RESEARCH ARTICLE

Model-free feature screening for categorical outcomes: Nonlinear effect detection and false discovery rate control

Qingyang Zhang[1]*, **Yuchun Du**[2]*

1 Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR, United States of America, 2 Department of Biological Sciences, University of Arkansas, Fayetteville, AR, United States of America

* qz008@uark.edu (QZ); ydu@uark.edu (YD)

Abstract

Feature screening has become a real prerequisite for the analysis of high-dimensional genomic data, as it is effective in reducing dimensionality and removing redundant features. However, existing methods for feature screening have been mostly relying on the assumptions of linear effects and independence (or weak dependence) between features, which might be inappropriate in real practice. In this paper, we consider the problem of selecting continuous features for a categorical outcome from high-dimensional data. We propose a powerful statistical procedure that consists of two steps, a nonparametric significance test based on edge count and a multiple testing procedure with dependence adjustment for false discovery rate control. The new method presents two novelties. First, the edge-count test directly targets distributional difference between groups, therefore it is sensitive to nonlinear effects. Second, we relax the independence assumption and adapt Efron's procedure to adjust for the dependence between features. The performance of the proposed procedure, in terms of statistical power and false discovery rate, is illustrated by simulated data. We apply the new method to three genomic datasets to identify genes associated with colon, cervical and prostate cancers.

OPEN ACCESS

Citation: Zhang Q, Du Y (2019) Model-free feature screening for categorical outcomes: Nonlinear effect detection and false discovery rate control. PLoS ONE 14(5): e0217463. https://doi.org/10.1371/journal.pone.0217463

Editor: Fabio Rapallo, Universita degli Studi del Piemonte Orientale Amedeo Avogadro, ITALY

Received: August 23, 2018

Accepted: May 13, 2019

Published: May 31, 2019

Data Availability Statement: The colon cancer data can be loaded from R package plsgenomics, which is freely available at https://cran.r-project.org/web/packages/plsgenomics. The cervical cancer data can be downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20592. The prostate cancer data can be loaded from R package oompaData, which is available at https://cran.r-project.org/web/packages/oompaData.

Funding: This study was funded by the Arkansas Biosciences Institute, the major research

Introduction

Feature screening, as a key and inevitable step in many bioinformatics applications, is effective in reducing dimensionality and removing redundant features. Because the quality of selected features may greatly affect the subsequent analysis and conclusions, a reliable screening procedure is essential in practice. In general, the ideal feature screening should have high sensitivity and specificity simultaneously, as too many false positives could result in poor model interpretability while too many false negatives may cause lack of fit and inaccurate prediction. In statistics and bioinformatics literature, there has been a wealth of feature screening techniques that can be roughly classified into two categories, namely model-based screening and model-free screening. The model-based methods often rely on a class of specific models such as generalized linear model and nonparametric regression model [1–4]. However with a large

component of the Arkansas Tobacco Settlement Proceeds Act of 2000. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: BH, Benjamini-Hochberg; BHY, Benjamini-Hochberg-Yekutieli; FDR, False discovery rate; MST, Minimum spanning tree.

number of predictors, it can be very challenging to specify the model structure without prior information. The model-free methods do not require any parametric assumption or model structure, therefore they are more flexible and more efficient than model-based methods for high-dimensional data [5–7].

Different types of data require different feature screening techniques. For instance, the dependence between a continuous response and continuous features could be quantified by correlation-based measures such as Pearson's correlation, rank-based correlation, and distance correlation [8–10]. There have been a number of model-free procedures recently developed based on these measures. For instance, Li et al. (2012) developed a rank-based feature selector that is robust to outliers and influential points [5]. Li, Zhong and Zhu (2012) introduced a sure independence screening procedure based on distance correlation [6]. Another type of problem is selecting continuous features for a categorical outcome, which is more common in genomic research. For example, it is often of interest to identify genes associated with cancer or certain cancer subtype. Existing approaches for such data type mainly rely on normal-based tests such as two-sample t test (for binary response), Hotelling's t test and F test (for multi-category response) [11, 12]. These tests are powerful in detecting the mean difference between phenotypes, however, they have several major drawbacks in real genomic applications. Firstly, these tests are normal-based and only targeting linear effects, thus may fail to detect important nonlinear effects. Nonlinear relations are very common in gene regulatory network [13], therefore should be taken into account for feature screening. Secondly, existing approaches have been mostly relying on some classic multiple testing procedures to control the false discovery rate (FDR), such as Benjamini-Hochberg (BH) procedure [14]. However, such procedures control FDR only when the test statistics are independent or weakly dependent, which might not be the case in gene selection problem (genes are often strongly associated with each other). In this paper, we aimed to develop a model-free screening procedure to overcome the two challenges, namely the nonlinear effect detection and FDR control under feature dependencies. To capture nonlinear associations between a categorical response and continuous features, we transformed the problem to testing the equality of two or multiple distributions, and a recently developed nonparametric test was used to evaluate the statistical significance. In addition, we adapted Efron's multiple testing procedure to control false discovery rate with feature dependence adjustment.

