Biomedical Signal Processing and Control 91 (2024) 105965

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Controlled variable selection with nonconvex regularization for identifying biomarkers

Shoujiang Li [a], Hui Zhang [b,*], Yong Liang [c]

[a] Department of Engineering Science, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, 999078, China
[b] School of Mathematics and Data Science, Shaanxi University of Science and Technology, Xi'an, 710021, China
[c] Peng Cheng Laboratory, Shenzhen, 518005, China

ARTICLE INFO

Keywords:
False discovery rate
Nonconvex regularization
Knockoff filter
Variable selection

ABSTRACT

Biomedical big data has revolutionized biomarker identification and has become a key driver for the development of precision medicine applications. However, existing computational methods have been able to rapidly identify biomarkers (variable selection), but true validation of biomarkers is still hampered by low statistical power and poor reproducibility of results. To address the above issues, in this paper, we propose two knockoff-based nonconvex regularization methods for identifying biomarkers. These two methods can perform variable selection while rigorously guaranteeing the false discovery rate (FDR) at a given desired level with high statistical power. We combine two nonconvex regularization methods, Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP), with the knockoff framework, respectively. Knockoff variables are first constructed to mimic the correlation structure of the original variables while maintaining independence from the response, and then the original and knockoff variables are used as augmentation matrices for variable selection. Since the nonconvex regularization method has good statistical theoretical properties such as unbiasedness, sparsity and Oracle, the proposed methods are better able to deal with heavy-tailed distributions, high noise and high correlation data. We verify the effectiveness of the proposed methods through numerical simulation experiments, and the results show that the proposed methods have strong statistical power while controlling the FDR compared to the comparison baseline method. We also apply the proposed methods to identify Human Immunodeficiency Virus (HIV) drug resistance-related gene mutations, Alzheimer's disease brain lesion regions, and purity-related genes in tumor samples, which can provide references and help for clinical diagnosis and treatment.

1. Introduction

Biomarkers typically refer to indicators that can be objectively measured and evaluated to reflect physiologic or pathologic processes, as well as biological effects on exposure or therapeutic interventions [1]. Precision medicine can rely on validated biomarkers to better classify patients based on their probable disease risk, prognosis and/or response to treatment. With the development of high-throughput technology, biomedical big data [2] has brought great revolution to the identification of biomarkers and has become a key driver in the development of precision medicine applications [3]. However, the full potential of big data cannot be mined without computational or statistical methods to carry out the process of recognizing reliable patterns and extracting useful information. Dramatic increases in computational power and resource availability have driven the development of such approaches over the past few decades [4], and now the discovery and identification of biomarkers from big data, either through computational or statistical methods, is increasingly becoming one of the key technologies in precision medicine approaches [5].

However, clinically validated biomarkers represent a very small percentage of biomarkers that have been discovered or have been reported in the literature. As of 2010, only 24 cancer biomarkers had been approved by the FDA [6], and of the 1,261 biomarker proteins cited in the literature, approximately only 5% have been studied in greater depth, with less than 3% used in the clinic [7]. And another paper noted that there are more than 150,000 publications documenting thousands of biomarkers, but no more than 100 have been validated in actual clinical practice [8].

The reasons for this may be, on the one hand, the under-promotion of the application of current computational methods to more complex real-world medical data [9], and on the other hand, the massive amount of data that increases the burden of the multiple testing problem, the

[*] Corresponding author.
E-mail address: zhanghui9@sust.edu.cn (H. Zhang).

https://doi.org/10.1016/j.bspc.2024.105965
Received 13 November 2023; Received in revised form 13 January 2024; Accepted 28 January 2024
Available online 1 February 2024

fact that the real data has a more complex correlation as well as an increase in the computational cost [10]. Although existing computational methods have been able to rapidly identify putative biomarkers, true validation of biomarkers is still hampered by low statistical power and poor reproducibility of results.

To reduce time-consuming and expensive experimental validation, researchers require more reliable biomarkers that contain few false positives. Generally, the identification of biomarkers is statistically referred to as variable selection or feature selection, also known as "discoveries". Accordingly, the false discovery rate (FDR) [11] was developed as a statistical criterion to ensure the reliability of discoveries. Existing methods of controlling for FDR typically rely on p-values to compute tests of variable significance. The Benjamini–Hochberg (BH) method [11] tests for the significance of a variable by calculating the p-values that satisfy collateral conditions such as positive correlation or independence [12]. However, these conditions are difficult to ensure, especially in the nonlinear or high-dimensional case. The Benjamini–Yekutieli (BY) [13] method ensures FDR control when the p-value has an arbitrary form of dependence. However, it may suffer from statistical power loss compared to the BH method. Therefore, it cannot guarantee that FDR control is at the target or desirable level, which limits the wide application of FDR control in healthcare big data.

The knockoff framework [14] is a recent statistical breakthrough aimed at controlling FDR under arbitrary correlation structures and without computing p-values. The key idea of the knockoff framework is to construct knockoff variables that mimic the correlation structure of the original variables but are independent of the response conditions of the original variables, i.e., the knockoff variables are similar to the original variables in terms of correlation structure, but given the original variables they are conditionally independent of the response outcome. Thus, knockoff variables can be used as negative controls for the original covariates in order to separate the true variables from the redundant or noisy ones, thus ensuring FDR control. In contrast to the well-known BH, the knockoff framework appropriately accounts for arbitrary correlations between the original variables while ensuring FDR control. Furthermore, it is not limited to the use of calibrated p-values and can be flexibly applied to feature importance scores computed based on a variety of machine learning methods with rigorous finite-sample statistical guarantees. The model-X knockoff [12] is shown to ensure that the FDR is controlled at a given desired level in any dimension and in any dependency structure between the variable and the response. Candès et al. tested model-X knockoff in conjunction with statistics from the Least Absolute Shrinkage and Selection Operator (LASSO) [15] for high-dimensional nonparametric conditional modeling, showed superior performance through numerical experiments, and obtained twice as many findings as the original analyses in a study of Crohn's disease.

LASSO, also known as L_1 regularization, is a convex optimization problem which can be solved by convex optimization tools with sparsity and is widely used in the study of various variable selection problems. However, in practice LASSO usually fails to produce the sparsest solution and does not handle error data with heavy-tailed distributions well. In addition, LASSO may lead to the selection of redundant or noisy variables due to the overshrinking of the model's nonzero coefficients.

To address the above issues, in this paper, we propose two knockoff-based nonconvex regularization methods and apply them to identify biomarkers. These two methods are able to perform variable selection while rigorously ensuring that the FDR is at a given desired level with high statistical power. We combine two nonconvex regularization methods, Smoothly Clipped Absolute Deviation (SCAD) [16] and Minimax Concave Penalty (MCP) [17], with the knockoff framework, respectively. Compared with the LASSO method, the proposed methods have good statistical theoretical properties, such as unbiasedness, sparsity, and Oracle, and are able to deal with the heavy-tailed distributions, highly noisy, and highly correlated data in a better way. We design a series of numerical simulation experiments to analyze the effectiveness of the proposed methods. Further, we apply the proposed methods to biomarker studies of Human Immunodeficiency Virus (HIV) drug resistance-related gene mutations, Alzheimer's disease brain lesion regions, and purity-related gene identification in tumor samples. The source code is available from https://github.com/shoujiang/knockoff_nonconvex.

2. Methods and materials

2.1. The variable selection problem

The general variable selection problem is the following. Suppose that there are p potential explanatory variables $\boldsymbol{X} = (X_1, \dots, X_p)$ and an observed response Y. Given n samples, we need to know which predictors (variables) are important for the response Y.

We assume that, conditionally on the predictors, the responses are independent and the conditional distribution of Y_i only depends on its corresponding vector of predictors. Formally, We denote the following conditional distribution.

$$Y_i|(X_{i,1}, \dots, X_{i,p}) \overset{\text{ind.}}{\sim} F_{Y|X}, \quad i = 1, \dots, n. \tag{1}$$

The variable selection problem is motivated by the belief that, in many practical applications, $F_{Y|X}$ actually only relies on a (small) subset $\mathcal{S} \subset \{1, \dots, p\}$ of the predictors, such that Y is independent of all other variables, conditionally on $\{X_j\}_{j\in\mathcal{S}}$. This is a very intuitive definition, that can be informally restated by saying that the other variables are not important because they do not offer any extra information on Y. The smallest set $\mathcal{S}$ with this property is often referred to as a Markov blanket [18]. The variables of set $\{X_j\}_{j\in\mathcal{S}}$ are truly significant for the response and those $\{X_j\}_{j\notin\mathcal{S}}$ are truly not significant for the response. Under very mild conditions on $F_{Y|X}$, this can be shown to be unique and the variable selection problem is cleanly defined. In order to avoid any ambiguity in those pathological cases in which the Markov blanket is not unique, we refer to as the jth predictor is *null* if and only if Y is independent of X_j, conditionally on all other predictors $X_{-j} = \{X_1, \dots, X_p\} \setminus \{X_j\}$. We denote the subset of null variables by $\mathcal{S}^0 \subset \{1, \dots, p\}$ and call the jth variable relevant (or non-null) by $\mathcal{S}$ if $j \notin \mathcal{S}^0$.

We aim to identify the set $\mathcal{S}$ with a theoretically guaranteed false discovery rate. More precisely, we are able to select the variables while ensuring that the FDR is under target level $\alpha \in (0, 1)$. Here, the FDR is defined as

$$\mathrm{FDR} = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{S}^0|}{|\hat{\mathcal{S}}| \vee 1}\right]. \tag{2}$$

where $\hat{\mathcal{S}}$ denotes a subset selected from the observed data $(\boldsymbol{X}, Y)$. Controlling the FDR is actually to control Type I error. The power is another measurement of the performance of variable selection procedure, which is the expectation of the proportion of the selected true variables among the true variables. The power is defined as

$$\mathrm{Power} = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{|\mathcal{S}|}\right]. \tag{3}$$

The aim of study is to achieve high power for variable selection with exactly guaranteed FDR control.

2.2. Nonconvex regularization based on model-X knockoffs

We study the SCAD and MCP based on the model-X knockoffs for the variable selection problem, which theoretically guaranteed FDR control and meanwhile achieving high power.

According to the assumptions of variable selection problem in Section 2.1, we consider the general linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon, \tag{4}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the random matrix of explanatory potential variables, $\boldsymbol{y} \in \mathbb{R}^n$ is a responses vector, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown coefficients vector, $\epsilon \in \mathbb{R}^p$ is the vector of random errors, and n denotes the sample size and p denotes the dimensionality of variable. Suppose that $\boldsymbol{\beta}$ is sparse and k is the number of nonzero $\boldsymbol{\beta}$, $k \ll n$, $k \ll p$, that is, we hope to select significant variables set $\mathcal{S}$, k is the size of $\mathcal{S}$. Specifically, $\beta_j \neq 0$ for some (unknown) index $j \in \mathcal{S}$ and $\beta_j = 0$ for all $j \in \mathcal{S}^0$.

The nonconvex regularization methods provide an effective way to solve the above problem, which have the following form,

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + P(\boldsymbol{\beta};\lambda) \right\}. \tag{5}$$

where $(\boldsymbol{y}, \boldsymbol{X})$ represents a data set, $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ is a square loss function. The regularization parameter $\lambda \in (0, +\infty)$ controls the complexity of model and thus alleviates the overfitting problem that tends to arise due to high dimensionality of the variables and small sample size, i.e., $n \ll p$. The $P(\boldsymbol{\beta};\lambda)$ represents the penalty function. Here, we consider SCAD and MCP penalty function. Statistically, Fan and Li [16] theoretically proved that SCAD possesses sparsity, unbiasedness, continuity and Oracle properties of variable selection and overcomes some limitations of L_1 regularization. The SCAD penalty is defined as

$$\begin{aligned} P_{\mathrm{SCAD}}(\beta;\lambda) = \lambda|\beta| I_{\{0\le|\beta|\le\lambda\}} + \left(\frac{(a-1)\lambda^2}{2} + \lambda^2\right) I_{\{|\beta|\ge a\lambda\}} \\ + \left(\frac{2a\lambda(|\beta|-\lambda) - (|\beta|^2-\lambda^2)}{2(a-1)} + \lambda^2\right) I_{\{\lambda\le|\beta|\le a\lambda\}}, \end{aligned} \tag{6}$$

where a is a constant, for some $a > 2$. MCP was proposed by Zhang et al. [17] for variable selection of high-dimensional data. MCP satisfies the Oracle property of variable selection, i.e., the correct model is selected consistently and the estimation of parameters satisfies asymptotic normality. The MCP penalty is expressed as

$$P_{\mathrm{MCP}}(\beta;\lambda) = \lambda \int_0^{\beta} (1 - \frac{x}{\gamma\lambda})_+ dx, \tag{7}$$

where $(1 - \frac{x}{\gamma\lambda})_+ = \max\{1 - \frac{x}{\gamma\lambda}, 0\}$, for some $\gamma > 1$. Both SCAD and MCP penalty function were performed in an attempt to avoid excessive penalization for coefficients β of model. However, SCAD and MCP cannot guarantee the FDR control while performing variable selection.

Then, we utilize the model-X knockoffs to construct knockoff variables $\tilde{\boldsymbol{X}}$ for the original variables $\boldsymbol{X}$. The construction of model-X knockoff variables is the key ingredient of the model-X knockoff procedure proposed by Candès et al. [12]. We construct the model-X knockoff variable defined as follows

Definition 1. The model-X knockoff variables $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ of a set of random variables $\boldsymbol{X} = (X_1, \ldots, X_p)$ is required to satisfy the following properties:

- For any subset $\mathcal{S} \subset \{1, \ldots, p\}$

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\mathrm{swap}(\mathcal{S})} \overset{d}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}]. \tag{8}$$

$[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ denotes that matrix $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are concatenated by columns. $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\mathrm{swap}(\mathcal{S})}$ is obtained by swapping the entries X_j and $\tilde{X}_j$ for any $j \in \mathcal{S}$ and $\overset{d}{=}$ denotes equal in distribution.

