BMC Bioinformatics

RESEARCH

Open Access

Grace-AKO: a novel and stable knockoff filter for variable selection incorporating gene network structures

Peixin Tian[1], Yiqian Hu[1], Zhonghua Liu[2*] and Yan Dora Zhang[1,3*]

*Correspondence:
zl2509@cumc.columbia.edu;
doraz@hku.hk

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China
[2] Department of Biostatistics, Columbia University, New York, NY, USA
[3] Centre for PanorOmic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

Abstract

Motivation: Variable selection is a common statistical approach to identifying genes associated with clinical outcomes of scientific interest. There are thousands of genes in genomic studies, while only a limited number of individual samples are available. Therefore, it is important to develop a method to identify genes associated with outcomes of interest that can control finite-sample false discovery rate (FDR) in high-dimensional data settings.

Results: This article proposes a novel method named Grace-AKO for graph-constrained estimation (Grace), which incorporates aggregation of multiple knockoffs (AKO) with the network-constrained penalty. Grace-AKO can control FDR in finite-sample settings and improve model stability simultaneously. Simulation studies show that Grace-AKO has better performance in finite-sample FDR control than the original Grace model. We apply Grace-AKO to the prostate cancer data in The Cancer Genome Atlas program by incorporating prostate-specific antigen (PSA) pathways in the Kyoto Encyclopedia of Genes and Genomes as the prior information. Grace-AKO finally identifies 47 candidate genes associated with PSA level, and more than 75% of the detected genes can be validated.

Keywords: False discovery rate, High-dimensional data, Variable selection, Graph-constrained

Background

Identifying genes and pathways associated with complex traits is a primary focus for advancing the scientific understanding in genomic studies, for which extensive clinical experiments and genetic counseling are required [1]. A major challenge is that genomic data is usually high-dimensional but with a limited sample size. Currently, there are rich public genomic data, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/), which provides gene-regulatory pathways including biological regulatory relationships between genes or gene products. These gene-regulatory pathways form a network that can be represented as a graphical structure, where the vertices are genes or gene products, and the edges are gene-regulatory pathways.

Inspired by the nature of genomic data, graph-constrained estimation (Grace) offers a novel perspective on variable selection for genomic data by taking graphical structures into account. The predictors in the network are gene-expression data linked by gene-regulatory pathways for a specific clinical outcome, which can also be called graph-structured covariates [2]. These graphical structures, such as gene-regulatory pathways, can improve the sensitivity of detecting pathways [3]. Therefore, [4] developed a Grace model, network-constrained regularization procedure by encoding the graphical structures into a Laplacian matrix to incorporate this kind of prior biological information into regression analysis. The network-constrained regularization procedure includes the Lasso [5], and the elastic net regulation procedure [6] as special cases. Particularly, the network-constrained penalty may be transformed to a Lasso-type problem by constructing an augment dataset, which can also retain the automatic variables selection property [4]. The augment dataset can extend the sample size from n to $n + p$, allowing the model to choose p variables despite the fact that $n \ll p$. Moreover, because the loss function of the network-constrained penalty is a convex function, it can ensure the grouping effect of the regression in the case of identical predictors. In accordance with the theorem of [4], the quantitative description of the grouping effect is measured beyond a half of the elastic net model $\frac{1}{2\lambda_2}\sqrt{2(1-\rho)}$, where λ_2 is a fixed scalar and ρ is the correlation between the vertices, which means that if two vertices are highly correlated (e.g., $\rho = 1$), the difference between their coefficient paths could be almost 0. According to [1], the optimal policy for this variable selection effort would be to identify significant relevant genes and provide error control for these discoveries (both genes and variants). However, these conventional regression approaches only control false discovery rate (FDR) asymptotically with no guarantee in finite-sample settings. Specifically, there is a lack of effective approaches for variable selection which can not only integrate the graphical structure but also provide a guarantee of finite-sample FDR control.

To address this challenge, we propose Grace-AKO, a novel method for identifying genes associated with complex traits of scientific interest that integrates the core concept of aggregation of multiple knockoffs (AKO, [7]) with the network-constrained regularization procedure [4]. The salient idea of knockoff inference is to generate knockoff variables by mimicking the correlation structure of the original variables without considering the response variable (conditionally on the original variables) [8]. Model-X knockoffs, as opposed to fixed-X knockoffs, regarded the original variables as random and relied on the specific stochastic properties of the linear model, thus extending knockoff inference to high-dimensional data [8]. These knockoff variables are applied to control finite-sample FDR served as negative controls so that the original variables are selected if they are considerably more connected with the response variable than their knockoff variables. Specifically, the knockoff inference uses various types of feature statistics to determine which variables are significant and which are not. The feature statistics impose a flip-sign property, which implies that swapping the variables and their knockoffs alters the sign of the feature statistics. The methods for constructing the feature statistics are i.i.d. random for the "null hypothesis" whose coefficients are zero [8, 9]. To control FDR, [9] developed a data-dependent threshold whose derivation formula may be regarded as an estimate of the fraction of false discoveries. In addition, variables whose feature statistics are larger than the threshold may be selected, and estimations of the FDR can be converted

to provide finite-sample FDR control with a high degree of accuracy. However, model-X knockoffs were generated using Monte Carlo sampling, which made it challenging to reproduce the results. Thus, multiple knockoffs were proposed to address this limitation, allowing for a more stable finite-sample FDR control and more reproducible findings [7, 10, 11]. In particular, statistical aggregation is a typical statistical approach to solve instability by aggregating model-X knockoffs inference. Aggregation of multiple knockoffs (AKO) was proposed by [7], which rested on a reformulation of model-X knockoffs to introduce an intermediate feature statistic. It brought the idea from [12] to replicate model-X knockoff procedure multiple times, and then performed statistical aggregation to generate new intermediate feature statistics. Hence, it is more stable than model-X knockoff filter.