The remainder of the paper is structured as follows: In Section Methods, we formulate the problem and introduce the two-step procedure including edge-count test and Efron's multiple testing procedure. In Section Results, we conduct a simulation study to evaluate the performance of the proposed procedure in terms of statistical power and false discovery rate control under various settings. The new method is applied to three real genomic datasets to search genes that differentiate cancer and normal subjects. We discuss the new method with some future work perspectives in Section Discussion and conclude the paper in Section Conclusions.

Methods

Problem formulation and edge-count test

We consider a general setting where the outcome variable is discrete with J categories ($J < \infty$) and the features are continuous. For example in genomics, the outcome response can be normal/diseased, cancer subtypes or tumor stages and each feature can be the expression level of a gene. Existing model-free screening based on correlation measures [5, 6] were developed for continuous outcomes, therefore not suitable for this problem [15]. In this paper, we introduced a novel graph-based method to select continuous features that are associated with a categorical response. Our method is model-free and does not depend on any hypothesis on the form of dependence. To begin with, let $\{X_1, \ldots, X_p\}$ be p features (p can be large), and $\{1, \ldots, J\}$ be the sampling space of response variable Y. With N independent observations of $\{Y, X_{i,1\leq i\leq p}\}$, we test the independence between Y and X_i, which is equivalent to testing equality of J conditional distributions, i.e.,

$$\begin{aligned} H_{0i} &: F_{X_i|Y=1}(x) = \ldots = F_{X_i|Y=J}(x), \text{for any } x \in \mathbb{R} \\ H_{\alpha i} &: F_{X_i|Y=j}(x) \neq F_{X_i|Y=j'}(x), \text{for some } x \text{ and } (j, j'), \end{aligned}$$

where $F_{X_i|Y=j}(x)$ stands for the cumulative distribution function of X_i in group $Y = j$. To test if H_{0i} is true, we employed a modified edge-count test which is proved more powerful in detecting difference between multiple multivariate distributions [13, 16, 17]. This test has resulted in several successful applications. For instance, Zhang (2018) [13] applied this method to search differentially co-expressed gene pairs from high-dimensional data. Zhang, Mahdi & Chen (2017) [17] employed this test to identify pathways that contribute to ovarian cancer progression. The motivation of the edge-count test is that if samples in difference groups have different distributions, they would be preferentially closer to others from the same group than those from the other group. The distance between samples can be represented by a regular similarity graph. For instance, Chen and Friedman (2017) [16] suggested a minimum spanning tree (MST) or a more general d-MST (a union of d disjoint MSTs). The edge-count test rejects the null hypothesis if the number of between-group edges in the similarity graph is significantly less than what we expected. To implement the graph-based test, we first pooled samples from all J groups and indexed them by $1, 2, \ldots, N = \sum_{j=1}^{J} n_j$. The group index for sample k was denoted by y_k. A d-MST is then constructed on the pooled samples using the standard Kruskal's algorithm [18]. Unless otherwise specified, G simultaneously represents the similarity graph and the set of all edges, while $|G|$ denotes the total number of edges throughout the paper. For the edge connecting samples k and k', i.e., (k, k'), we define R_j as the number of edges connecting samples from same group j, i.e.,

$$R_j = \sum_{(k,k')\in G} I(y_k = y_{k'} = j), \quad (1)$$

and the test statistic has the following quadratic form:

$$S := [\boldsymbol{R} - E(\boldsymbol{R})]^T \boldsymbol{V}^{-1}(\boldsymbol{R})[\boldsymbol{R} - E(\boldsymbol{R})], \quad (2)$$