- The response Y should be conditionally independent of the knockoff variable $\tilde{X}_j$ given the original variable X_j.

The same distribution of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\mathrm{swap}(\mathcal{S})}$ and $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ is equivalent to the fact that they have the same first two moments, i.e., the same mean and covariance.

Suppose that $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$, then the covariance of $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ is defined as

$$\mathrm{cov}(\boldsymbol{X}, \tilde{\boldsymbol{X}}) = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{s\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{s\} & \boldsymbol{\Sigma} \end{bmatrix}, \tag{9}$$

where s is chosen to obtain a positive semidefinite covariance matrix. Theoretically, we can generate model-X knockoff variables satisfying the above properties by using the Sequential Conditional Independent Pairs (SCIP) algorithm proposed by Candès et al. in [12]. However, since the SCIP algorithm generates knockoff variables depending on the exact distribution, the computational complexity is rather high. Therefore, we utilize an approximate construction method called the approximate semidefinite program (ASDP) to generate the model-X knockoff variable. The ASDP program construction is as follows,

Step 1. Choose an approximation $\boldsymbol{\Sigma}_{\mathrm{approx}}$ of $\boldsymbol{\Sigma}$ and solve:

$$\begin{array}{ll} \text{minimize} & \sum_j \left|1 - \hat{s}_j\right| \\ \text{subject to} & \hat{s}_j \ge 0 \\ & \mathrm{diag}\{\hat{s}\} \preceq 2\boldsymbol{\Sigma}_{\mathrm{approx}} \end{array} \tag{10}$$

Step 2. Solve:

$$\begin{array}{ll} \text{maximize} & \gamma \\ \text{subject to} & \mathrm{diag}\{\gamma\hat{s}\} \preceq 2\boldsymbol{\Sigma} \end{array} \tag{11}$$

and set $s^{\mathrm{ASDP}} = \gamma\hat{s}$.

We concatenate the original variables $\boldsymbol{X}$ and the generated model-X knockoff variables $\tilde{\boldsymbol{X}}$ into a new $n \times 2p$ augmented matrix $[\boldsymbol{X}, \tilde{\boldsymbol{X}}] \in \mathbb{R}^{n\times 2p}$ by columns. Then we reconsider and rewrite the nonconvex regularization (SCAD and MCP) form of (5) so that $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$ replaces $\boldsymbol{X}$ as follows

$$\hat{\boldsymbol{\beta}}^{\mathrm{aug}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{2p}} \left\{ \frac{1}{2}\|\boldsymbol{y} - [\boldsymbol{X}, \tilde{\boldsymbol{X}}]\boldsymbol{\beta}\|_2^2 + P(\boldsymbol{\beta};\lambda) \right\}. \tag{12}$$

Using the data set $([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$, we solve the nonconvex optimization problem of Eq. (12) and obtain the model solution as $\hat{\boldsymbol{\beta}}^{\mathrm{aug}}(\lambda) = (\hat{\boldsymbol{\beta}}_1^{\mathrm{aug}}(\lambda), \ldots, \hat{\boldsymbol{\beta}}_{2p}^{\mathrm{aug}}(\lambda))$. In order to select the significant variables, we measure the importance of variables by constructing the statistics V. There are various methods for constructing the knockoff statistics for each variable X_j, such as $|\hat{\beta}_j^{\mathrm{aug}}(\lambda)| - |\hat{\beta}_{j+p}^{\mathrm{aug}}(\lambda)|$, $\log(|\hat{\beta}_j^{\mathrm{aug}}(\lambda)|)$-$\log(|\hat{\beta}_{j+p}^{\mathrm{aug}}(\lambda)|)$ or $\mathrm{sign}(|\hat{\beta}_j^{\mathrm{aug}}(\lambda)| - |\hat{\beta}_{j+p}^{\mathrm{aug}}(\lambda)|)\max\{|\hat{\beta}_j^{\mathrm{aug}}(\lambda)|, |\hat{\beta}_{j+p}^{\mathrm{aug}}(\lambda)|\}$.

We use the following statistic V_j to measure the difference between the magnitude of coefficients of original variables and the corresponding knockoff variables.

$$V_j = |\hat{\beta}_j^{\mathrm{aug}}(\lambda)| - |\hat{\beta}_{j+p}^{\mathrm{aug}}(\lambda)| \tag{13}$$

It satisfies the requirements of the sign-flip property described in [12]. Given by a data-dependent knockoff threshold

$$\Gamma = \min\left\{ t > 0 : \frac{\#\{j : V_j \le -t\}}{\#\{j : V_j \ge t\}} \le \alpha \right\} \tag{14}$$

where $\alpha \in [0, 1]$. Candès et al. [12] have proved that the model-X knockoffs control the FDR under a desired α for finite sample size and any dimensionality. We select the set $\hat{\mathcal{S}}$ of significant variables by $\hat{\mathcal{S}} = \{j : V_j > \Gamma\}$ with controlling the FDR level α.

Intuitively, the reason why this procedure can control the FDR is that the sign of null V_j's can be shown to be the result of an independent coin flips, by the exchangeability of the null $|\hat{\beta}_j^{\mathrm{aug}}(\lambda)|$ and $|\hat{\beta}_{j+p}^{\mathrm{aug}}(\lambda)|$, conditional on the absolute value $|V|$. Therefore, it can be shown that the adaptive threshold Γ is a conservative estimate of proportion of false discoveries.

3. Simulation studies

In this section, we conduct numerical experiments to investigate the performance of the proposed methods that can control the false discovery rate well in the procedure of variable selection on the simulated datasets. We will compare the performance of SCAD and MCP based on model-X knockoffs with several other methods for variable selection, which are LASSO with knockoff filter, SCAD regularization method, MCP regularization method, and random forests method [19]. We will first describe the model setup and simulation settings in detail, and then compare and analyze the experimental results.

3.1. Simulation designs and settings

3.1.1. Effect of signal amplitude

The signal magnitude is defined as the magnitude of the nonzero coefficient $\boldsymbol{\beta}$ in a linear model, also known as the coefficient magnitude. In the model Eq. (5) for identifying biomarkers, the signal magnitude indicates the correlation between the selected biomarker and the response outcome. We consider the effect of varying the signal amplitude A, the row of the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}$ is drawn from i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, we normalize each column of $\boldsymbol{X}$.

Gaussian linear model. We simulate the response $\boldsymbol{y}$ from Gaussian linear model, that is, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ is generated by randomly selecting k nonzero coefficients and randomly positive and negative signs. The element of the model error $\epsilon \in \mathbb{R}^n$ is drawn from $\mathcal{N}(0, 1)$.

Binomial linear model. We link potential variables and responses in a nonlinear fashion. More specifically, we assume the following binomial linear model to link variables and potential responses $\boldsymbol{y} = (y_1, \dots, y_p)^\top$,

$$\Pr\left(y_i = 1 \mid \boldsymbol{x}_i\right) = \frac{\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)} \tag{15}$$

where the coefficient vector $\boldsymbol{\beta}$ is the same as above the settings of Gaussian linear model.

3.1.2. Effect of variable correlation

When we consider the effect of varying the variable correlation, the row of the design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is drawn from i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Theta}_\rho)$, where $(\boldsymbol{\Theta}_\rho)_{jk} = Cov(\boldsymbol{x}_j, \boldsymbol{x}_k) = \rho^{|j-k|}$, $j, k = 1, \dots, p$, denotes the correlation and then we normalize each column of $\boldsymbol{X}$. $Cov(\boldsymbol{x}_j, \boldsymbol{x}_k)$ denotes the covariance of $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ and is defined as $Cov(\boldsymbol{x}_j, \boldsymbol{x}_k) = E[(\boldsymbol{x}_j - E(\boldsymbol{x}_j))(\boldsymbol{x}_k - E(\boldsymbol{x}_k))]$, where $E(\cdot)$ is the expectation.

In this simulation, we simulate the response vector $\boldsymbol{y}$ from the same Gaussian linear model and binomial linear model as in Section 3.1.1. The remaining settings are also the same as those in the previous Section 3.1.1.

3.1.3. Simulation settings

The desired target FDR level $\alpha = 0.1$ in all simulations. We consider $n > p$ for the low-dimensional setting, where the sample size $n = 3000$, the variables dimension $p = 1000$, and $n < p$ for the high-dimensional setting, where $n = 500$ and $p = 1000$. For simulations that $\boldsymbol{y}$ comes from the Gaussian linear model, the $k = 60$ is the number of nonzero coefficients $\boldsymbol{\beta}$. For simulations that $\boldsymbol{y}$ comes from the binomial linear model, $k = 60$ is in low-dimensional settings and $k = 40$ is in high-dimensional settings.

For effect of signal amplitude, the correlation coefficient $\rho = 0$, that is, the variables are independent of each other and the varying signal amplitude A is set differently for different simulations, refer to the caption of each simulation for details.

For effect of variable correlation, the varying correlation coefficient $\rho = (0.1, 0.2, \cdots, 0.8)$ in all simulations and the signal amplitude A has different fixed values for different simulations.

3.2. Simulation results

We generate simulated datasets to calculate the mean FDR and mean power for each method over 30 repetitions respectively. For the illustration purposes, we denote the SCAD, MCP, and LASSO methods based on model-X knockoffs as Knockoff SCAD, Knockoff MCP and Knockoff LASSO, respectively.

3.2.1. Analysis of effect of signal amplitude

With the effect of varying signal amplitude, the FDR and power variation curves of various methods are shown in Figs. A.12–2.

Fig. A.12 shows the power and FDR curves as the coefficient amplitude is varied under the Gaussian linear model and low-dimensional settings. Although the power of SCAD and MCP are relatively high, the FDR of these methods is very high as well. When we apply the knockoff filter, we can see from Fig. A.12 that all methods successfully control FDR at the desired level. It can be seen that the power of the Knockoff SCAD and Knockoff MCP is slightly higher than that of the Knockoff LASSO with guaranteed control of the FDR.

In Gaussian linear model and the high-dimensional setting, the results are shown in Fig. 1. Similarly, the FDR of SCAD and MCP greatly exceeded the desired FDR level ($\alpha = 0.1$), and all knockoff methods indeed control the FDR, especially, the power of Knockoff SCAD and Knockoff MCP are higher than Knockoff LASSO when coefficient amplitude $A \geq 4.5$.

Fig. A.15 shows the performance of different methods in binomial linear model and low-dimensional settings. The results show that the power of Knockoff SCAD is higher than Knockoff LASSO for all the coefficient amplitude, and all knockoff methods indeed control the FDR. The power of Knockoff MCP is higher than Knockoff LASSO when coefficient amplitude $A \geq 7$. Similarly, although the power of SCAD and MCP methods are relatively high, their FDR is high as well. In other words, SCAD and MCP completely failed to control the FDR.

As shown in Fig. 2, all knockoff methods indeed control the FDR at the target level in binomial linear model and high-dimensional settings. The results also show that the power of Knockoff SCAD and Knockoff MCP are higher than Knockoff LASSO for all the coefficient amplitude. In particular, for all the coefficient amplitude, the power of Knockoff MCP is particularly much higher than Knockoff LASSO.

3.2.2. Analysis of effect of variable correlation

With the effect of varying the variable correlation, the FDR and power variation curves of various methods are shown in Figs. A.13–4.

Figs. A.13 and 3 show the performance of the various methods in the Gaussian linear model for low and high dimensional settings, respectively. Fig. A.13 shows that Knockoff SCAD, Knockoff MCP and Knockoff LASSO have the same performance in low-dimensional settings. Similarly, neither SCAD nor MCP select variables under the target FDR control. Fig. 3 shows that Knockoff SCAD and Knockoff MCP continue to be much more powerful than other methods with the consideration of variable dependencies in high-dimensional settings. It can be seen that the performance of SCAD and MCP is the same as Fig. A.13.