Specifically, the key contribution of our proposed Grace-AKO is that we integrate the graphical structure to conduct variable selection with finite-sample FDR control. Based on the knockoff inference property, we update the Laplacian matrix in the network-constrained penalty and perform variable selection with the original explanatory variables and their knockoffs. The primary steps of Grace-AKO are summarized as follows. First, we generate model-X knockoffs according to the correlation structure of the graph-structured covariates [8] and encode the graphical structures into a Laplacian matrix [13]. Second, we simultaneously fit the graph-structured covariates and their model-X knockoffs into the network-constrained regularization procedure to multiple feature statistics: Lasso coefficient-differences (LCDs). Third, we repeat the above procedures multiple times and employ the statistical aggregation approach [12] to transform the LCDs into new intermediate feature statistics, Aggregated Grace Coefficients (AGCs). Fourth, we conduct the Benjamini–Hochberg (BH) procedure [14] on the multiple AGCs to select the candidate variables. In our simulation studies, we show that Grace-AKO has satisfactory performance, allowing for higher reproducibility of results, and can control the FDR in finite-sample settings. We further analyze a prostate cancer data set from The Cancer Genome Atlas (TCGA) program using Grace-AKO, and then identify 47 candidate genes, of which 75% were also found in the previous literature.

The remainder of this article is organized as follows. In "Method" section, we describe the method of Grace-AKO under the linear regression framework. In "Results" section, we assess the performance of Grace-AKO using simulation studies. In "Application to the TCGA Prostate Cancer Data" section, we apply Grace-AKO to a prostate cancer data set from the TCGA program by incorporating the KEGG pathways as prior information. In "Conclusions" section, we briefly summarize our method.

Method

In genomic studies, we usually apply a regression model to identify genes and pathways associated with the trait of interest by linking high-dimensional data (e.g., microarray gene-expression data) to the trait. Consider the following linear model where $\boldsymbol{X}$ is a $n \times p$ design matrix with n observations and p predictors, and $\mathbf{y}$ is the response:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$ and $\mathbf{x}_j = (\mathrm{x}_{1j}, \mathrm{x}_{2j}, \cdots, \mathrm{x}_{nj}), j = 1, \cdots, p$ represents a vector of the graph-structured covariates from genomic data (e.g.,

gene-expression data), and the coefficient $\boldsymbol{\beta}$ represents the contribution of the graph-structured covariates to the trait of interest, and $\boldsymbol{\epsilon}$ is a vector of random errors. We further assume that the response is centred and the predictors are standardized,

$$\sum_{i=1}^{n} \mathrm{y}_i = 0, \quad \sum_{i=1}^{n} \mathrm{x}_{ij} = 0 \quad \text{and} \sum_{i=1}^{n} \mathrm{x}_{ij}^2 = 1, \quad \text{for} j \in [p],$$

where $[p]$ denotes the set including $\{1, 2, \cdots, p\}$. The main goal is to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ and select the predictors with nonzero contribution. The regulatory relationships of biological networks for some complex traits, represented as graphical structures between genes or gene products in genomic studies, shed light on underlying biological knowledge, where the covariates are the graph's nodes and the edges indicate functional relationships between two genes. The biological networks can be utilized to identify the differentially expressed genes [15, 16]. Specifically, in such a graphical structure, the genes are linked by edges with certain probabilities, where the edge probability is interpreted as a weight in an undirected graph to form a weighted graph [2]. To incorporate the prior information about the biological networks, [4] proposed a network-constrained regularization criterion:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{X\beta})^T (\mathbf{y} - \boldsymbol{X\beta}) + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 \boldsymbol{\beta}^T \boldsymbol{L\beta}, \tag{2}$$

where $\boldsymbol{L}$ is a non-negative Laplacian matrix of a weighted graph containing biological networks information, and $||\cdot||_1$ indicates the L_1 norm, and λ_1 is the Lasso penalty [5], and λ_2 is the penalty for the Laplacian matrix $\boldsymbol{L}$. Specifically, the Laplacian matrix was first introduced by [13], which included numerous properties of the graph by its consistent set of the eigenvalues or spectrum. When p is large, the model (1) is treated as "sparse", in which most elements of the coefficient $\boldsymbol{\beta}$ are zero [2]. In equation (2), the L_1 norm deals with sparse matrices, and $\boldsymbol{\beta}^T \boldsymbol{L\beta}$ induces a smooth solution of coefficients of the graph-structured covariates. Additionally, $\boldsymbol{L}$ depicts the graphical structure assuming that set V includes vertices corresponding to the graph-structured covariates, and W is the weights of the edges in which $w(u, v)$ denotes a weight of the edge between the graph-structured covariates u and v, and the degree of vertex v is represented as $d_v = \sum_{u \sim v} w(u, v)$. In the genomic data, $w(u, v)$ quantifies the uncertainty of the edge between two vertices, such as the probability of an edge connecting two graph-structured covariates when the graphical structure is constructed from data. Motivated by [13], we apply the normalized $\boldsymbol{L}$ [4]:

$$L(u, v) = \begin{cases} 1 - w(u, v)/d_u, & \text{if } u = v \text{ and } d_u \neq 0; \\ -w(u, v)/\sqrt{d_u d_v}, & \text{if } u \text{ and } v \text{ are adjacent;} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Therefore, we can rewrite the second penalty term $\boldsymbol{\beta}^T \boldsymbol{L\beta}$ of Eq. (2) as follows [4]:

$$\boldsymbol{\beta}^T \boldsymbol{L\beta} = \sum_{u \sim v} \left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}} \right)^2 w(u, v).$$

The network-constrained regularization procedure integrates the known biological network's information for variable selection. By introducing the network-constrained

penalty, the network-constrained regularization procedure was able to identify more interpretable genes and sub-networks related with the outcome of interest, while simultaneously inducing sparsity and smoothness of the biological network and the coefficients. However it does not ensure false discovery rate (FDR) control in finite-sample settings [4]. The network-constrained regularization procedure only has the asymptotic property only when $n \to \infty$ and p is fixed and suffers from identifying numerous false positive discoveries when p is large and the number of samples n is limited. To address this issue, we introduce knockoff inference and multiple knockoffs to control finite-sample FDR to achieve a stable performance. The knockoff filter procedure was first proposed in [17], and [8] further proposed model-X knockoff filter to extend its application to high-dimensional data. Model-X knockoffs, $\tilde{\boldsymbol{X}}$ are generated from the original data by Monte Carlo sampling and retain the same data structure as the originals $\boldsymbol{X}$, in which $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are pairwise exchangeable. We summarize the properties of model-X knockoffs as in [8]:

(1) Swapping the locations of related elements, $\mathbf{x}_j$ and $\tilde{\mathbf{x}}_j$, would not change the joint distribution of $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ conditional on $\mathbf{y}$.
(2) Once the original covariates $\boldsymbol{X}$ are known, their model-X knockoffs, $\tilde{\boldsymbol{X}}$ provides no extra information on the response variable $\mathbf{y}$.

