearly and the number of boosting iteration is less critical. A Manhattan plot was given in Figure 2 with the dashed horizontal line corresponding to the estimated threshold $\widehat{\Pi}_{thres}(0.2) = 72\%$. Three variables have selection frequencies larger than this dashed horizontal line. The vertical blue lines highlight the selection frequencies of the four previously-detected SNPs that are associated with overall survival of CM patients by Yin et al. 2014. The red vertical lines highlight the SNPs whose selection frequencies pass the estimated threshold. The lower panel of Figure 2 illustrates pairwise correlations across the $2,339$ SNPs with the strength of the correlation, from positive to negative, indicated by the color spectrum from red to dark blue. One of the top SNPs in our finding, rs74189161 (with selection frequency = 72% and $\overline{Fdr} = 0.16$) is strongly correlated with rs3752447 identified by Yin et al. (2014), with correlation coefficients $r^2 = 1$ (calculated with plink v1.07; Purcell et al., 2007). Besides confirming the previously reported SNP, we also found some novel signals. For example, we identified a cluster of signals around SNP rs356665 in gene FANCC and a SNP rs3087374 in gene FANC1. Both two genes have previously been reported having regulation effects with the FA pathway (Kao et al., 2011; Ella et al., 2012; Jenkins et al., 2012). Mutations in the FA pathway are identified in diverse cancer types (Hucl and Gallmeier 2011) and therefore are likely to modulate the survival of CM patients.

Summary of selected SNPs by Lasso (sorted by the magnitude of coefficients; only predictors with absolute coefficients larger than 0.01 are included), their estimated coefficients by boosting without stability selection, P-values based on univariate approach, selection frequencies based on stability selection.

SNPs	Chromosome	Gene	$\hat{\beta}_{Lasso}$	$\hat{\beta}_{Boosting}$	P-value	Frequency (%)
rs74189161	13	*BRCA2*	-0.11	-0.10	0.002	72*
rs356665	9	*FANCC*	-0.09	-0.04	0.03	88*
rs11649642	16	*FANCA*	-0.08	-0.05	0.01	27
rs9567670	13	*BRCA2*	-0.07	-0.03	0.01	51
rs8081200	17	*BRIP1*	-0.06	-0.02	0.05	38
rs3087374	15	*FANC1*	-0.06	-0.01	0.02	73*
rs35322368	9	*FANCC*	0.06	0	0.03	65
rs57119673	16	*FANCA*	-0.04	-0.01	0.03	54
rs8061528	16	*BTBD12*	-0.03	0	0.12	36
rs2247233	15	*FANC1*	0.02	0	0.15	39
rs848286	2	*FANCL*	0.02	0	0.02	23
rs62032982	16	*PALB2*	0.01	0	0.04	34

$\hat{\beta}_{Lasso}$: coefficients from Lasso; $\hat{\beta}_{Boosting}$: coefficients from boosting; P-value: calculated from univariate approach; Frequency (%): selection frequencies across 100 subsampling; $\overline{Fdr}$: estimated empirical Bayes false discovery rate (based 500 permuted samples); the false discovery control of the predictors under stability selection are coded by (*) to indicate that the selection frequencies pass the $\overline{Fdr}$ threshold.

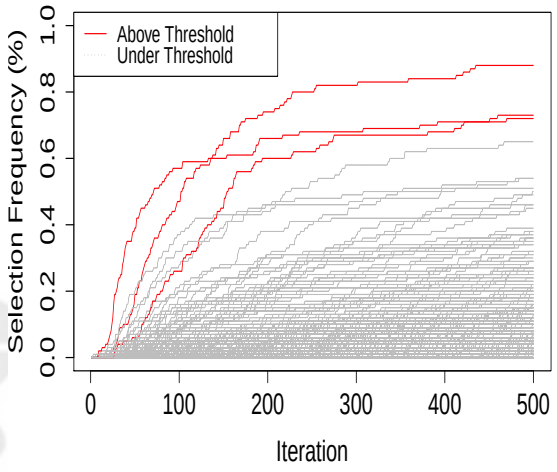

Figure 1: Selection Path: selection frequencies across 500 boosting iterations; Threshold: estimated empirical Bayes false discovery rate $\overline{Fdr} \leq 0.2$ (based 500 permuted samples)

6 Discussion

Reducing the number of false discoveries is often very desirable in biological applications since follow-up experiments can be costly and laborious. We have proposed a boosting method with stability selection to analyze high-dimensional GWAS data. We demonstrated and compared performances of the proposed method and the commonly used univariate approaches or Lasso for variable selection. The proposed method outperformed other methods in terms of substantially reduced false positives and low false negatives.

Finally, it is worth mentioning that gradient descent (ascent) works in flexible parameter spaces, even including infinite-dimensional functional spaces. In the latter case, as the search space is typically a functional space, one needs to calculate the Gâteaux derivative of the functional in order to determine the optimal descent direction. We will report the work elsewhere.