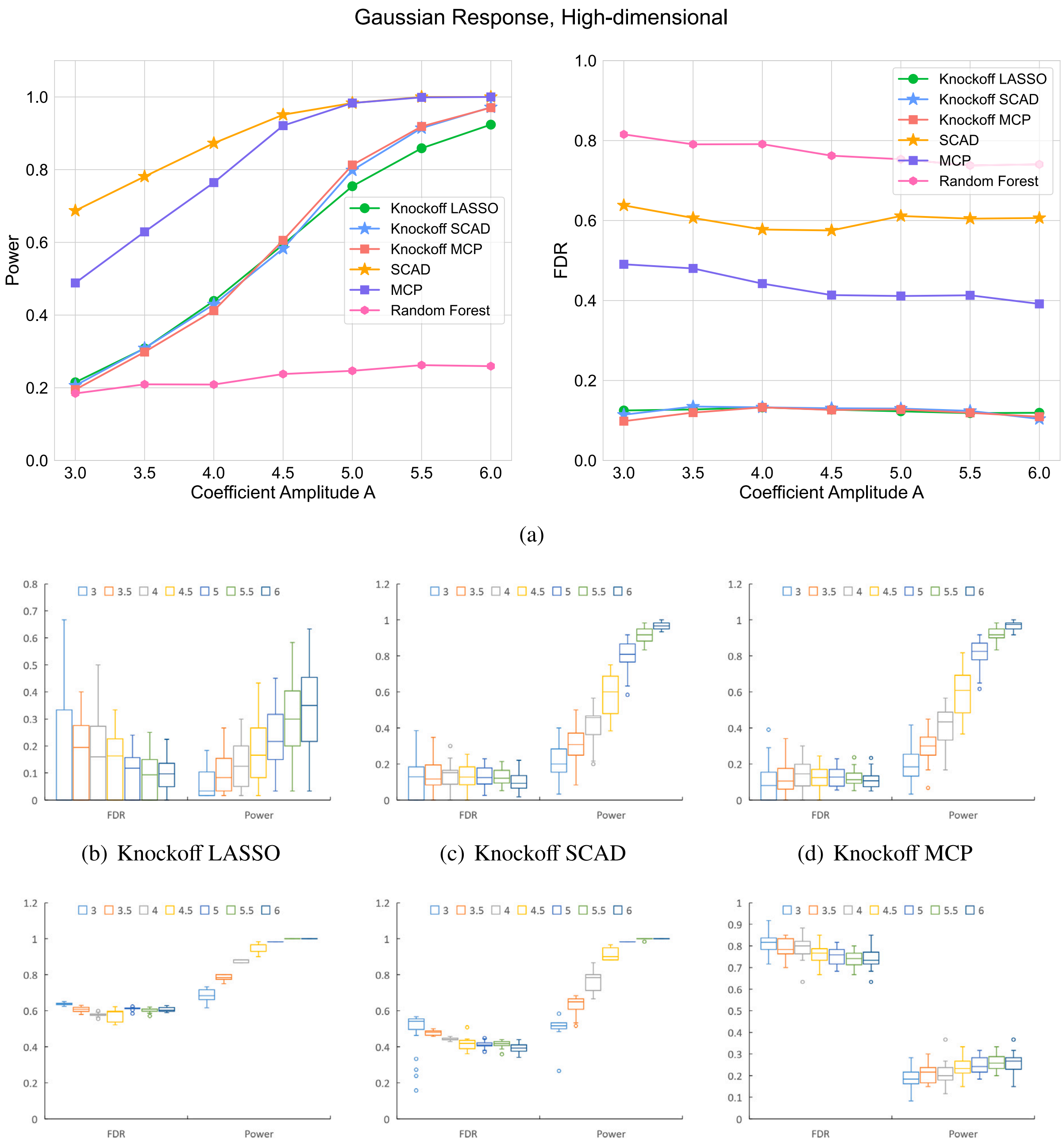
Figs. A.14 and 4 are devoted to the case of binomial linear model. Fig. A.14 shows that when $\rho < 0.4$, the power of Knockoff SCAD and Knockoff MCP are higher than Knockoff LASSO, and when $\rho > 0.4$, the power of them is almost the same for low-dimensional settings. Fig. 4 shows that the power of Knockoff SCAD and Knockoff MCP are higher than Knockoff LASSO when $\rho < 0.7$ in high-dimensional settings. Same as other simulation settings, SCAD and MCP still cannot control target FDR in both low and high dimensional settings.

However, for random forests, coefficient magnitudes and variable correlations are not significant for both FDR and Power, but FDR for random forests is consistently large while statistical Power is small, thus suggesting that random forests are not able to control for FDR and thus obtain high statistical power.

4. Real data applications

4.1. Application to HIV-1 drug resistance

In this section, we apply the proposed methods (Knockoff SCAD and Knockoff MCP) to HIV-1 data [20] to identify mutations associated with drug resistance.



(a)

(b) Knockoff LASSO (c) Knockoff SCAD (d) Knockoff MCP

(e) SCAD (f) MCP (g) Random Forest

Fig. 1. Comparison of Power and FDR for various variable selection procedures with varying signal amplitude in the high-dimensional setting of the Gaussian linear model.

4.1.1. *Dataset and settings*

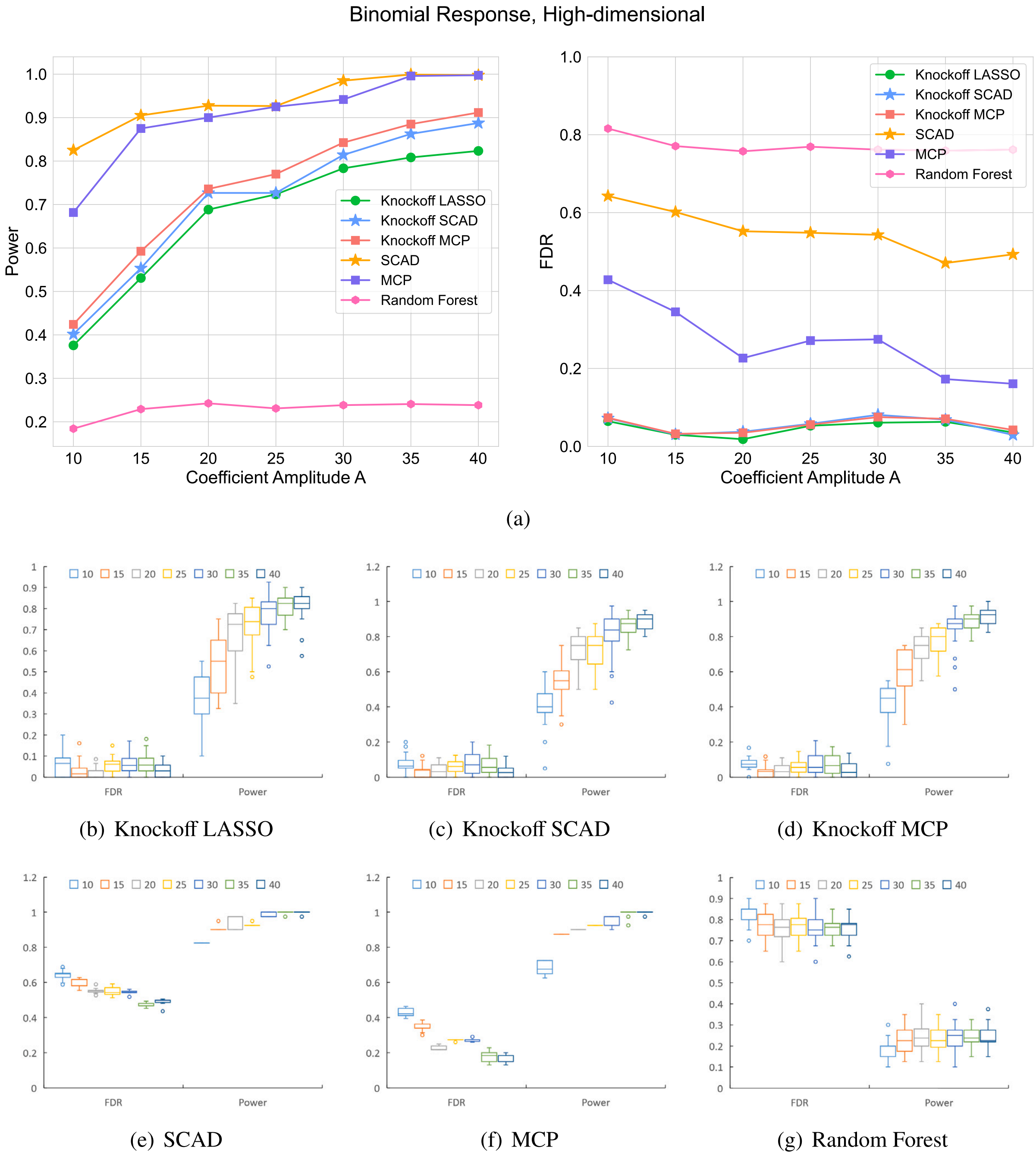
The dataset includes genotype information and drug resistance measurements. There are three different classes of drugs, protease inhibitors (PIs), nucleoside reverse-transcriptase inhibitors (NRTIs), and nonnucleoside reverse transcriptase inhibitors (NNRTIs) each with a corresponding dataset. Details of the three datasets are shown in Table 1.

We remove samples with missing resistance information during data preprocessing while retaining only those mutations that had more than three in the sample, as shown in the Table 1. The variable X_i is a marker for the presence or absence of the ith mutation, and the log-transformed drug resistance level is the response $\boldsymbol{y}$. Rhee et al. [21] created the treatment-selected mutation (TSM) panel containing mutations associated with the treatment of each class of drugs. In fact, TSM is a good approximation of ground truth and can be used to evaluate our methods. We compare the mutations selected by the proposed methods with TSM. For each drug class, as long as the ith mutation is selected for any of the drugs in that class, we considered it as a discovery.

Table 1
The description of HIV-1 drug resistance datasets.

Drug type	Numbers of drugs	Sample size n	Mutations appearing ≥ 3 times in sample
PI	7	846	209
NRPI	6	634	287
NNRPI	3	745	319

In the selection procedure, we desire target FDR level $\alpha = 0.2$. We note the selected mutations in the TSM list as true discoveries and those not in the TSM list as false discoveries. We perform the experiment 100



(a)

(b) Knockoff LASSO

(c) Knockoff SCAD

(d) Knockoff MCP

(e) SCAD

(f) MCP

(g) Random Forest

Fig. 2. Comparison of Power and FDR for various variable selection procedures with varying signal amplitude in the high-dimensional setting of the binomial linear model.

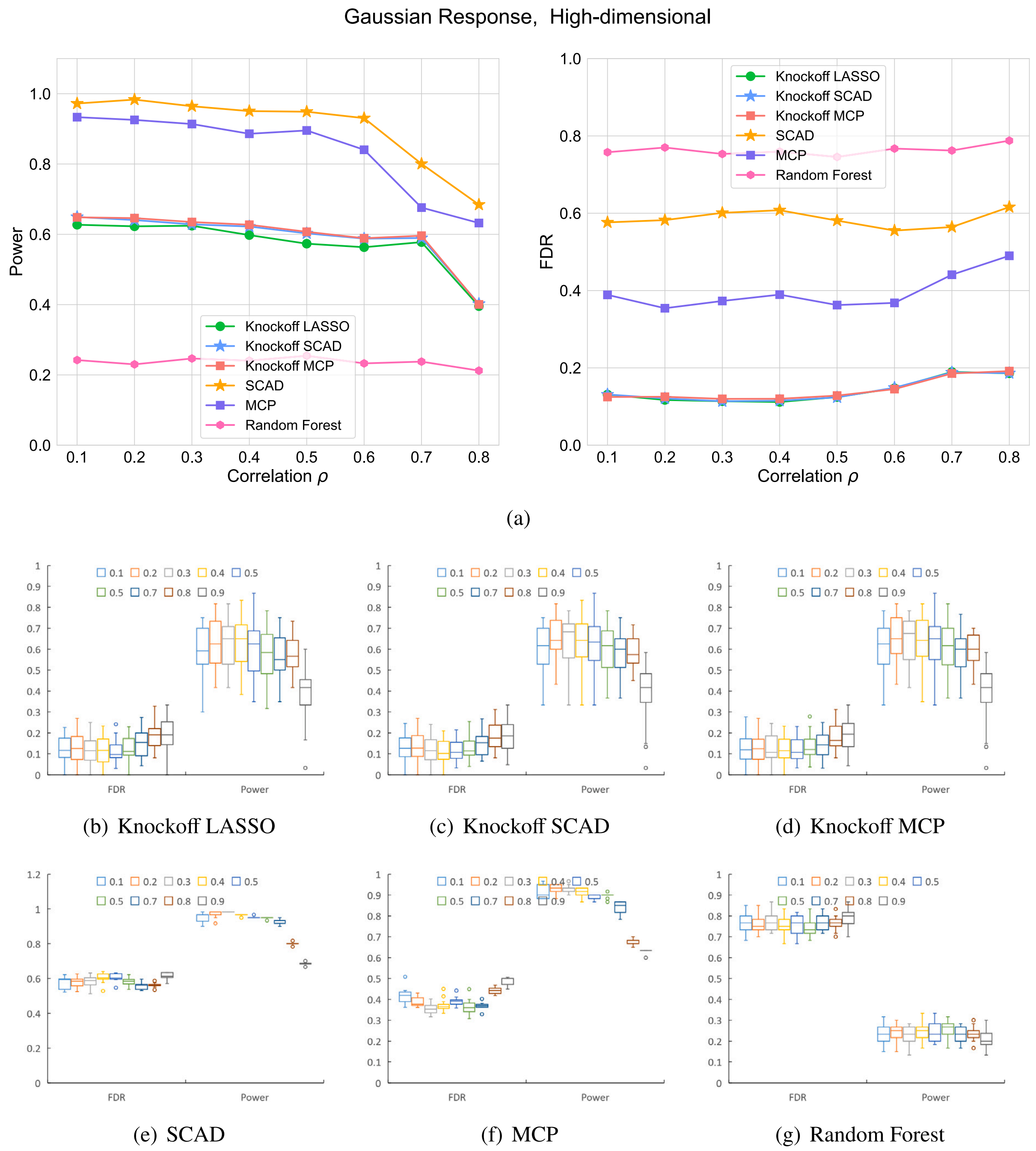
times separately and then compare the results of the different methods by the mean of the number of true discoveries and false discoveries.

4.1.2. Analysis of experimental results

We compare the performance of the three methods (Knockoff SCAD, Knockoff MCP and Knockoff LASSO) applied to different types of HIV drug resistance-associated mutations in experiments. Figs. 5, 6, and 7 summarize the mutations selected by all methods for the three different types of drugs PI, NRTI and NNRTI, respectively. Compared to Knockoff LASSO, Knockoff SCAD and Knockoff MCP, two nonconvex regularization based on knockoffs methods, obtained better performance with much better controlled proportion of false discoveries in 15 out of 16 cases. As can be seen from Fig. 5, especially in the analysis of both APV and ATV cases, the proportion of false discoveries has been quite low. In summary, the proposed methods show slightly better agreement with TSM lists compared to Knockoff LASSO.