The knockoffs filter is a cheap method to control finite-sample FDR since it does not require strong assumptions about the design matrix $\boldsymbol{X}$. Due of the random nature of model-X knockoffs sampling, however, the outcome would be unstable and cannot be guaranteed to be reproduced. To increase the stability, multiple knockoffs approaches were developed, and aggregation of multiple knockoffs (AKO) is one of them [7]. In this article, we present a novel method for variable selection termed Grace-AKO, which combines the biological network information for improved variable selection with finite-sample FDR control using knockoff filter technique. Our proposed Grace-AKO has four major steps.

First, we generate model-X knockoffs, $\tilde{\boldsymbol{X}}$ from the original data matrix $\boldsymbol{X}$ using R package "knockoff" [8]. $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ is further regarded as a new design matrix of the predictors. Based on the properties of model-X knockoffs whose correlation structure is the same as that of the original variables, we generate a new normalized Laplacian matrix $\boldsymbol{L}$. In summary, we generate knockoff variables and update the Laplacian matrix to include the graphical structure of knockoff variables in this step. The knockoff variables are introduced into the model based on their properties.

Second, the response variable $\boldsymbol{y}$ and the new predictors $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ are fitted in Eq. (2). Inspired by [4], a natural solution of Grace-AKO is equivalent to the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}\{\|\mathbf{y} - (\boldsymbol{X}, \tilde{\boldsymbol{X}})(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}})^T\|_2\}, \tag{4}$$

subject to:

$$(1-\alpha)\sum_{j=1}^{p}\|\beta_j\|_1+\alpha\sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}}-\frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v)\le t, \tag{5}$$

where $\alpha=\lambda_2/(\lambda_1+\lambda_2)$, and t is a constant value, and $(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})^T$ is a vector of the new coefficients. Furthermore, in the function $(1-\alpha)\sum_{j=1}^{p}\|\beta_j\|_1+\alpha\sum_{u\sim v}(\frac{\beta_u}{\sqrt{d_u}}-\frac{\beta_v}{\sqrt{d_v}})^2 w(u,v)$, $\|\cdot\|_1$ deals with sparse data matrix consistent with its function in the network-constrained penalty, and the second term penalizes the weighted sum of the squares of the difference of coefficients between the graph-structured covariates, which is scaled by the degree of the associated vertices in the network. Specifically, when two genes are connected, it is expected that their coefficients would be similar rather than identical, which is accomplished by applying the second term of the penalty [4].

To identify relevant variables, we compute the Lasso coefficient-difference (LCD) as the feature statistic to measure the evidence against null hypothesis ($\beta_j=0$) [8]:

$$W_j=|\hat{\beta}_j|-|\hat{\beta}_{j+p}|, \tag{6}$$

where $\hat{\beta}_j$, and $\hat{\beta}_{j+p}$ are the estimated coefficients of $\mathbf{x}_j$ and $\tilde{\boldsymbol{x}}_j$, respectively. Due to the symmetric distribution of W_j under the null, W_j equally takes on positive and negative values [8]. Moreover, a large positive value of W_j suggests that the distribution of $\mathbf{y}$ is statistically dependent on $\mathbf{x}_j$ and that there is a strong probability that $\mathbf{x}_j$ is a relevant gene associated with the response $\mathbf{y}$. In this step, the network-constrained penalty is integrated with knockoff variables to control finite-sample FDR. The natural solution of Grace-AKO follows the same optimization problem as the network-constrained regularization procedure. However, the model-X knockoffs procedure only generates knockoff variables once by Monte Carlo sampling, which leads to instability. To solve this challenge, we repeat the knockoff generation process multiple times and apply the statistical aggregation strategy to increase stability [12].

Third, following [7], Grace-AKO transforms the feature statistic W_j into a new intermediate feature statistic q_j:

$$q_j=\begin{cases}\frac{1+\#\{k:W_k\le -W_j\}}{p}, & W_j>0;\\ 1, & W_j<0.\end{cases} \tag{7}$$

We repeat the aforementioned steps B times, including the generation of knockoffs and the calculation of the intermediate feature statistic q_j, to generate a $B\times p$ matrix of the intermediate feature statistics. Then, we propose a new feature statistic, Aggregated Grace Coefficient (AGC), which is derived by applying the quantile aggregation algorithm [12] to the $B\times p$ matrix:

$$\bar{q}_j=\min\left\{1,\frac{Q_\gamma\left(\left\{q_j^{(b)}:b=\{1,2,\cdots,B\right\}\right)}{\gamma}\right\}, \tag{8}$$

where γ is the quantile point, and $Q(\cdot)$ denotes the quantile function, and B is the prespecified replication times. To summarize the third step, we generate model-X knockoffs B times and then use the quantile aggregation procedure to yield AGCs, $\bar{\boldsymbol{q}}$. The quantile aggregation approach introduces intermediate feature statistics and aggregated feature

statistics, which are based on the concept of statistical aggregation [12] to improve stability.

Fourth, we apply the Benjamini–Hochberg procedure (BH) [14] to the AGCs, $\bar{\boldsymbol{q}} = (\bar{q}_1, \bar{q}_2, \cdots, \bar{q}_p)$ to compute a data-dependent threshold:

$$t_{BH} = \max\left\{k : \bar{q}_{(k)} \leq \frac{k}{\alpha p}\right\},$$

where α is the user-specified nominal FDR level. We finally choose the candidate variables satisfying the following requirement:

$$\hat{S} = \{j \in [p] : \bar{q}_{(j)} \leq \bar{q}_{(\hat{t}_{BH})}\}.$$

In this article, we measure the performance using the modified false discovery rate (mFDR):

$$\text{mFDR} = \mathbb{E}\left[\frac{|\{j \in \hat{S} \cap S_0\}|}{|\hat{S}| + 1/\alpha}\right],$$

where $S_0 = \{j \in [p] : \beta_j = 0\}$ includes the predictors that have no effect on the trait of interest. The implementation details are available in Algorithm 1.

Algorithm 1: Algorithm of Grace-AKO

Data: $(\boldsymbol{X}, \boldsymbol{Y}, \gamma, q)$: $\boldsymbol{X}$-original data, $\boldsymbol{Y}$-response variable, B-loop times, γ-pre-specified quantiles, q-pre-specified FDR level.