Acknowledgement

The work is partially supported by an NIH grant.

References

Balch, C. M. and Gershenwald, J. E. and Soong, S. J. and et al. (2009). Final version of 2009 AJCC melanoma staging and classification. *Journal of Clinical Oncology*, **27**, 6199-206.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300.

Bishop, C. (1995) *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

Boser, B. E. and Guyon, I. M. and Vapnik, V. N. (1992) A training algorithm for optimal margin classifiers. *In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 144-152.

Breiman, L. and Friedman, J. and Olshen, R. and Stone, C. (1984) *Classification and Regression Trees*, Wadsworth, New York.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer.

Bühlmann, P., and Yu, B. (2003) Boosting with the $L2$ loss: Regression and classification *Journal of the American Statistical Association*, **98 (462)**, 324-339.

Bühlmann, P., and Yu, B. (2006) Boosting for high-dimensional linear models *Annals of Statistics*, **34**, 559-583.

Bühlmann, P., and Hothorn, T, (2007) Boosting algorithms: Regularization, prediction and model fitting *Statistical Science*, **22(4)**, 477-505.

Efron, B., and Hastie, T. and Johnstone, I. and Tibshirani, R. (2004) Least angle regression *Annals of Statistics*, **32(2)**, 407-499.

Efron B (2008). Microarrays, empirical Bayes and the two groups model. *Statistical Science*, **23**, 122.

Efron, B. (2012) Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction *Institute of Mathematical Statistics Monographs*, Cambridge University Press.

Ella, R., Thompson, M.A., Doyle, G.L., Ryland, S.M., Rowley, D.Y., H. Choong et al. (2012) Exome Sequencing Identifies Rare Deleterious Mutations in DNA Repair Genes FANCC and BLM as Potential Breast Cancer Susceptibility Alleles. *PLOS genetics.*, DOI: 10.1371/journal.pgen.1002894

Fan, J., and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties *Journal of the American Statistical Association*, **96 (456)**, 1348-1360.

Fan, J., and Li, R. (2002) Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, **30(1)**, 74-99.

Freund, Y., and Schapire, R. (1996) Experiments with a new boosting algorithm *Machine Learning: Proceedings of the Thirteenth International Conference*, Morgan Kauffman, San Francisco, 148-156.

Friedman, J.H., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion) *Annals of Statistics*, **28(2)**, 337-407.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine *Annals of Statistics*, **29(5)**, 1189-1232.

Gui, J., and Li, H. (2005) Penalized cox regression analysis in the high-dimensional and low-sample size settings with application to microarray gene expression data. *Bioinformatics*, **21**, 3001-3008.

Hastie, T. and Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York

Hucl, T., and Callmeier, E.(2010) DNA Repair: Exploiting the Fanconi Anemia Pathway As a Potential Therapeutic Target. *Physiological Research*, **60**, 453-465.

Hunter, D. R. and Lange, K. (2004) A tutorial on MM algorithms *The American Statistician*, **58**, 30-37.

Jenkins, C., Kan, J., and Hoatlin, M.E. (2012) Targeting the Fanconi Anemia Pathway to Identify Tailored Anticancer Therapeutics. *Anemia.*, Article ID 481583. http://dx.doi.org/10.1155/2012/481583

Kao, W.H., Riker, A.I., Kushwaha, D.S., Ng, K., Enkemann, S.A., Jove, R., Buettner, R., Zinn, P.O., Sanchez, N.P., Villa, J.L., d'Andrea, A.D., Sanchez, J.L., Kennedy, R.D., Chen, C.C., Matta, J.L. (2011) Upregulation of Fanconi anemia DNA repair genes in melanoma compared with non-melanoma skin cancer. *Journal of Investigative Dermatology*, **131(10)**, 2139-2142.

Lange, K.. Optimization, 2nd Edtition. *Springer Texts in Statistics,* Springer, New York (2012)

Li, H. and Luan, Y. (2005) Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data *Bioinformatics*, **21,** 2403-2409.

Liu, H. and Wei, Q and Gershenwald, J. E. and et al. (2012) Influence of single nucleotide polymorphisms in the MMP1 promoter region on cutaneous melanoma progression, *Melanoma Research*, **22** 169-75.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion) *Journal of the Royal Statistical Society: Series B*, **72**, 417-473.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, **81(3)**, 559-575.

Rendleman, J. and Shang, S. and Dominianni, C. and et al. (2013) Melanoma risk loci asdeterminants of melanoma recurrence and survival, *Journal of Translational Medicine*, **11(279)**,

Ridgeway, G. (1999) The State of Boosting *Computing Science and Statistics*, **31,** 172-181.