4.2. Application to Alzheimer’s Disease

In this application, we apply the proposed methods to study lesion regions of brains in Alzheimer’s Disease (AD), which is an irreversible



(a)

(b) Knockoff LASSO

(c) Knockoff SCAD

(d) Knockoff MCP

(e) SCAD

(f) MCP

(g) Random Forest

Fig. 3. Comparison of Power and FDR for various variable selection procedures varying variable correlation in the high-dimensional setting of the Gaussian linear model. Here the signal amplitude $A = 4.5$.

neurodegenerative brain disease and the most common form of dementia in the elderly population. With the increasing aging of the population, AD has attracted a lot of attention in recent years.

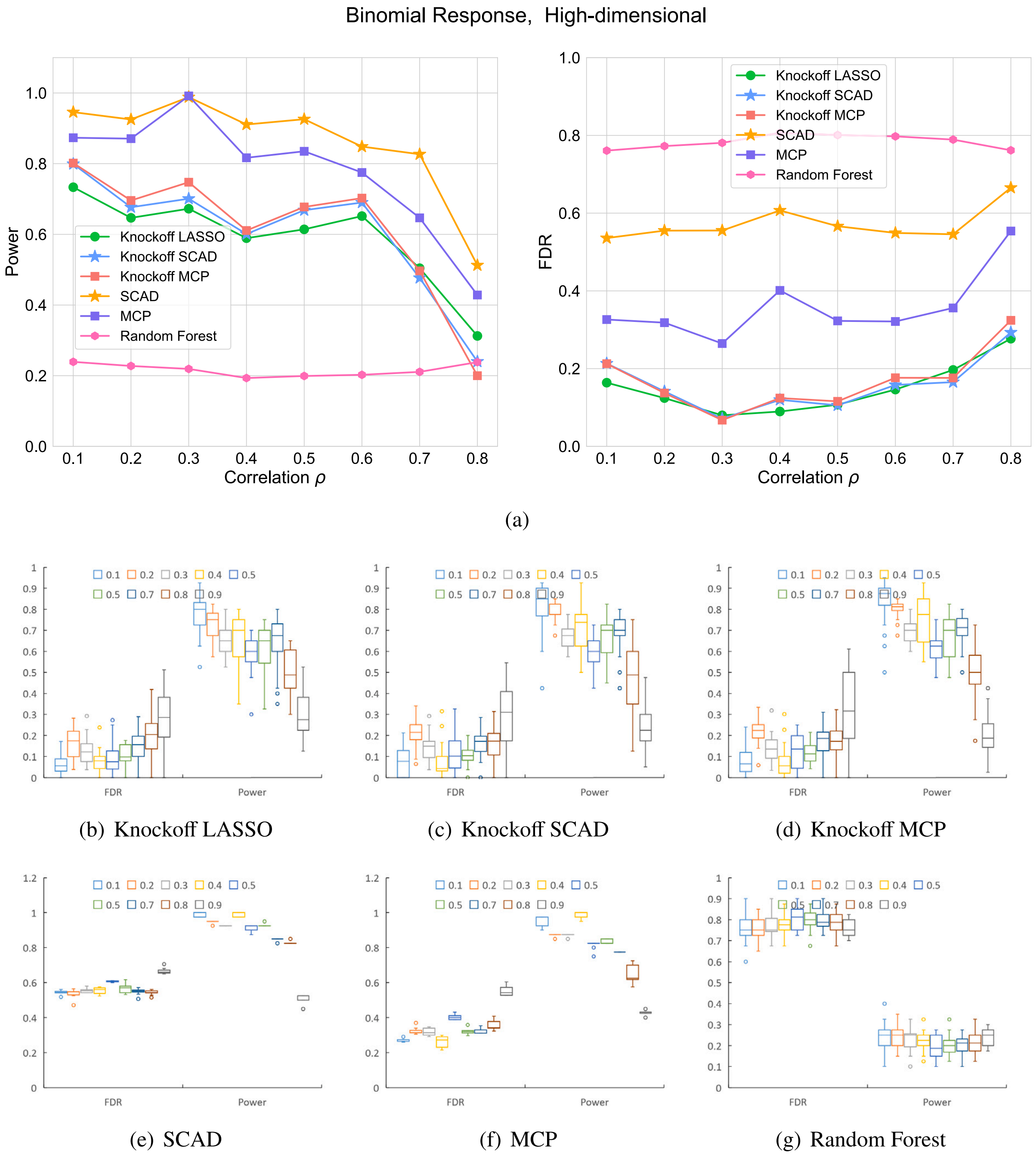
4.2.1. Structural magnetic resonance imaging dataset and settings

We obtain the dataset acquired by structural Magnetic Resonance Imaging (MRI) scan from ADNI.[1] The dataset consist of 126 AD, 433 Mild Cognitive Impairment (MCI) and 193 Normal Controls (NC), for a total of 752 samples. We first pre-processed each image with the Dartel VBM [22] and then segmented the gray matter (GM), white matter (WM) and cerebral spinal fluid (CSF) using the toolbox Statistical Parametric Mapping (SPM).[2] Finally, the entire brain was divided into 90 Cerebrum anatomical regions using Automatic Anatomical Labeling (AAL) atlas [23], and the sum of all GMs within each region was provided as its volume.

The variable X_i denotes the volume of the ith region that is normalized. The Alzheimer's Disease Assessment Scale(ADAS) was initially used by Rosen et al. [24] to assess the severity of cognitive dysfunction and was later found by Zec et al. [25] to be able to clinically differentiate between patients with AD and normal controls. We use ADAS as the response $\boldsymbol{y}$. The desired target FDR α equal to 0.2. We applied the Knockoff SCAD, Knockoff MCP and Knockoff LASSO methods to

[1] http://adni.loni.ucla.edu

[2] https://www.fil.ion.ucl.ac.uk/spm/



(a)

(b) Knockoff LASSO

(c) Knockoff SCAD

(d) Knockoff MCP

(e) SCAD

(f) MCP

(g) Random Forest

Fig. 4. Comparison of Power and FDR for various variable selection procedures varying variable correlation in the high-dimensional setting of the binomial linear model. Here the signal amplitude $A = 30$.

identify lesion regions associated with Alzheimer's Disease. We perform 100 repetitions of the experiment and we consider the lesion region is selected if it has been selected more than half the times.

4.2.2. Analysis of region selection

Table 2 shows the results of the lesion regions of the brain selected by various methods. The (L) represents the left brain and (R) represents the right brain. The three overlapping lesion regions, Hippocampus (L), Hippocampus (R) and Middle Temporal Gyrus (L), have been extensively reported in the literature to have significant gray matter degeneration in AD patients [26–28]. The Alzheimer's Disease usually begins and is ultimately most severe in the medial temporal lobe, particularly the entorhinal cortex and hippocampus was reported in [26]. Schuff et al. [27] showed that Alzheimer's Disease patients have a smaller hippocampus on average and greater volume loss over time than normal subjects, whereas mild cognitive impairment patients have a volume of hippocampus between patients of Alzheimer's Disease and normal subjects. Roquet et al. [28] reported that dementia revealed atrophy around the left Middle Temporal Gyrus in mild-AD patients.

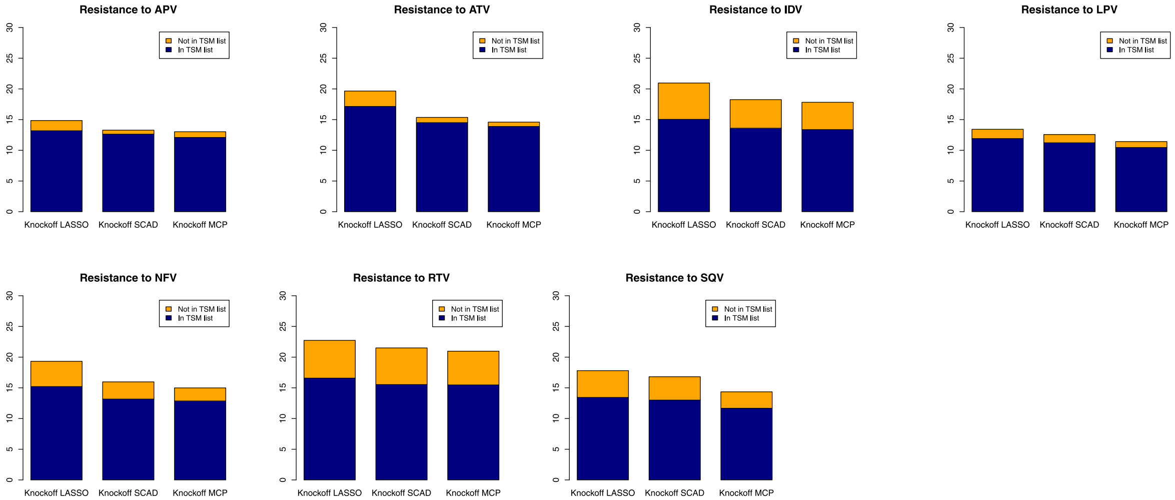


Fig. 5. Results of applying Knockoff SCAD, Knockoff MCP and Knockoff LASSO on PI-type drug resistance of HIV-1 based on genetic mutations. Dark blue indicates the selected mutations is in the TSM list (true discoveries), orange indicates the selected mutations is not in the TSM list (false discoveries).

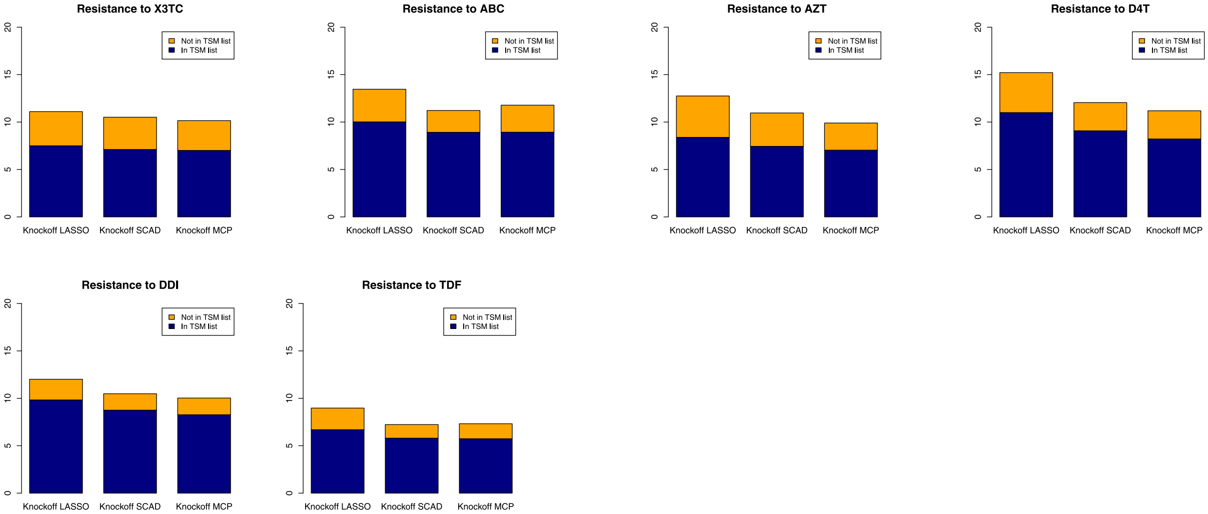


Fig. 6. Results of applying Knockoff SCAD, Knockoff MCP and Knockoff LASSO on NRTI-type drug resistance of HIV-1 based on genetic mutations.

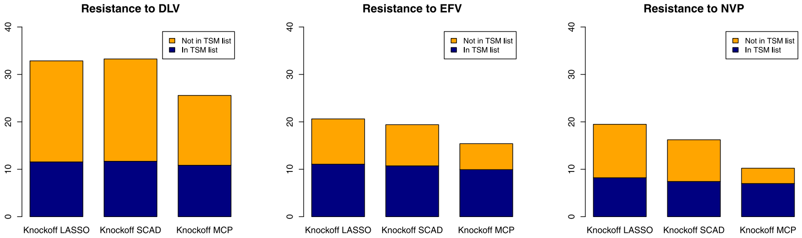


Fig. 7. Results of applying Knockoff SCAD, Knockoff MCP and Knockoff LASSO on NNRTI-type drug resistance of HIV-1 based on genetic mutations.

Table 2
The lesion regions are selected using Knockoff SCAD, Knockoff MCP and Knockoff LASSO in structural MRI dataset.

Selected regions	Knockoff SCAD	Knockoff MCP	Knockoff LASSO
Hippocampus (L) [26,27]	✓	✓	✓
Hippocampus (R) [26,27]	✓	✓	✓
Middle Temporal gyrus (L) [28]	✓	✓	✓
Middle Temporal gyrus (R) [29]	✓	✓	
Inferior Temporal gyrus (L)	✓	✓	
Inferior parietal, but supramarginal and angular gyri (R) [30]	✓	✓	
Superior frontal gyrus, orbital part (L) [30,31]		✓	

Compared to Knockoff LASSO, Knockoff SCAD and Knockoff MCP can additionally select region “Inferior Temporal gyrus” in left brain and “Middle Temporal gyrus”, “Inferior parietal, but supramarginal and angular gyri” in the right brain. Bakkour et al. reported that Inferior Temporal gyrus much more affected by Alzheimer’s Disease than by aging in [30]. Dong et al. [29] suggest that hypometabolism of right Middle Temporal Gyrus may be a typical feature of subjective cognitive decline, and that massive hypometabolism in patients in the symptomatic phase of Alzheimer’s Disease may begin with right Middle Temporal Gyrus and develop gradually from the preclinical stage. Bakkour et al. [30] also the Inferior parietal region is affected by both aging and Alzheimer’s Disease.