Result: Variable selection set: $\hat{S}$.

while $b \leq B$ **do**

- Draw knockoffs variables $\tilde{\boldsymbol{X}}$;
- Fit $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ into Grace model;
- Calculate W_j based on (6); Calculate new statistics q_j based on (7);

Do aggregation by quantile function (8) with a $B \times p$ matrix of q_j and γ to get AGCs;

Use the BH procedure to compute the threshold: $t_{BH} = max\{k : \bar{q}_{(k)} \leq \frac{k}{\alpha p}\}$;

Variable selection set: $\hat{S} = \{j \in [p] : \bar{q}_{(j)} \leq \bar{q}_{(\hat{t}_{BH})}\}$.

As an illustration in [8], the primary objective of the knockoff filter procedure was to build an as permissive as possible data-dependent threshold. The threshold can provide a controllable estimation of FDR to ensure model-X knockoffs property. [7] was based on a reformulation of the original knockoff inference and developed an intermediate feature statistics to replace the feature statistics of [8]. Specifically, AKO still provided a data-dependent threshold based on the specification in [8] which is used by Grace-AKO.

Results

Simulation Studies

In this section, we performed a wide range of simulation studies to evaluate our proposed method, Grace-AKO, and compared it to the network-constrained regularization procedure (namely Grace) [4]. We supposed that there were 10 transcription factors (TFs) and each regulated 10 genes. The graphical structure included g unconnected regulatory modules with p genes in total and edges linked each of the TFs and 10 genes

regulated. Here we denoted the first 44 genes related to the response variable $\mathbf{y}$. We assumed that the data were simulated from the following settings:

(1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{\beta} = (5, \overbrace{\frac{5}{\sqrt{10}}, \cdots, \frac{5}{\sqrt{10}}}^{10}, -5, \overbrace{\frac{-5}{\sqrt{10}}, \cdots, \frac{-5}{\sqrt{10}}}^{10}, 3, \overbrace{\frac{3}{\sqrt{10}}, \cdots, \frac{3}{\sqrt{10}}}^{10}, -3, \overbrace{\frac{-3}{\sqrt{10}}, \cdots, \frac{-3}{\sqrt{10}}}^{10}, 0, \cdots 0)$$

, and $\boldsymbol{\epsilon}$ was generated from $N(0, \boldsymbol{\sigma}^2)$;

(2) The noise level was denoted as $\sigma^2 = \sum_u^p \beta_u^2/4$;

(3) For each expression level TF, X was drawn from normal distribution: $\boldsymbol{X}_g \sim N(0, 1)$, and conditional on the TF, the expression levels of genes which regulated to the specific TF were drawn from a conditional normal distribution with correlation of 0.7.

We set $n = (100, 200, 300)$, $g = (10, 20, 40, 60)$ and $p = (110, 220, 440, 660)$, where g represented the total number of unconnected regulatory modules, to simulate 12 settings in total. Figure 1 shows the sub-matrix in order to provide a more accurate depiction of the graphical structure, given that every 11 of the first 44 variables form an identical sub-network. As the empirical research in [7] demonstrated, performance was stable and robust once the iteration B approached 25 times. In the meanwhile, FDR could be empirically controlled under the pre-specified level. Therefore, we fixed $B = 25$, and $\gamma = 0.1$ in all scenarios. Additionally, we set the mFDR control level α at 0.1. According to [4], Grace performed better than the Lasso [5] and the elastic net [6] in term of the combination of $\|\cdot\|_1$ and $\|\cdot\|_2$ and the Laplacian matrix $\boldsymbol{L}$. We thus focused on the

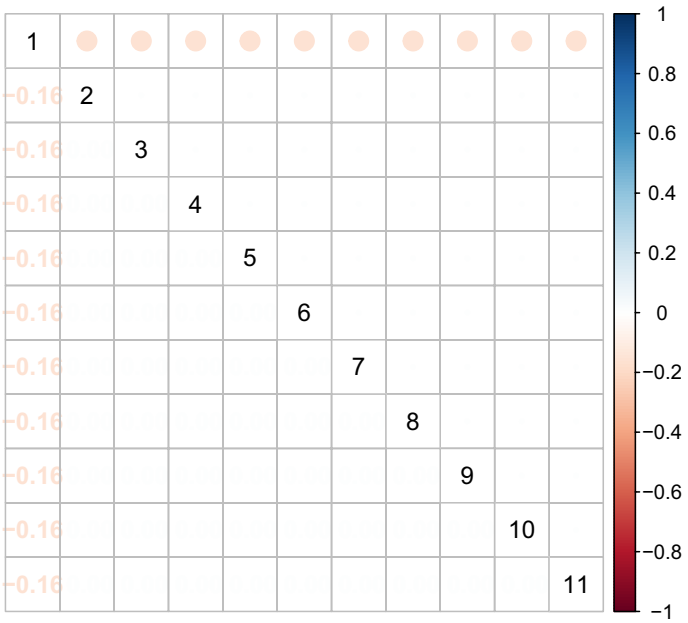

Fig. 1 Figure of 11 × 11 sub-matrix of Laplacian matrix. The upper half of the matrix is color-coded to indicate correlation. The below half of the matrix is numeric-coded to indicate correlation

comparison of Grace-AKO and Grace in our simulation studies. We assessed the regression performance using the average values of mFDR, standard errors (calculated by Monte Carlo method), and the number of variables selected.

To tune the parameters, we applied 10-fold cross-validation. The values of tuning parameter λ_L were specified from the range of 0.1 to 2.0 with a step size of 0.1, which ensured that the $\boldsymbol{L}$ matrix was non-negative. Moreover, the values of λ_1 were drawn from the range of 110 to 200 with a step size of 5. The validation set of λ_2 ranged from 1 to 10 with a step size of 1. Each simulation setting was repeated for 30 times.

As Table 1 showed, the mFDR values of Grace-AKO were controlled at the pre-specified level, and Grace could not control mFDR in finite-sample settings. Furthermore, as the data dimension p increased with a fixed sample size n, the mFDR for Grace increased which even reached at 0.67. However, Grace-AKO could always control the mFDR under 0.1. Moreover, the standard errors for Grac-AKO were always smaller than Grace's, which indicated that Grace-AKO could also improve the stability. In Fig. 2, we depicted a boxplot for the findings when $n = 300$. The upper figure's black line represented the pre-specified FDR level. Grace, in particular, had inflated power at cost of high mFDR.