Schramm, S. J. and Mann, G. J. (2011) Melanoma prognosis: a REMARK-based systematic review and bioinformatic analysis of immunohistochemical and gene microarray studies *Molecular Cancer Therapeutics*, **10,** 1520-1528.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011) Regularization paths for Cox's proportional hazards model via coordinate descent *Journal of Statistical Software*, **39(5),** 1-13.

Storey, John D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, **31 (6**, 20132035.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16(4),** 385–395.

Tusher, V.G. et al. (2001) Significane analysis of microarrays applied to the ionizing radiation repsonse. *Proceedings of the National Academy of Sciences USA*, **98,** 5116-5121.

Yin, J., Liu, H., Liu, Z., Wang, L.E., Chen, W,V., Zhu, D., Amos, C.I., Fang, S., Lee, J.E., and Weo, Q. (2014) Genetic Variants in Fanconi Anemia Pathway Genes BRCA2 and FANCA Predict Melanoma Survival. *Journal of Investigative Dermatology.*

Zhao, D.S., and Li, Y. (2010), Principled Sure Independence Screening for Cox Models With Ultra-High-Dimensional Covariates, manuscript, Harvard University

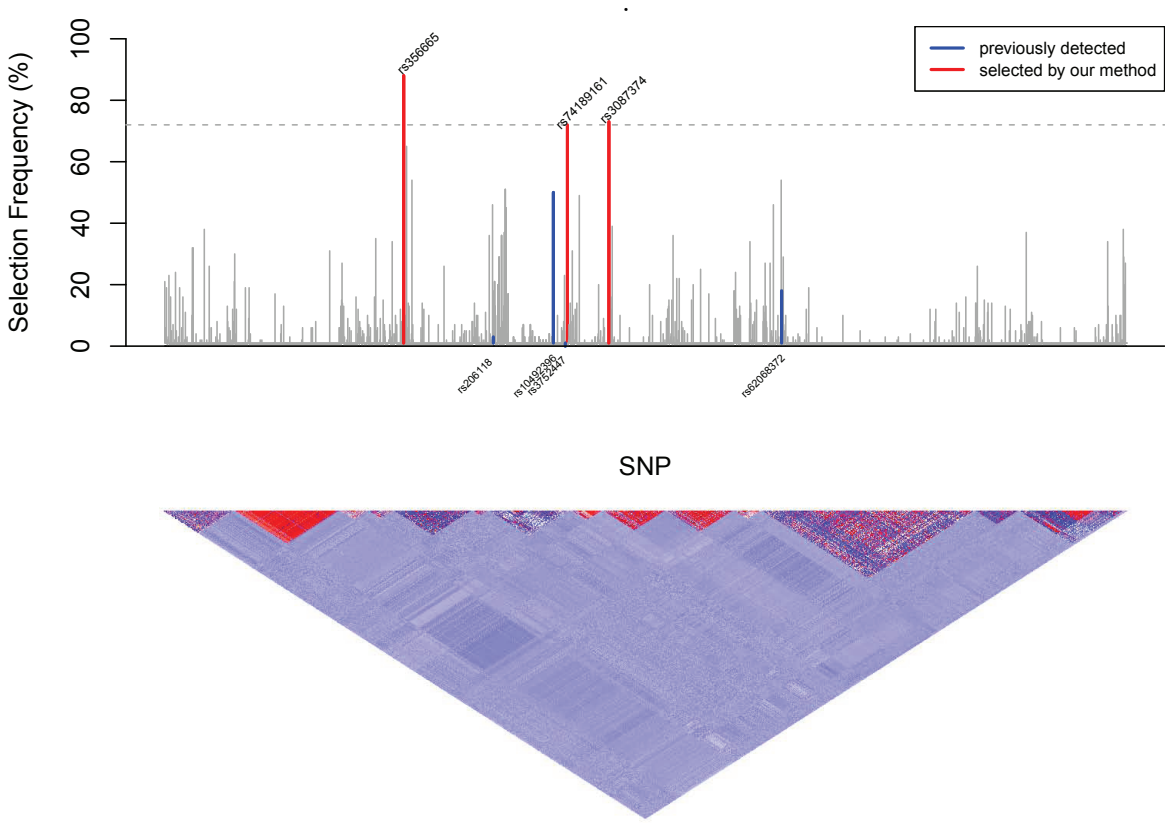

Figure 2: Manhattarn Plot for Selection Frequency (%); dashed horizontal line: estimated threshold $\widehat{\Pi}_{thres}(0.2) = 72\%$; vertical blue lines: selection frequencies of the four previously-detected SNPs that are associated with overall survival of CM patients by Yin et al. 2014; red vertical lines: the SNPs whose selection frequencies pass the estimated threshold; the lower panel: pairwise correlations across the $2,339$ SNPs with the strength of the correlation, from positive to negative, indicated by the color spectrum from red to dark blue