Further, Cajanus et al. [31] studied the effect of Superior Frontal gyrus on disinhibition and aberrant motor behavior, and the results of the study showed that the left Superior Frontal gyrus showed a correlation with elation and disinhibition, while aberrant motor behavior was associated with right Superior Frontal gyrus. Of the three methods we compared, the unique Knockoff MCP selects the “Superior frontal gyrus, orbital part (L)” associated with Alzheimer’s Disease.

These results show that the proposed methods have more powerful than Knockoff LASSO in this experiment. This is due to the nonconvex regularization possesses excellent statistical theoretical properties such as unbiasedness, sparsity and Oracle properties that guarantee that nonsignificant variables are missed.

4.3. Application to tumor sample purity estimation

Tumor sample purity is defined as the percentage of cancer cells in the tumor sample and is utilized to detect non-cancer cells and cancer cells in the tumor microenvironment. Current studies show that tumor sample purity is significantly correlated with gene expression data [32,33]. In recent years, RNA-seq gene expression data have been used to estimate tumor purity and to discover genetic signatures for individual cancer types [34,35]. Tumor purity and gene expression are positively correlated, then it is highly likely that the gene is predominantly expressed in cancer cell. In this application, we are interested in the important genes identified by applying the proposed methods to estimate tumor sample purity using RNA-seq expression data.

4.3.1. RNA-seq expression dataset and settings

RNA-seq gene expression data and tumor purity data required for the experiment can be downloaded from the Genomic Data Commons[3] [36]. We utilize the proposed methods for estimation of tumor purity to identify genes associated with skin cutaneous melanoma (SKCM) and breast invasive carcinoma (BRCA), respectively. First, we eliminate samples with missing values and 0 variance in the RNA-seq gene expression data. The three datasets we obtained contain 444 SKCM samples and 1000 BRCA samples respectively, and each sample consists of 20506 expressed genes. The tumor purity corresponding to the sample in the gene expression data are used as the response variable. Before performing gene identification procedure, we first reduce the expression data from more than 20,000 dimensions to about 500 dimensions by using LASSO to save computational costs. We set the desired FDR target level α equal to 0.1. Further, we analyzed the genes identified by the Knockoff SCAD, Knockoff MCP and Knockoff LASSO, respectively. Each method is run for 100 repetitions of the experiment on each cancer dataset, and then the gene is identified if it has been selected more than half the times.

Table 3
There are genes identified by Knockoff SCAD and Knockoff MCP respectively that are associated with SKCM tumor purity.

Knockoff SCAD	Knockoff MCP
ACTL8, APOL6, BMP15, DDX4, DMRTC1, EFEMP1, GLRX, GPR109A, GPR120, HSPA12A, ISG20, KRTAP19.7, LOC440461, LOXL1, PPP1R16B, RNF144A, SLC1A7, VGLL4	ACTL8, APOL6, BMP15, DDX4, DMRTC1, DNTT, EFEMP1, GLRX, GPR109A, GPR120, GPR27, HSPA12A, ISG20, KRTAP19.7, LOC440461, LOXL1, PPP1R16B, RNF144A, SLC1A7, VGLL4

4.3.2. Analysis of identified genes related to tumor sample purity

Tables 3 and 4 show the genes identified by Knockoff SCAD, Knockoff MCP and Knockoff LASSO, respectively. The expression of these genes are related to SKCM tumor purity. The correlation between relative gene expression level and SKCM tumor purity is depicted in Figs. 9 and A.16, and these genes are common among the genes identified by the three methods. In this experiment, all SKCM samples with tumor purity of the top $1/3$ are divided into one group (High purity) and the last $1/3$ are divided into one group (Low purity). Each high-low pair is tested by nonparametric Wilcoxon signed-rank test and the p-value is shown at the top of box plot. The box plots 9 and A.16 show that the expression levels of most of the identified genes are highly correlated with SKCM tumor purity. Most of these genes are derived from stromal cells, as their expression levels are negatively correlated with SKCM tumor purity, such as ISG20, GLRX, APOL6. Then, all SKCM samples with gene expression levels of the top $1/3$ are divided into one group (High level) and the last $1/3$ are divided into one group (Low level). We perform survival analysis on these identified genes (ISG20, GLRX, APOL6, RNF144 A). The genes in Fig. 9 correspond to genes with significant p-value in the coefficients of the Cox regression model. The Kaplan–Meier curve of the survival analysis is shown in Fig. 8. Cheng et al. [37] reports the high expression of ISG20 is related to the long overall survival in SKCM and suggests that ISG20 may be a good marker of virus prevention and cancer progression. The high expression level of APOL6 may detect malignant transformation at its earliest occurrence, which is a necessary condition for improving preventive interventions and reducing the incidence of cancer [38]. RNF144 A is specifically upregulated in melanocytes and has the function of avoiding uncontrolled proliferation. It is found to be a part of embryonic development and is a regulator of cancer development [39].

Tables 5 and 6 reports identified genes related to BRCA tumor purity by using Knockoff SCAD, Knockoff MCP and Knockoff LASSO. The box plots in Figs. 11 and A.17 show that the expression levels of most of the identified genes are highly correlated with BRCA tumor purity. The settings of this experiment are the same as that of SKCM. The survival analysis is shown in Fig. 10. Among such genes, TARS, TPRXL, SPOCK3, PAK7, IL12B and CCL21 have significant p-value in the coefficients of the Cox regression model. TARS is correlated to shorter OS time in BRCA [40]. Hicks et al. [41] report that there is a significant positive correlation between IL12B gene expression levels and tumor infiltration of $CD8^{+}\ T$ cells and M1 macrophages. CCL21 is ranked as one of the top important genes for pan-cancer tumor purity prediction in [35]. In summary, the partial genes identified by our method reproduce previous work utilizing diverse methods.

5. Discussion and conclusion

Aiming at the low statistical power and poor reproducibility of results of existing variable selection methods for identifying biomarkers, we propose two knockoff-based nonconvex regularization methods and apply them to identify biomarkers. These two methods are able to perform variable selection while rigorously guaranteeing the FDR at a given desired level with high statistical power. The proposed methods have good statistical theoretical properties such as unbiasedness, sparsity and Oracle, which can better handle heavy-tailed distributions,

[3] https://gdc.cancer.gov/about-data/publications/pancanatlas

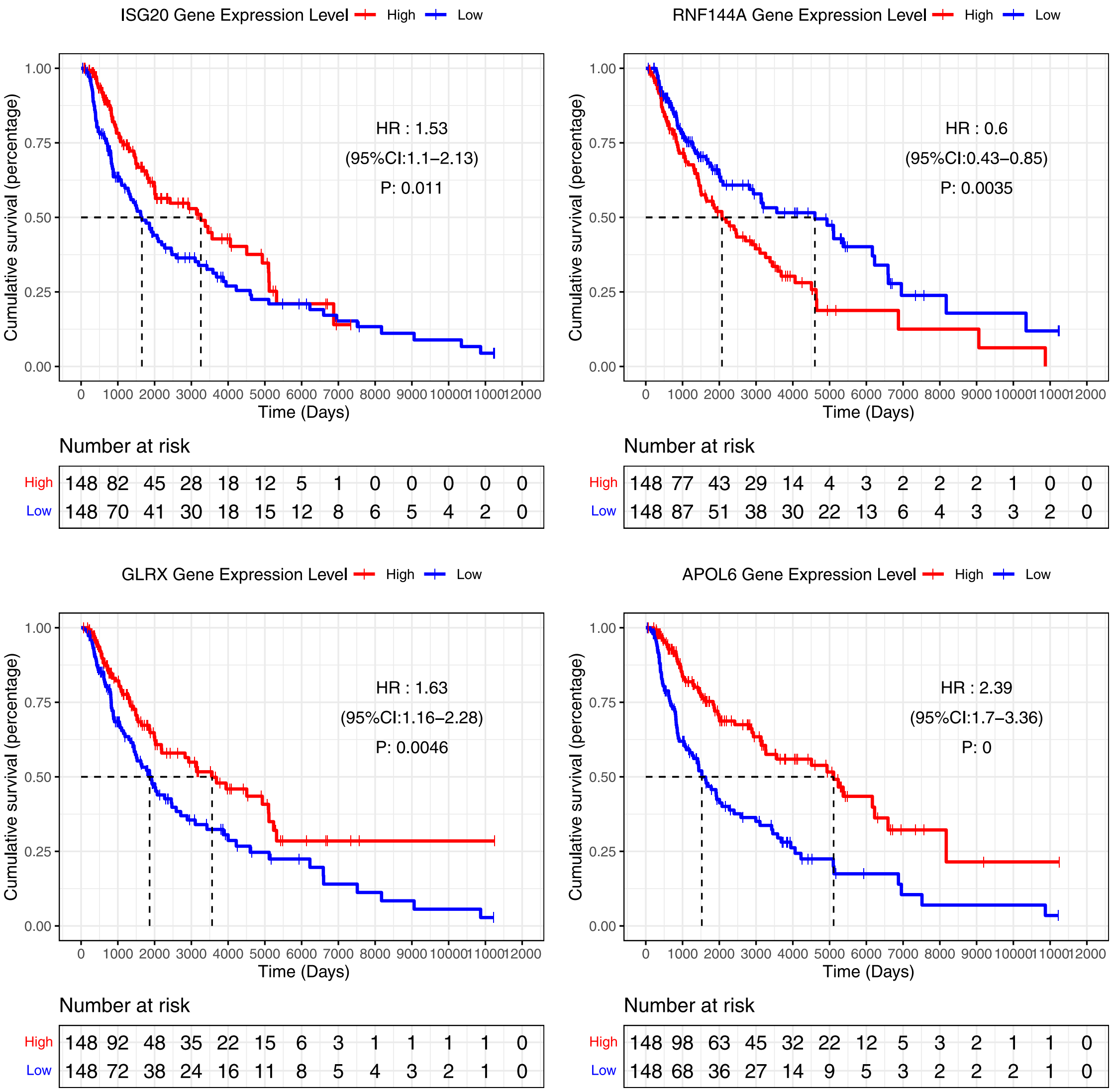


Fig. 8. Kaplan–Meier curve of SKCM survival analysis, hazard ratio (HR) and its 95% confidence interval in cox regression, and the statistical test p-value of the overall model importance.

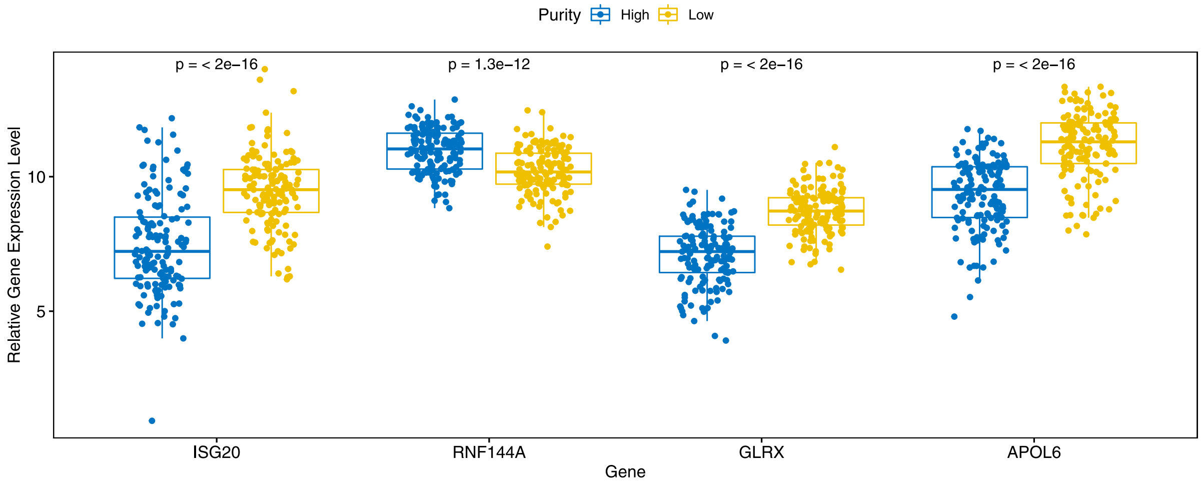


Fig. 9. The relative gene expression level of identified genes corresponding survival analysis Fig. 8 in SKCM. Yello indicates low purity and blue indicates high purity. The relative gene expression level is normalized by log_2 transformation.