As shown in Table 2, the number of selected variables varied significantly between these two models. The number of true variables was fixed at 44. When n and p increased, Grace selected many false discoveries. Notably, when $n = 300$ and $p = 660$, Grace identified 138 candidate variables. This was the reason why Grace's power was inflated in Fig. 2. Grace-AKO showed considerably more stable performance, and it can guarantee finite-sample FDR control under all data settings with slightly conservative power.

We further assessed the robustness and single knockoff (model-X knockoffs) performance on simulated data ($n = 100$, $p = 110$, and a pre-specified mFDR = 0.1). To evaluate the robustness of Grace-AKO, we randomly selected 20 vertices from the first 44 elements (true candidate genes) of the Laplacian matrix and set their degrees to zero. Additionally, we set the first and third TFs to have false degrees of 1 and 4, respectively. The mFDR and TPP of Grace-AKO were 0.013 and 0.516, respectively. It indicated that Grace-AKO could still control the mFDR under the pre-specified FDR level (mFDR = 0.1), despite the fact that some information of graphical structure was misspecified for the true variables. Moreover, prior researches demonstrated that the findings might be robust to the misspecification of the graphical structure [4, 15, 16]. As there were few genes associated with the response variable, the majority of coefficients would be zero. In addition, we examined the performance of Grace incorporating with model-X knockoffs, termed as Grace-KO over 200 simulations. We conducted simulations with

Table 1 The mean and standard errors (in brackets) of modified false discovery rate (mFDR) over 30 replications for Grace-AKO versus Grace model with a pre-specified mFDR level at 0.1

p	$n = 100$		$n = 200$		$n = 300$	
	Grace-AKO	**Grace**	**Grace-AKO**	**Grace**	**Grace-AKO**	**Grace**
10	0.01 (0.018)	0.15 (0.032)	0.03 (0.032)	0.22 (0.049)	0.03 (0.025)	0.26 (0.053)
20	0.02 (0.027)	0.26 (0.045)	0.02 (0.023)	0.37 (0.045)	0.03 (0.040)	0.43 (0.010)
40	0.03 (0.030)	0.24 (0.062)	0.04 (0.030)	0.48 (0.040)	0.06 (0.047)	0.61 (0.030)
60	0.04 (0.026)	0.40 (0.053)	0.03 (0.026)	0.55 (0.032)	0.06 (0.034)	0.67 (0.018)

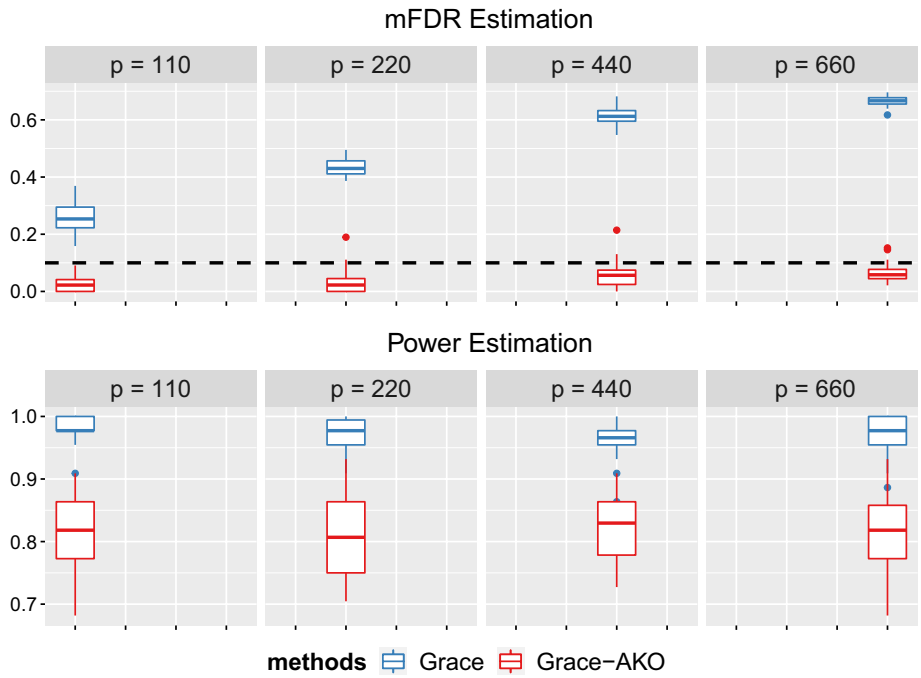

Fig. 2 Figures of mFDR and Power for Grace-AKO v.s Grace model when $n = 300$. The upper figure plots mFDR and lower figure plots power. The red color represents our new method, Grace-AKO and the blue color represents Grace. Each setting has been replicated for 30 times. The black dashed line is the pre-specified mFDR level which equals to 0.1

Table 2 The average numbers of variables selected for Grace-AKO vesus Grace over 30 replications with a pre-specified mFDR level at 0.1

p	$n = 100$		$n = 200$		$n = 300$	
	Grace-AKO	**Grace**	**Grace-AKO**	**Grace**	**Grace-AKO**	**Grace**
10	24	45	33	56	37	62
20	25	52	32	71	37	83
40	24	59	34	90	39	118
60	24	63	32	101	39	138

The number of true variables is 44

Table 3 The mean and standard errors (in brackets) of mFDR, power and the number of variables selected over 200 replications for Grace-AKO versus Grace-KO model with a pre-specified mFDR level at 0.1

	mFDR	**Power**	**Number of variables selected**
Grace-KO	0.018 (0.0277)	0.373 (0.2390)	17 (11)
Grace-AKO	0.015 (0.0204)	0.526 (0.0927)	23 (4)

200 replications when $n = 100$ and $p = 110$. The results were reported in Table 3. We observed that Grace-KO and Grace-AKO were both able to control the mFDR under the pre-specified mFDR level, and Grace-AKO shown more stable performance with a lower

standard deviation. Moreover, Grace-KO performed more conservatively than Grace-AKO, which was able to identify fewer candidate genes. Furthermore, we also assessed the computational cost of Grace-AKO when $n = 100$ and $p = 110$. The knockoff generation and inference steps of Grace-AKO could be conducted by parallel running, which took about 30 seconds with a server of Intel Xeon Silver 4116 CPU 2.10 GHz and 64 GB RAM memory. Consequently, we concluded that Grace-AKO was robust to the incorrect information of the graphical structure and was more powerful in identifying candidate variables.

Application to the TCGA Prostate Cancer Data

To demonstrate the usefulness of Grace-AKO, we applied it to a gene-expression data of prostate-specific antigen (PSA) level from The Cancer Genome Atlas (TCGA) program. The TCGA program is a landmark cancer genomics program with over 11,000 cases of primary cancer samples spanning 33 cancer types [18]. Additionally, the Kyoto Encyclopedia of Genes and Genomes (KEGG), as a public database, contains rich information about the graphical structures of genes [19]. It contributes to understanding various aspects of biological systems and pathways. Prostate cancer is the most common malignancy in mid-aged males and the second leading malignancy [20]. This external graphical structure is represented by a penalty weight matrix, which is the Laplacian matrix $\boldsymbol{L}$ constructed in equation (3). Ref. [21] indicated that metastatic prostate cancer remained incurable even in patients who finished intensive multimodal therapy. It is an urgent challenge to propose a novel approach for disease management via identifying prognostic determinant genes. Moreover, [22] indicated that the statistical significance of differential expression might require abundant experiments, and the probability of type I error increased as the multiple hypotheses were tested. In this article, we thus were more concerned with the accuracy of gene selection in tumor investigations (assessed by mFDR) than the power of detection.

We first removed the samples with missing measurements and then encoded PSA pathways [23] from the KEGG to construct the normalized Laplacian matrix for Grace and Grace-AKO. We finally obtained the data with sample size $n = 339$ and dimension $p = 5,947$. We denoted the PSA level as our response. To reduce computational demands, we first performed variable screening via correlation learning following [24]. The final sample size and dimension were $n = 339$ and $p = 600$, respectively, consisting of the data structure in simulation studies. We then standardized the explanatory variables and centered the response. The parameter λ_1 ranged from 1 to 40 length out as 4, and λ_2 was among 1, 2, 3, 4, and 5. We denoted the target FDR level at 0.2 and $B = 25$ for the iterations of AKO procedure. The quantile point was $\gamma = 0.3$. For a fair comparison, we both conducted 10-fold cross-validation to select the tuning parameters. We showed the details in Fig. 3.

There were 90 genes selected by Grace in total, and 47 genes selected by Grace-AKO. Ref. [2] indicated that Grace might lose some accuracy when the coefficients' signs are different. After checking the intersection of variable selection sets, we found that the genes selected by Grace-AKO were a subset of Grace's. However, due to the inflated mFDR in the simulation studies, the results of Grace might identify some false discoveries. Moreover, previous studies confirmed 35 out of 47 candidate genes detected by

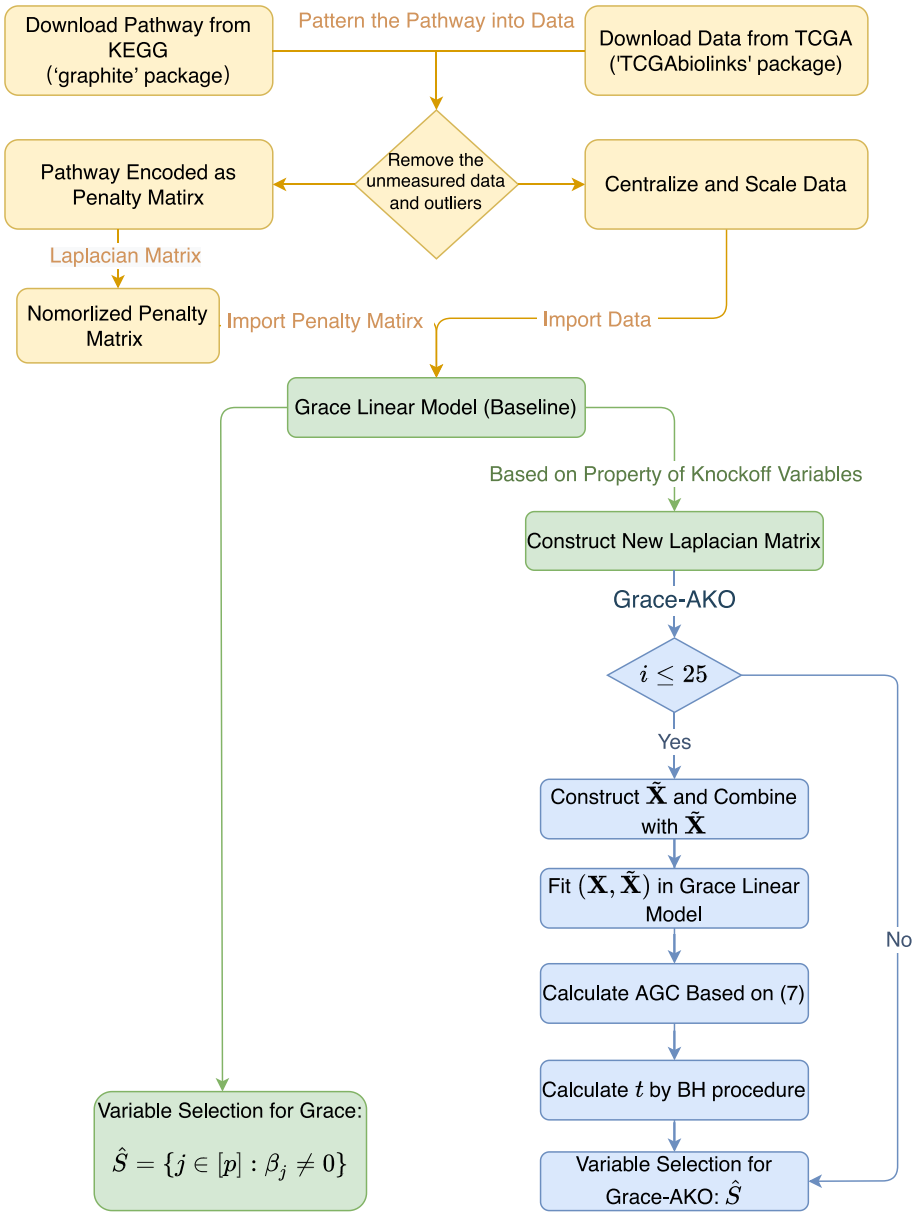

Fig. 3 Flowchart of Grace-AKO. These two procedures are differentiated by colors, in which the green one is Grace and the blue one is Grace-AKO. First, we encode the Laplacian matrix account for the network structure and sample model-X knockoffs. Second, we compute the feature statistics LCDs. Third, we aggregate the LCDs to get AGCs. Fourth, we select the variables whose AGCs satisfy the requirements