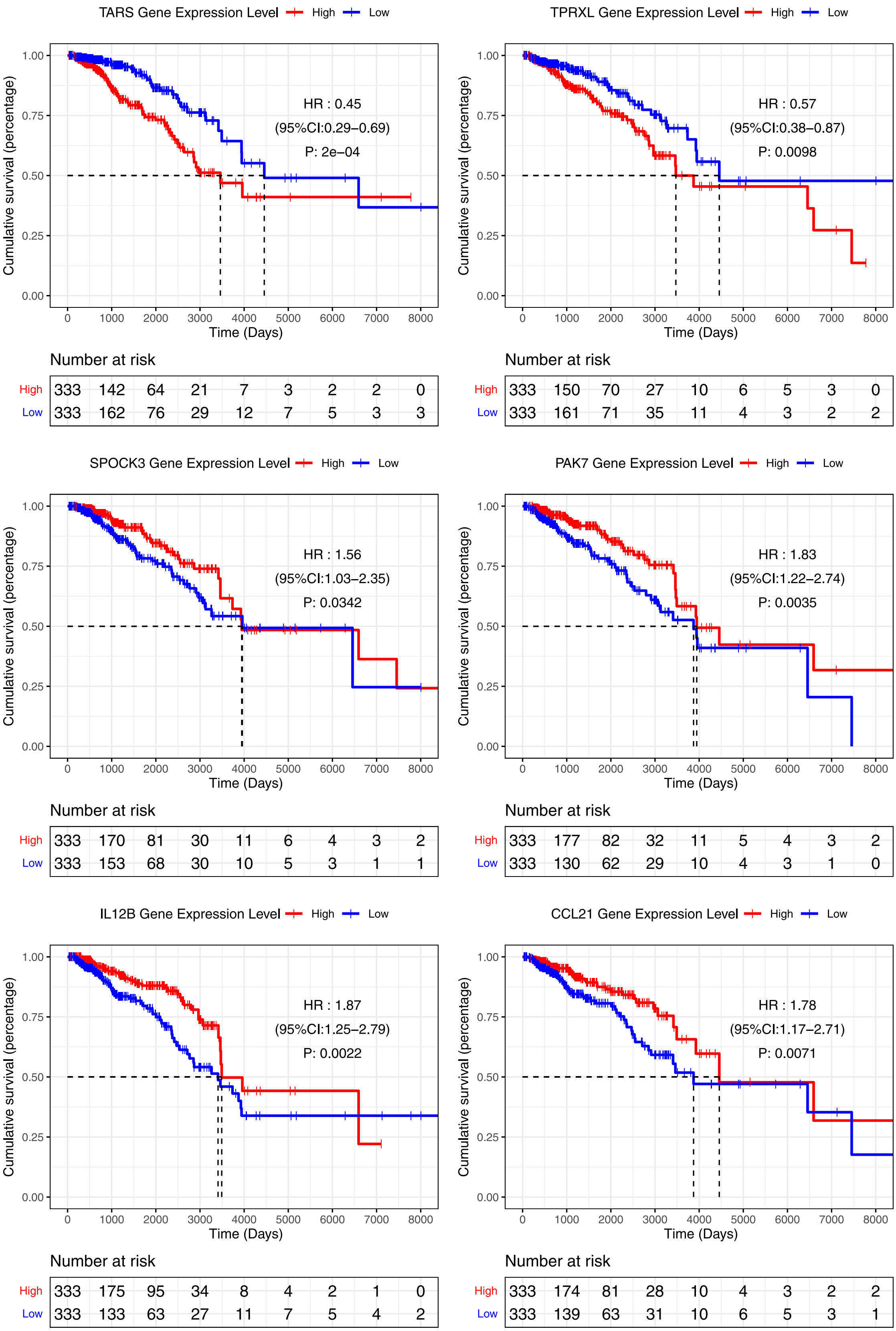


Fig. 10. Kaplan–Meier curve BRCA survival analysis, hazard ratio (HR) and its 95% confidence interval in cox regression, and the statistical test *P*-value of the overall model importance.

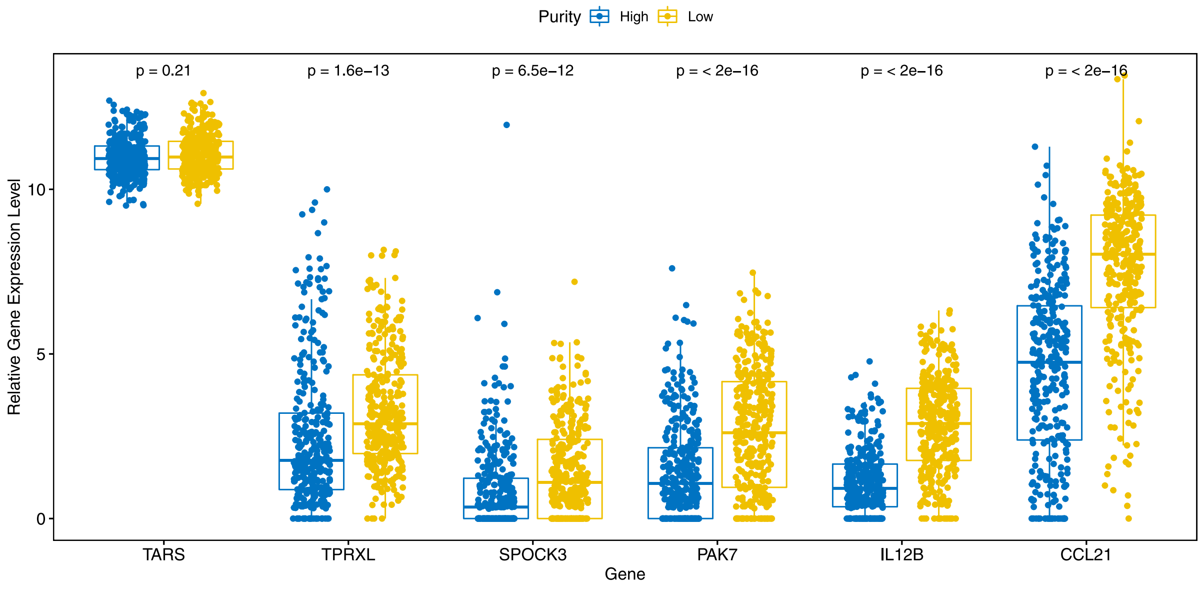


Fig. 11. The relative gene expression level of identified genes corresponding survival analysis Fig. 10 in BRCA. Yello indicates low purity and blue indicates high purity. The relative gene expression level is normalized by log_2 transformation.

Table 4
There are genes identified by Knockoff LASSO that are associated with SKCM tumor purity.

Knockoff LASSO
ACSL5, ACTL7B, ACTL8, ACVR1C, ADCY3, ADPGK, AKAP8L, ALDH4A1, ANO2, APOBEC1, APOL6, ARC, BCORL2, BCYRN1, BEST4, BMP15, C11orf85, C15orf53, C17orf57, C6orf223, C7orf16, CAMP, CBWD6, CCDC65, CD163L1, CDK5R1, CHRM4, CRTC3, CXCL6, CXorf66, DDX4, DEFB109P1, DEFB126, DMRTC1, DNTT, EFEMP1, EMR3, EPDR1, FAM151A, FFAR2, FGFBP2, FKBP9L, FLJ32063, FRAS1, GALNT13, GLRX, GPHN, GPR109A, GPR120, GPR27, GRAPL, HCG4P6, HCP5, HLA.DRB6, HSPA12A, IL6, IQCF3, ISG20, KRT76, KRTAP19.7, LAMP3, LDHD, LIMS3, LOC100268168, LOC340094, LOC375190, LOC440461, LOC441204, LOC729121, LONRF2, LOXL1, LUZP4, LYZL6, MBD3L1, MBD6, MIPOL1, MPP4, MS4A4A, MX1, MYH2, MYO1H, NBPF16, NFKB2, NRAP, OR13F1, OR1D2, OR8G5, PADI6, PARP8, PCDHGA9, PDE6C, PDX1, PI16, PNKD, PPP1R16B, PRODH, PSG2, RARRES1, RELB, RNF113B, RNF126P1, RNF144A, RORC, RPA1, RPS26P11, SLC1A7, SNORA62, SNRPN, SNTG2, SPATA9, SRBD1, TBC1D28, TNFRSF18, TNIP2, TSPAN33, TXNDC11, UFSP2, UNC13A, VGLL4, ZFAND3, ZNF366, ZNF619

Table 5
There are genes identified by Knockoff LASSO that are associated with BRCA tumor purity.

Knockoff LASSO
ANP32A, C14orf23, C1orf127, C20orf152, CARD10, CCL21, CD300E, CDH9, CDRT15, CST1, CTAGE4, CYP2A6, DEFB119, DMGDH, FAM118A, FAM193A, FGF1, FNIP2, FOXP2, GRIA2, IL12B, IL18RAP, ISM1, KRTAP19.3, LCN12, MAGEB16, MIAT, NAALADL1, OR10A7, OR5AR1, OVCH2, PAK7, PCK1 , PECAM1, PIM2, PPIAL4B, PPIAL4E, SAA4, SLC37A1, SMTN, SPANXB2, SPOCK3, TARS, TEPP, TPRXL, TSNAX.DISC1

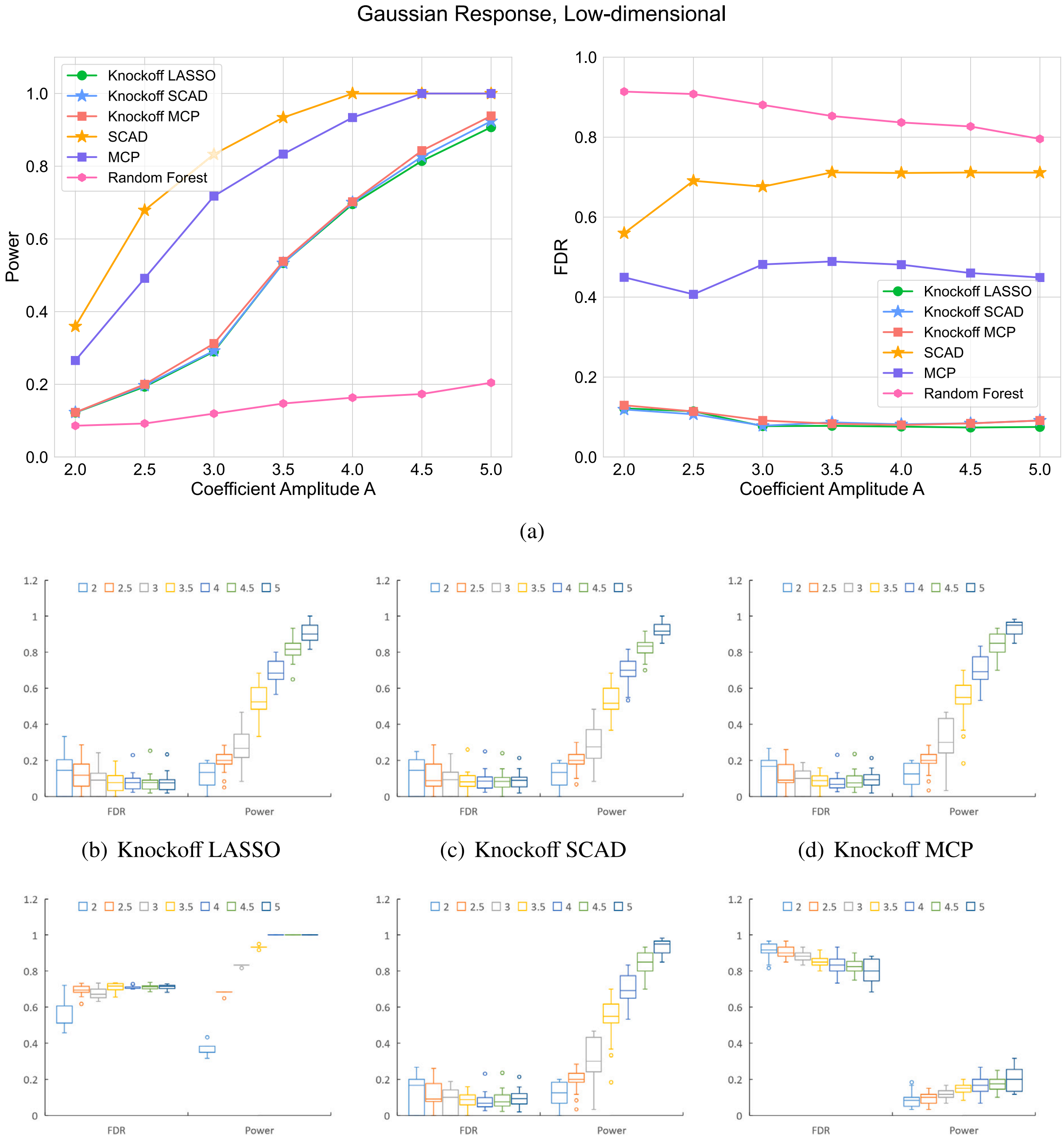
Table 6
There are genes identified by Knockoff SCAD and Knockoff MCP respectively that are associated with BRCA tumor purity.

Knockoff SCAD	Knockoff MCP
C14orf23, C20orf152, CCL21, CD300E, CDH9, CDRT15, IL12B, LCN12, MAGEB16, MIAT, NAALADL1, OR10A7, OR5AR1, PAK7, PPIAL4B, PPIAL4E, SPANXB2, SPOCK3, TARS, TPRXL	C14orf23, C20orf152, CCL21, CD300E, CDH9, CDRT15, CST1, CYP2A6, GRIA2, IL12B, LCN12, MAGEB16, MIAT, NAALADL1, OR10A7, OR5AR1, PAK7, PECAM1, PIM2, PPIAL4B, PPIAL4E, SPANXB2, SPOCK3, TARS, TPRXL

high noise and high correlation data. We use comprehensive simulation data studies to validate the proposed methods for FDR control in different settings. We verify that the proposed methods outperform the compared baseline methods in terms of statistical power and robustness. We further apply the proposed methods to the identification of three biomarkers: identification of mutations in Human Immunodeficiency Virus (HIV) drug resistance-related genes, identification of brain lesion regions in Alzheimer's disease, and identification of purity-related genes in tumor samples. Our results show that the proposed methods can provide a powerful add-on to existing bioinformatics tools to improve the reliability of FDR control, thereby increasing the reproducibility of scientific discoveries and providing a reference and aid for clinical diagnosis and treatment by reducing time-consuming and expensive experimental validation.

In experiments simulating synthetic data, the proposed methods have a significant advantage over methods that do not take FDR control into account, such as SCAD regularization, MCP regularization, and random forests, in controlling the FDR at a given desired level and also obtaining good statistical power. Compared to the Knockoff LASSO method, although it is a slight advantage in part of the experiments, it has significantly higher statistical efficacy in part of the experiments at the same level of FDR control. However, we also find that the performance of the simulation experiment results is also unstable due to the unstable generation of the knockoff variables, which requires the development of more stable and fast methods for generating the knockoff variables, which will be a possible future research work. In addition, the proposed methods are a linear model, which may not be able to capture the nonlinear effects, while deep neural networks have excellent nonlinear modeling capabilities, therefore, extending the idea of the proposed methods to deep neural networks is also a very worthwhile research topic, and how to design statistics in the face of the huge model parameters of deep neural networks is a challenge.