Grace-AKO were related to prostate cancer. Some details of them were listed in Table 4. For example, IL9 was recently assigned as an essential gene in tumor immunity, and AK6 was already regarded as a biomarker in prostate cancer treatments [25, 26]. HLA-DRB5 was highly related to MHC-II genes, which were the related genes of prostate cancer, and so was the CHRNB2 [27]. Moreover, a high concentration of IFNA2 was related to advanced prostate carcinoma [28]. PLK1 was over-expressed in many cancers, including prostate cancer, and scientists found that translation of the PP2A-PLK1 SDL interaction caused the expression of PLK1 and PP2A, which were commonly regarded as a

Table 4 Genes Selected by Grace-AKO in the application of prostate cancer and PSA pathways

Genes name	CHR	Genes Expression
ANAPC7	12	Ubiquitous expression in testis (RPKM 11.7), esophagus (RPKM 9.0) and 25 other tissues
AK6	5	Ubiquitous expression in testis (RPKM 18.6), adrenal (RPKM 17.2) and 25 other tissues
CHRNB2	1	Biased expression in brain (RPKM 9.5), adrenal (RPKM 0.7) and 1 other tissue
COL4A5	X	Broad expression in endometrium (RPKM 21.7), skin (RPKM 10.8) and 20 other tissues
PRLR	5	Biased expression in placenta (RPKM 18.9), endometrium (RPKM 11.6) and 8 other tissues
CAMK4	5	Broad expression in brain (RPKM 7.0), lymph node (RPKM 3.7) and 21 other tissues
CHRNB2	1	Biased expression in brain (RPKM 9.5), adrenal (RPKM 0.7) and 1 other tissue
PPP2R2D	10	Ubiquitous expression in pancreas (RPKM 6.3), ovary (RPKM 5.9) and 25 other tissues
IFNA13	9	–
COL1A1	17	Biased expression in gall bladder (RPKM 850.7), urinary bladder (RPKM 497.1) and 11 other tissues
RBPJ	4	Ubiquitous expression in placenta (RPKM 17.3), endometrium (RPKM 16.0) and 25 other tissues
PARD6G	18	Broad expression in skin (RPKM 15.6), esophagus (RPKM 5.7) and 15 other tissues
ZNF766	19	Ubiquitous expression in testis (RPKM 7.0), thyroid (RPKM 4.6) and 25 other tissues
PRKACG	9	–
SPDYE4	17	Restricted expression toward testis (RPKM 1.5)
PRKACA	19	Ubiquitous expression in heart (RPKM 62.0), adrenal (RPKM 43.9) and 25 other tissues
HLA-DRB5	6	Broad expression in lung (RPKM 275.5), lymph node (RPKM 163.3) and 14 other tissues
ZNF671	19	Ubiquitous expression in spleen (RPKM 4.7), lymph node (RPKM 4.6) and 25 other tissues
ZNF492	19	Broad expression in testis (RPKM 1.4), bone marrow (RPKM 0.9) and 16 other tissues
MAPK3	16	Ubiquitous expression in small intestine (RPKM 45.8), colon (RPKM 42.8) and 25 other tissues
ZNF461	19	Ubiquitous expression in testis (RPKM 1.8), thyroid (RPKM 1.6) and 25 other tissues
CCNE1	19	Biased expression in placenta (RPKM 16.5), bone marrow (RPKM 11.3) and 12 other tissues
CCR3	3	–
ZNF251	8	Ubiquitous expression in endometrium (RPKM 6.9), thyroid (RPKM 6.8) and 25 other tissues
CSH1	17	Restricted expression toward placenta (RPKM 9553.7)
ACTG1	17	Ubiquitous expression in ovary (RPKM 1227.2), esophagus (RPKM 970.4) and 25 other tissues
TH	11	Restricted expression toward adrenal (RPKM 42.8)
ANF853	–	–
CDC27	17	Ubiquitous expression in thyroid (RPKM 20.3), testis (RPKM 11.6) and 25 other tissues
H2AC14	6	–
H2AC7	6	–
IL9	5	–
IL5RA	3	Biased expression in lung (RPKM 1.2), prostate (RPKM 0.7) and 13 other tissues
IFNA7	9	–
IFNA2	9	–
FARR2	-	–
ZNF248	10	Ubiquitous expression in endometrium (RPKM 2.8), thyroid (RPKM 2.7) and 24 other tissues
IFNA17	9	–
ZNF331	19	Broad expression in adrenal (RPKM 22.0), placenta (RPKM 9.5) and 20 other tissues
ZNF473	19	Broad expression in testis (RPKM 14.8), spleen (RPKM 2.3) and 21 other tissues
ZNF713	7	Broad expression in brain (RPKM 2.1), testis (RPKM 1.9) and 25 other tissues
COL9A1	6	Biased expression in prostate (RPKM 6.1), testis (RPKM 1.3) and 6 other tissues
H2BC9	6	–
PDGFB	22	Broad expression in placenta (RPKM 18.6), fat (RPKM 12.7) and 23 other tissues
MYC	8	Ubiquitous expression in gall bladder (RPKM 49.6), esophagus (RPKM 44.6) and 25 other tissues

CHR stands for chromosome

biomarker in the cancer cells [29]. Ref. [30] indicated the percentage of the case with alteration of ANAPC7 achieved beyond 5.21 percentage points. The results demonstrated that the APC/C had a profound effect on cancer survival. MYC in 2010 had already been confirmed to be affected by the loss of the tumor in [31].

Furthermore, there were also some subnetworks in our findings. H2BC9, H2AC14, and H2AC7 consisted of a small subnetwork. Moreover, IFNA7, IL9, IL5RA, IFNA2, PDGFB, and KIT comprised a subnetwork of validated pathways. Except for KIT, the remaining genes were investigated as possible prostate cancer therapy genes. Additionally, there was a pathway between CAMK4 and PRKACF, where CAMKK2 was a significant androgen receptor target for prostate cancer tumor growth, according to [32]. MYC was linked with RXRA. In [33], RXRA, which was discovered as a novel target of miR-191, was conserved in a cell line derived from radio recurrent prostate cancer.

Conclusions

This article introduces Grace-AKO to perform variable selection by incorporating the network-constrained penalty [4] and AKO [7]. In contrast to the conventional variable selection process, Grace-AKO applies a normalized Laplacian matrix to encode the graphical structures between the potential genes or the gene products. It applies the L_1 penalty to selected variables and the L_2 penalty to degree-scaled differences of coefficients concerning the graphical structure. Moreover, our proposed Grace-AKO guarantees of FDR control in finite-sample settings by identifying variable employing multiple knockoff variables. Grace-AKO addresses the instability of model-X knockoffs by incorporating a statistical aggregation procedure and introducing a new feature statistics AGC. The simulation results indicated that Grace-AKO had superior performance in finite-sample FDR control in a wide range of simulation settings. In order to control the finite-sample FDR, Grace-AKO would be slightly conservative in terms of power [10]. Furthermore, the proposed general framework for variable selection with finite-sample FDR control can be broadly extended to other existing penalties (e.g., the Lasso penalty [5] and the elastic net penalty [6]).

Acknowledgements
Publication made possible in part by support from the HKU Libraries Open Access Author Fund sponsored by the HKU Libraries.

Author Contributions
Y.D.Z. and Z.L. conceived and supervised the study. P.T. and Y.H. implemented the software and conducted analysis. P.T. wrote the first draft of the manuscript. P.T., Z.L. and Y.D.Z. revised the manuscript. All authors approved the final manuscript.

Funding
This work was supported, in part, by the Hong Kong Research Grants Council (RGC) Early Career Scheme 2021/22 (project number 27305221).

Availability of data and materials
The data generated and analysis code are available in the GitHub repository https://github.com/mxxptian/GraceAKO.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 7 June 2022 Accepted: 29 October 2022
Published online: 14 November 2022

References

1. Katsevich E, Sabatti C. Multilayer knockoff filter: controlled variable selection at multiple resolutions. Ann Appl Stat. 2019;13(1):1.
2. Li C, Li H. Variable selection and regression analysis for graph-structured covariates with an application to genomics. Ann Appl Stat. 2010;4(3):1498–516.
3. Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. Stat Appl Genetics Mol Biol. 2004;3(1).
4. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008;24(9):1175–82.
5. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol). 1996;58(1):267–88.
6. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol). 2005;67(2):301–20.
7. Nguyen TB, Chevalier JA, Thirion B, Arlot S. Aggregation of multiple knockoffs. In: International Conference on Machine Learning. PMLR; 2020. p. 7283–93.
8. Candes E, Fan Y, Janson L, Lv J. Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. J R Stat Soc Ser B (Stat Methodol). 2018;80(3):551–77.
9. Barber RF, Candès EJ. Controlling the false discovery rate via knockoffs. Ann Stat. 2015;43(5):2055–85.
10. Gimenez JR, Zou J. Improving the stability of the knockoff procedure: multiple simultaneous knockoffs and entropy maximization. In: Proceedings of the twenty-second international conference on artificial intelligence and statistics. 2019 16–18 Apr;89:2184–2192. http://proceedings.mlr.press/v89/gimenez19b.html.
11. Emery K, Keich U. Controlling the FDR in variable selection via multiple knockoffs. arXiv e-prints. 2019. arXiv:1911.09442.
12. Meinshausen N, Meier L, Bühlmann P. P-values for high-dimensional regression. J Am Stat Assoc. 2009;104(488):1671–81.
13. Chung FRK. Spectral Graph Theory. vol 92. American Mathematical Society, Providence. 1997. https://books.google.co.jp/books?id=YUc38_MCuhAC.
14. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Methodol). 1995;57(1):289–300.
15. Wei Z, Li H. A Markov random field model for network-based analysis of genomic data. Bioinformatics. 2007;23(12):1537–44.
16. Wei Z, Li H. A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. Ann Appl Stat. 2008;2(1):408–29.
17. Barber RF, Candès EJ. Controlling the false discovery rate via knockoffs. Ann Stat. 2015;43(5):2055–85.
18. Koboldt D, Fulton R, McLellan M, Schmidt H, Kalicki-Veizer J, McMichael J, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
19. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
20. Stangelberger A, Waldert M, Djavan B. Prostate cancer in elderly men. Rev Urol. 2008;10(2):111–9.
21. Wang G, Zhao D, Spring DJ, DePinho RA. Genetics and biology of prostate cancer. Genes Dev. 2018;32(17–18):1105–40.
22. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003;19(3):368–75.
23. Kim I, Choi S, Kim S. BRCA-Pathway: a structural integration and visualization system of TCGA breast cancer data on KEGG pathways. BMC Bioinformatics. 2018;19(1):42. https://doi.org/10.1186/s12859-018-2016-6.
24. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B- Stat Methodol. 2008;70(5):849–911.
25. Wan J, Wu Y, Ji X, Huang L, Cai W, Su Z, et al. IL-9 and IL-9-producing cells in tumor immunity. Cell Commun Signal. 2020;18(1):50.
26. Liu C, Zhang L, Huang Y, Lu K, Tao T, Chen S, et al. MicroRNA-328 directly targets p21-activated protein kinase 6 inhibiting prostate cancer proliferation and enhancing docetaxel sensitivity. Mol Med Rep. 2015;12(5):7389–95.
27. Axelrod ML, Cook RS, Johnson DB, Balko JM. Biological consequences of MHC-II expression by tumor cells in cancer. Clin Cancer Res. 2019;25(8):2392–402.
28. Erb HH, Langlechner RV, Moser PL, Handle F, Casneuf T, Verstraeten K, et al. IL6 sensitizes prostate cancer to the antiproliferative effect of IFNα2 through IRF9. Endocr Relat Cancer. 2013;20(5):677.
29. Cunningham CE, Li S, Vizeacoumar FS, Bhanumathy KK, Lee JS, Parameswaran S, et al. Therapeutic relevance of the protein phosphatase 2A in cancer. Oncotarget. 2016;7(38):61544–61.
30. Melloy P. The Anaphase-Promoting Complex: a key mitotic regulator associated with somatic mutations occurring in cancer. Genes Chromosomes Cancer. 2019;59(3):189–202.
31. Koh CM, Bieberich CJ, Dang CV, Nelson WG, Yegnasubramanian S, De Marzo AM. MYC and prostate cancer. Genes Cancer. 2010;1(6):617–28.
32. Dadwal UC, Chang ES, Sankar U. Androgen receptor-CaMKK2 axis in prostate cancer and bone microenvironment. Front Endocrinol (Lausanne). 2018;9:335.

33. Ray H, Haughey C, Hoey C, Jeon J, Murphy R, et al. miR-191 promotes radiation resistance of prostate cancer through interaction with RXRA. Cancer Lett. 2020;473:107–17.

Publisher's Note