To evaluate the results Human Immunodeficiency Virus (HIV) drug resistance related gene mutation identification, we compare the selected mutations with the existing treatment-selective mutation (TSM) group, the actual real situation is unknown, but these panels provide a good approximation of the reference, and the proposed methods has fewer false discoveries than Knockoff LASSO. As for the identification of brain lesion regions in Alzheimer's disease and the identification of genes related to the purity of tumor samples, due to the lack of real reference biomarkers, we queried a number of literature reports for



(a)

(b) Knockoff LASSO

(c) Knockoff SCAD

(d) Knockoff MCP

(e) SCAD

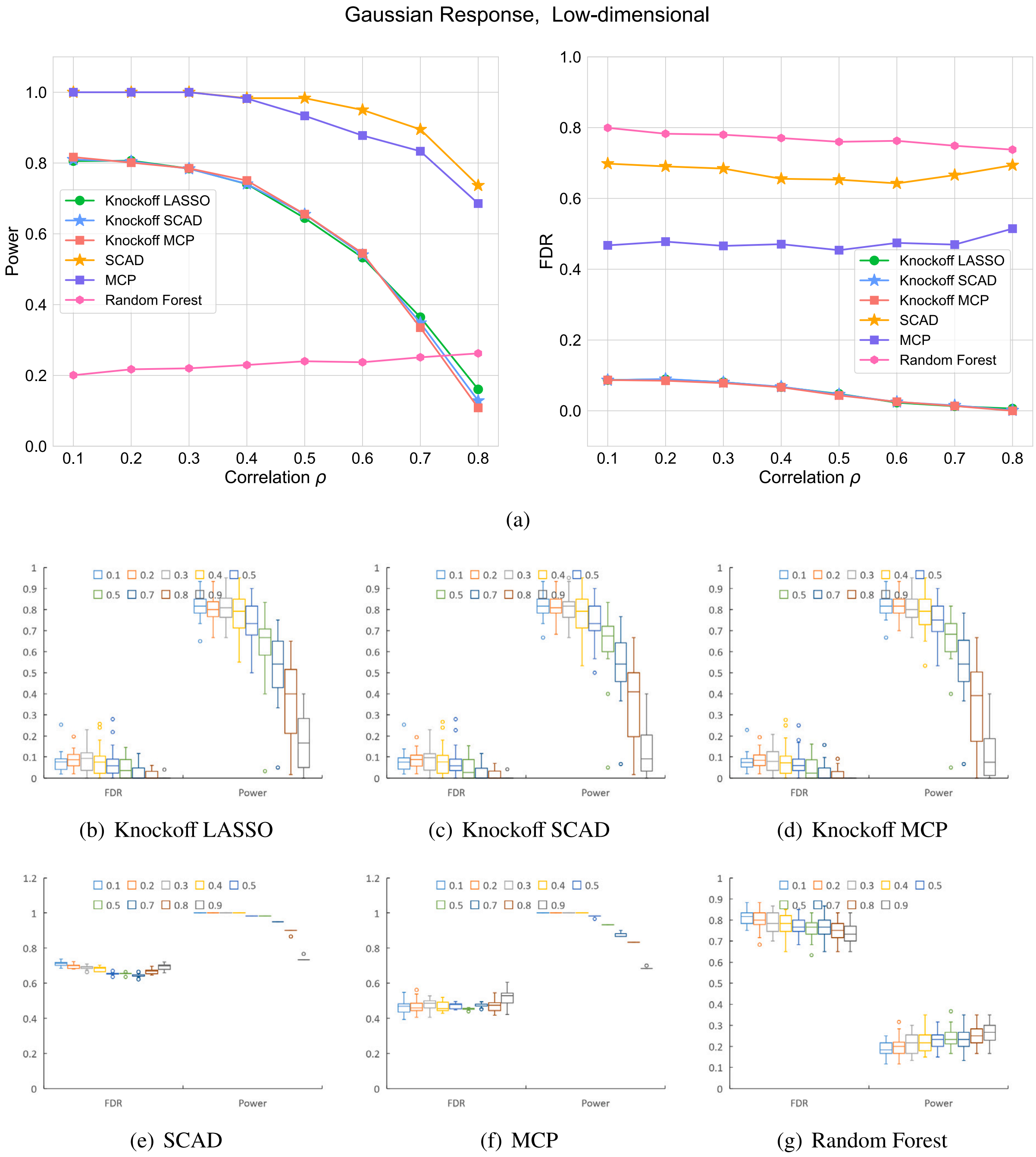
(f) MCP

(g) Random Forest

Fig. A.12. Comparison of Power and FDR for various variable selection procedures with varying signal amplitude in the low-dimensional setting of the Gaussian linear model.

side-by-side validation of the biomarkers identified by our method, and for more in-depth research on a particular topic, we need to further validate our selected biomarkers against wet experiments, which would be out of our the scope of this work is beyond our ability.

We have designed extensive and specific experiments to validate the FDR control of the proposed methods, both based on simulated synthetic data and based on real bioinformatic data. However, in most bioinformatics method papers, FDR control is only mentioned as a metric but rarely validated, thus, in the comparison method we use it can be seen that variable selection methods (e.g., random forests) that have not considered FDR control, which have a very high FDR, get many false positives. Many scholars have argued that the use of the BH procedure for p-values is effective in controlling FDR; however, because p-values will be invalidated when model assumptions are violated or p-values are computationally problematic. We use the knockoff framework to eliminate the need to compute p-values and combine it with the excellent statistical properties of non-convex regularization methods, making it a useful tool for identifying biomarkers in biomedical big data.



(a)

(b) Knockoff LASSO

(c) Knockoff SCAD

(d) Knockoff MCP

(e) SCAD

(f) MCP

(g) Random Forest

Fig. A.13. Comparison of Power and FDR for various variable selection procedures with varying variable correlation in the low-dimensional setting of the Gaussian linear model. Here the signal amplitude $A = 4.5$.

CRediT authorship contribution statement

Shoujiang Li: Investigation, Software, Writing – original draft. **Hui Zhang:** Methodology, Validation, Writing – review & editing, Visualization. **Yong Liang:** Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

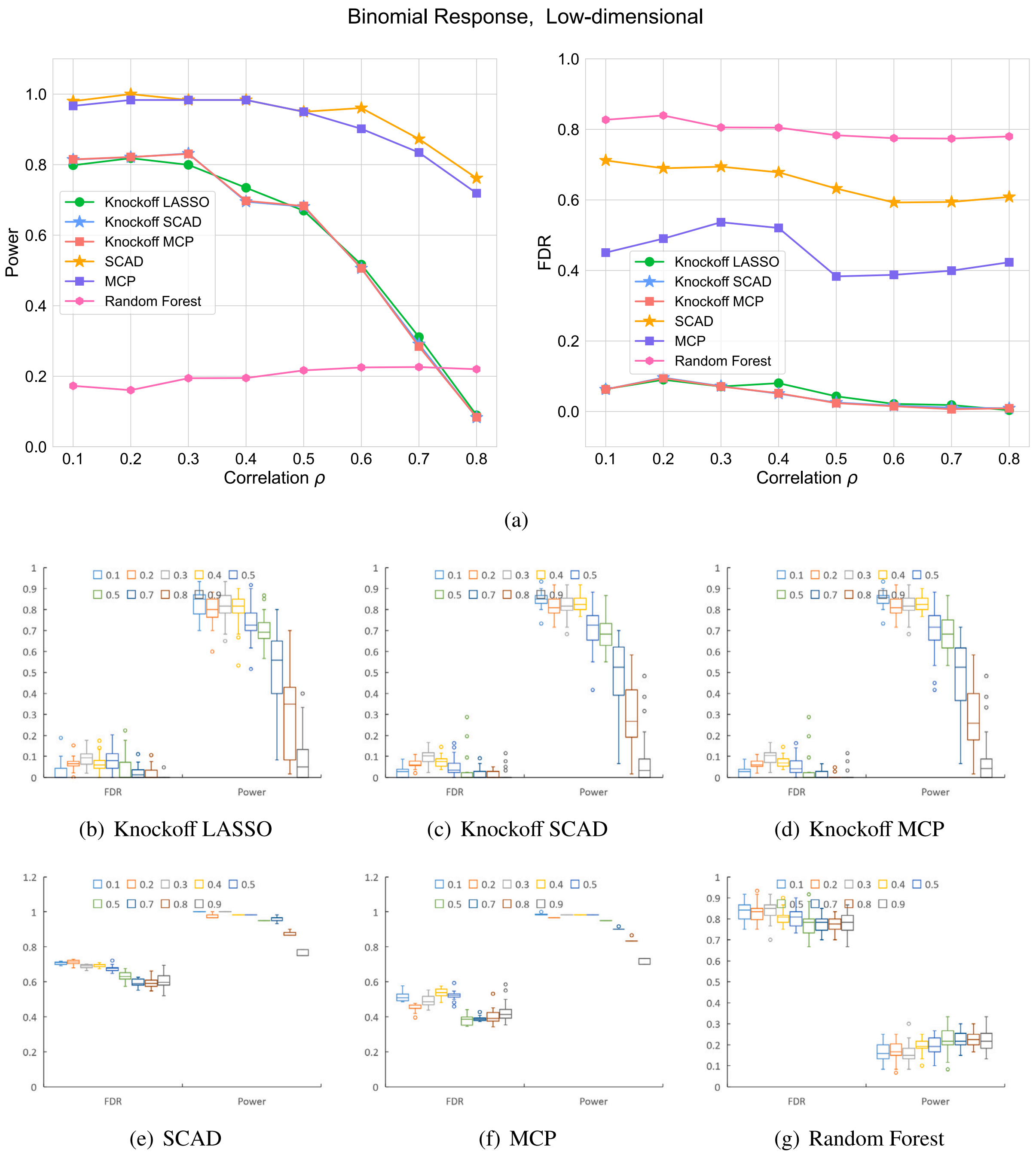
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported partially by the Natural Science Basic Research Program of Shaanxi (2024JC-YBQN-0717), the Shaanxi Provincial Joint Laboratory of Artificial Intelligence (2022JC-SYS05), the Major Key Project of Peng Cheng Laboratory (PCL2023AS1-2), and the National Key Research and Development Program of the Ministry of Science and Technology of China (2022ZD0116004).



(a)

(b) Knockoff LASSO

(c) Knockoff SCAD

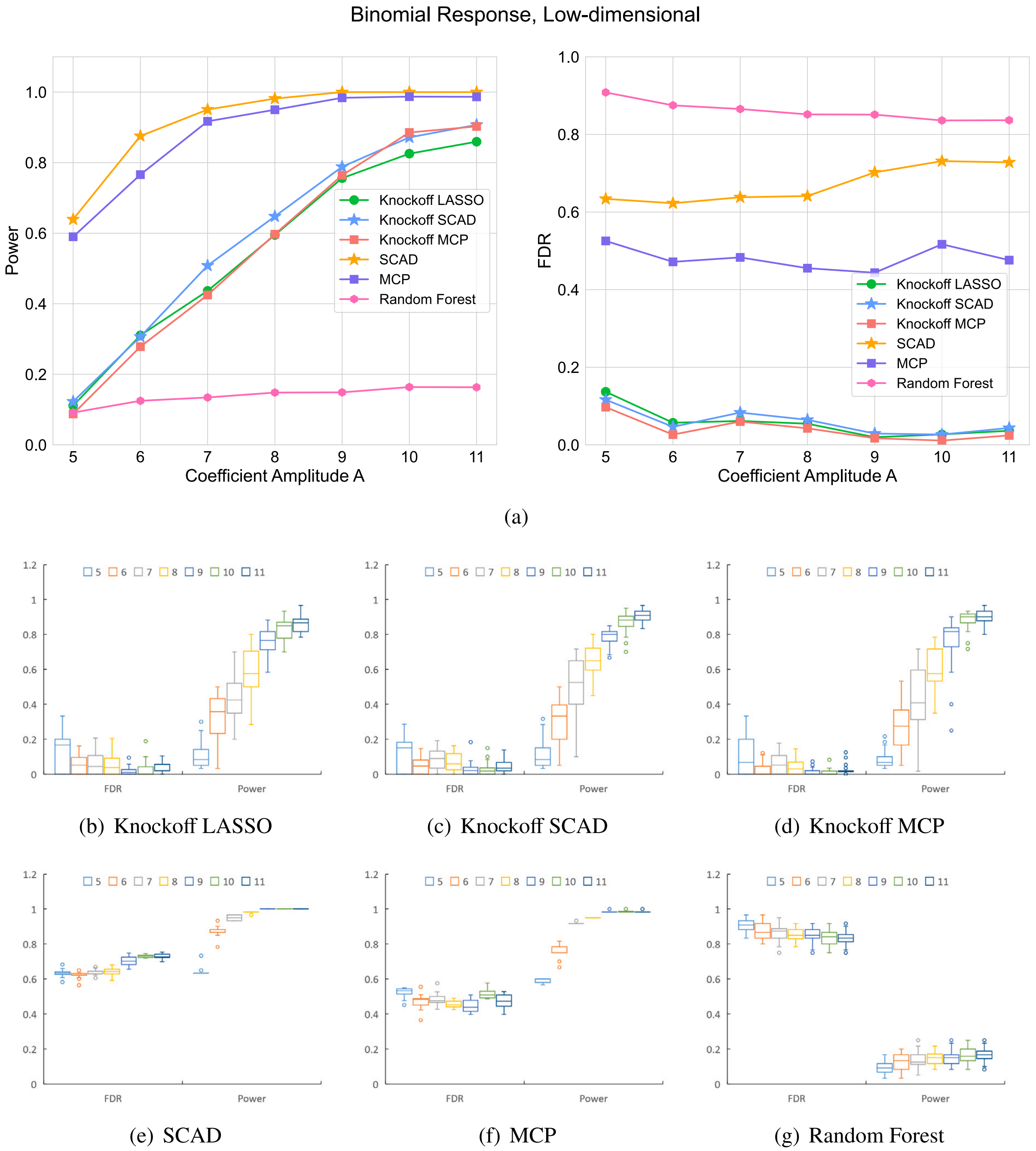
(d) Knockoff MCP

(e) SCAD

(f) MCP

(g) Random Forest

Fig. A.14. Comparison of Power and FDR for various variable selection procedures varying variable correlation in the low-dimensional setting of the binomial linear model. Here the signal amplitude $A = 10$.



(a)

(b) Knockoff LASSO

(c) Knockoff SCAD

(d) Knockoff MCP

(e) SCAD

(f) MCP

(g) Random Forest

Fig. A.15. Comparison of Power and FDR for various variable selection procedures with varying signal amplitude in the low-dimensional setting of the binomial linear model.

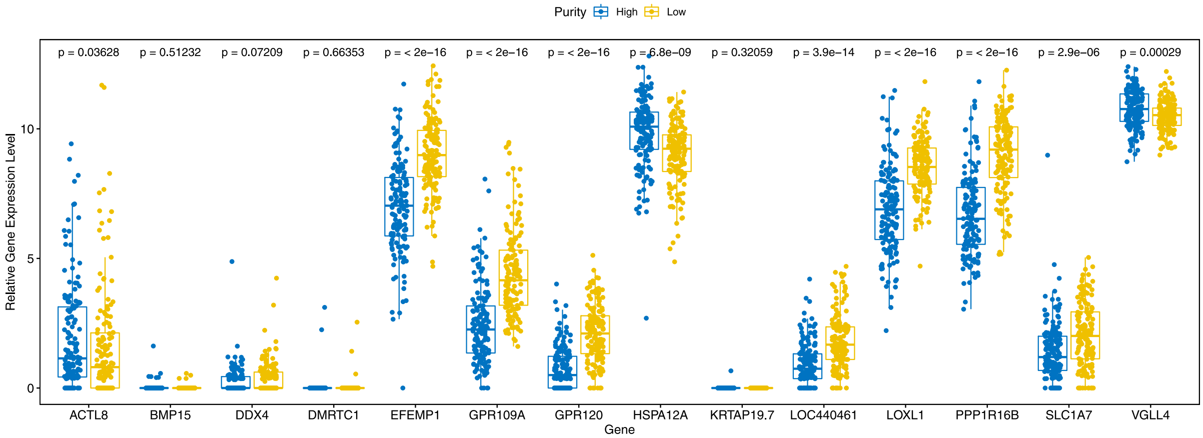


Fig. A.16. The relative gene expression level of identified genes in SKCM. Yello indicates low purity and blue indicates high purity. The relative gene expression level is normalized by log_2 transformation.

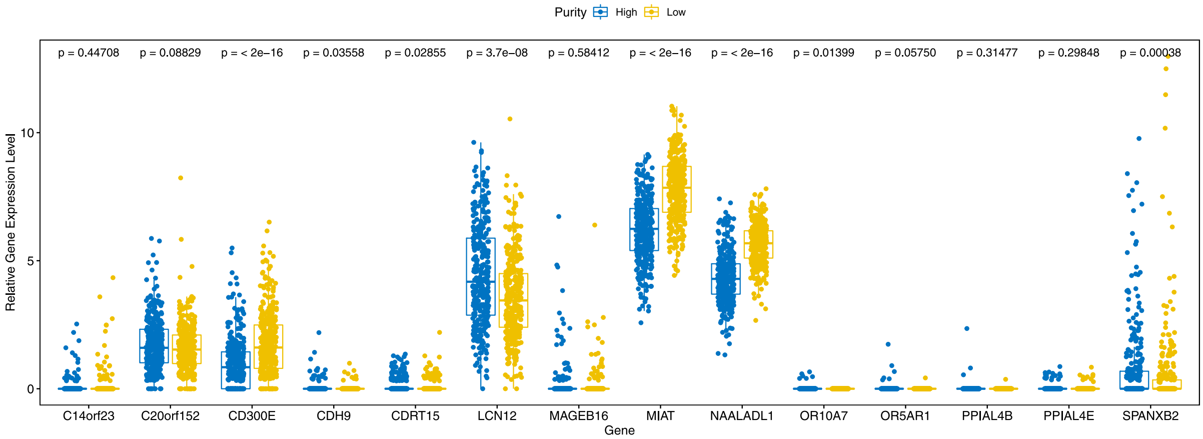


Fig. A.17. The relative gene expression level of identified genes in BRCA. Yello indicates low purity and blue indicates high purity. The relative gene expression level is normalized by log_2 transformation.

Appendix. Figures

See Figs. A.12–A.17.

References

[1] Biomarkers Definitions Working Group, A.J. Atkinson Jr., W.A. Colburn, V.G. DeGruttola, D.L. DeMets, G.J. Downing, D.F. Hoth, J.A. Oates, C.C. Peck, R.T. Schooley, et al., Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework, Clin. Pharmacol. Therapeutics 69 (3) (2001) 89–95.

[2] N. Catalyst, Healthcare big data and the promise of value-based care, NEJM Catal. 4 (1) (2018).

[3] N.P. Tatonetti, Translational medicine in the age of big data, Brief. Bioinform. 20 (2) (2019) 457–462.

[4] B. Berger, J. Peng, M. Singh, Computational solutions for omics data, Nat. Rev. Genet. 14 (5) (2013) 333–346.

[5] L. Bravo-Merodio, A. Acharjee, D. Russ, V. Bisht, J.A. Williams, L.G. Tsaprouni, G.V. Gkoutos, Translational biomarkers in the era of precision medicine, Adv. Clin. Chem. 102 (2021) 191–232.

[6] B.K. Dunn, P.D. Wagner, D. Anderson, P. Greenwald, Molecular markers for early detection, in: Seminars in Oncology, vol. 37, (no. 3) Elsevier, 2010, pp. 224–242.

[7] J.M. Rhea, R.J. Molinaro, Cancer biomarkers: Surviving the journey from bench to bedside, MLO: Med. Lab. Observer 43 (3) (2011) 10–12.

[8] G. Poste, Bring on the biomarkers, Nature 469 (7329) (2011) 156–157.

[9] N. Rifai, M.A. Gillette, S.A. Carr, Protein biomarker discovery and validation: The long and uncertain path to clinical utility, Nature Biotechnol. 24 (8) (2006) 971–983.

[10] Z. He, L. Liu, C. Wang, Y. Le Guen, J. Lee, S. Gogarten, F. Lu, S. Montgomery, H. Tang, E.K. Silverman, et al., Identification of putative causal loci in whole-genome sequencing data via knockoff statistics, Nat. Commun. 12 (1) (2021) 3152.

[11] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc.: Ser. B (Methodological) 57 (1) (1995) 289–300.

[12] E. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection, J. R. Stat. Soc. Ser. B Stat. Methodol. 80 (3) (2018) 551–577.

[13] Y. Benjamini, D. Yekutieli, et al., The control of the false discovery rate in multiple testing under dependency, Ann. Statist. 29 (4) (2001) 1165–1188.

[14] R.F. Barber, E.J. Candès, et al., Controlling the false discovery rate via knockoffs, Ann. Statist. 43 (5) (2015) 2055–2085.

[15] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.

[16] J. Fan, R. Li, Statistical challenges with high dimensionality: Feature selection in knowledge discovery, in: 25th International Congress of Mathematicians, ICM 2006, 2006.

[17] C. Zhang, et al., Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2) (2010) 894–942.

[18] J. Perl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kauffmann Publishers Inc, 1988.

[19] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R news 2 (3) (2002) 18–22.

[20] S.-Y. Rhee, R. Kantor, D.A. Katzenstein, R. Camacho, L. Morris, S. Sirivichayakul, L. Jorgensen, L.F. Brigido, J.M. Schapiro, R.W. Shafer, et al., HIV-1 pol mutation frequency by subtype and treatment experience: extension of the hivseq program to seven non-B subtypes, AIDS (London, England) 20 (5) (2006) 643.

[21] S.-Y. Rhee, W.J. Fessel, A.R. Zolopa, L. Hurley, T. Liu, J. Taylor, D.P. Nguyen, S. Slome, D. Klein, M. Horberg, et al., HIV-1 protease and reverse-transcriptase mutations: Correlations with antiretroviral therapy in subtype ?b isolates and implications for drug-resistance surveillance, J. Infect. Dis. 192 (3) (2005) 456–465.

[22] J. Ashburner, A fast diffeomorphic image registration algorithm, Neuroimage 38 (1) (2007) 95–113.

[23] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI mri single-subject brain, Neuroimage 15 (1) (2002) 273–289.

[24] W.G. Rosen, R.C. Mohs, K.L. Davis, A new rating scale for Alzheimer's disease, Am. J. Psychiatry 141 (11) (1984) 1356.

[25] R.F. Zec, E.S. Landreth, S.K. Vicari, E. Feldman, J. Belman, A. Andrise, R. Robbs, V. Kumar, R. Becker, Alzheimer disease assessment scale: Useful for both early detection and staging of dementia of the Alzheimer type, Alzheimer Dis. Assoc. Disorders (1992).

[26] P. Vemuri, C.R. Jack, Role of structural MRI in Alzheimer's disease, Alzheimer's Res. Ther. 2 (4) (2010) 1–10.

[27] N. Schuff, N. Woerner, L. Boreta, T. Kornfield, L. Shaw, J. Trojanowski, P. Thompson, C. Jack Jr., M. Weiner, Alzheimer's; Disease Neuroimaging Initiative, MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers, Brain 132 (4) (2009) 1067–1077.

[28] D. Roquet, V. Noblet, P. Anthony, N. Philippi, C. Demuynck, B. Cretin, C. Martin-Hunyadi, P.L. de Sousa, F. Blanc, Insular atrophy at the prodromal stage of dementia with Lewy bodies: A VBM DARTEL study, Sci. Rep. 7 (1) (2017) 1–10.

[29] Q. Dong, T. Li, X. Jiang, X. Wang, Y. Han, J. Jiang, Glucose metabolism in the right middle temporal gyrus could be a potential biomarker for subjective cognitive decline: A study of a han population, Alzheimer's Res. Therapy 13 (1) (2021) 1–12.

[30] A. Bakkour, J.C. Morris, D.A. Wolk, B.C. Dickerson, The effects of aging and Alzheimer's disease on cerebral cortical anatomy: Specificity and differential relationships with cognition, Neuroimage 76 (2013) 332–344.

[31] A. Cajanus, E. Solje, J. Koikkalainen, J. Lötjönen, N.-M. Suhonen, I. Hallikainen, R. Vanninen, P. Hartikainen, M. De Marco, A. Venneri, et al., The association between distinct frontal brain volumes and behavioral symptoms in mild cognitive impairment, Alzheimer's disease, and frontotemporal dementia, Front. Neurol. 10 (2019) 1059.

[32] E.A. Houseman, W.P. Accomando, D.C. Koestler, B.C. Christensen, C.J. Marsit, H.H. Nelson, J.K. Wiencke, K.T. Kelsey, DNA methylation arrays as surrogate measures of cell mixture distribution, BMC Bioinformatics 13 (1) (2012) 1–16.

[33] K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P.W. Laird, D.A. Levine, et al., Inferring tumour purity and stromal and immune cell admixture from expression data, Nature Commun. 4 (1) (2013) 1–11.

[34] D. Aran, M. Sirota, A.J. Butte, Systematic pan-cancer analysis of tumour purity, Nature Commun. 6 (1) (2015) 1–12.

[35] Y. Li, D.M. Umbach, A. Bingham, Q.-J. Li, Y. Zhuang, L. Li, Putative biomarkers for predicting tumor sample purity based on gene expression data, BMC Genomics 20 (1) (2019) 1–12.

[36] K.A. Hoadley, C. Yau, T. Hinoue, D.M. Wolf, A.J. Lazar, E. Drill, R. Shen, A.M. Taylor, A.D. Cherniack, V. Thorsson, et al., Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer, Cell 173 (2) (2018) 291–304.

[37] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Rev. 43 (1) (2001) 129–159.

[38] D. Ortega-Bernal, E. Arechaga-Ocampo, M.A. Alvarez-Avitia, N.S. Moreno, C. Rangel-Escareño, et al., A meta-analysis of transcriptome datasets characterizes malignant transformation from melanocytes and nevi to melanoma, Oncol. Lett. 16 (2) (2018) 1899–1911.

[39] Y. Xu, M. Wu, Q. Zhang, S. Ma, Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach, Genomics 111 (5) (2019) 1115–1123.

[40] H. Ye, N. Zhang, Identification of the upregulation of MRPL13 as a novel prognostic marker associated with overall survival time and immunotherapy response in breast cancer, Comput. Math. Methods Med. 2021 (2021).

[41] K.C. Hicks, P.L. Chariou, Y. Ozawa, C.M. Minnar, K.M. Knudson, T.J. Meyer, J. Bian, M. Cam, J. Schlom, S.R. Gameiro, Tumour-targeted interleukin-12 and entinostat combination therapy improves cancer survival by reprogramming the tumour immune cell landscape, Nature Commun. 12 (1) (2021) 1–18.