where $\boldsymbol{R} = (R_1, \ldots, R_J)^T$, $\boldsymbol{V}^{-1}(\boldsymbol{R})$ represents the inverse covariance matrix of $\boldsymbol{R}$. The test statistic defined here simply quantifies the deviation of $(R_1, \ldots, R_J)$ from their expected values under permutation null, i.e., H_0^*. Chen and Friedman (2017) [16] established the asymptotic distribution of S for $J = 2$, $S \rightarrow \chi^2_{df=2}$. In our technical report [17], it was proved that the test statistics for J groups asymptotically follows a Chi-square distribution with J degrees of freedom under mild regularity conditions. To illustrate the results, for an edge e in graph G, we let

$$\begin{aligned} A_e &= \{e\} \cup \{e' \in G : e' \text{ and } e \text{ share a node}\}, \\ B_e &= A_e \cup \{e'' \in G : \exists\, e' \in A_e, \text{ such that } e'' \text{ and } e \text{ share a node}\}, \end{aligned}$$

then the following theorem can be derived:

Theorem 1. *If* $|G| = O(N)$, $\sum_{k=1}^{N} |G_k|^2 - \frac{4|G|^2}{N} = O(N)$, $\Sigma_{e\in G}|A_e||B_e| = o(N^{3/2})$, $\lim_{N\to\infty} \frac{N_j}{N} = \lambda_j \in (0, 1)$, *then*

$$S := [\boldsymbol{R} - E(\boldsymbol{R})]^T \boldsymbol{V}^{-1}(\boldsymbol{R})[\boldsymbol{R} - E(\boldsymbol{R})] \xrightarrow{\mathcal{D}} \chi^2_J,$$

where $j = 1, \ldots, J$ *is the group index.*

The expected values and covariance matrix of $(R_1, \ldots, R_J)$ can be derived as follows:

$$\begin{aligned}
E(R_j)_{1\leq j\leq J} &= |G|\frac{n_j(n_j-1)}{N(N-1)}, \\
V(R_j)_{1\leq j\leq J} &= E(R_j)(1-E(R_j)) + 2C\frac{n_j(n_j-1)(n_j-2)}{N(N-1)(N-2)} \\
&\quad +(|G|(|G|-1)-2C)\frac{n_j(n_j-1)(n_j-2)(n_j-3)}{N(N-1)(N-2)(N-3)}, \\
Cov(R_j, R_{j'})_{j\neq j'} &= (|G|(|G|-1)-2C)\frac{n_j n_{j'}(n_j-1)(n_{j'}-1)}{N(N-1)(N-2)(N-3)} - E(R_j)E(R_{j'}),
\end{aligned}$$

where $N = \sum_{j=1}^{J} n_j$ and $C = \frac{1}{2}\sum_{k=1}^{N} |G_k|^2 - |G|$.

The convergence rate of the asymptotic result is the usual $n^{-1/2}$ and there are three conditions on the similarity graph (stated in the main Theorem above). $|G| \sim O(N)$ requires that the density of the graph is of the same order as the pooled sample size. $\sum_{k=1}^{N} |G_k|^2 \sim O(N)$ ensures that there is no large hubs nor many small hubs. $\Sigma_{e\in G}|A_e||B_e| \sim o(N^{3/2})$ requires there is no cluster of small hubs [16]. These conditions are satisfied by the k-MST based on Euclidean distance [16], we therefore recommend using k-MST as the similarity graph in edge-count test. Furthermore, we conducted a simulation study to evaluate the finite sample performance of the asymptotic null distribution under different sample sizes and different similarity graphs. Details of the simulation settings can be found in S1 File, and the results were summarized in Fig T in S1 File. It is found that under two different models (standard normal distribution and exponential distribution with $\lambda = 1$), the asymptotic chi-squared distribution works quite well in approximating p-values, even for relatively small sample size, e.g, 20 samples in each group of Y. Increasing sample size generally results in better accuracy of approximation, and the use of slightly denser graph (e.g., 3-MST or 5-MST) may result in better accuracy. These findings are consistent with the simulation results for two groups ($J = 2$) [16]. For small sample sizes (e.g., $n_j \leq 10, j = 1, \ldots, J$), however, the asymptotic distribution might not work well, and in such cases, it is safer to use a permutation p-value based on our test statistic S.

It is noteworthy to mention that the main theorem also applies to multi-dimensional features (X_i can be a random vector), i.e., our method can be used to select feature sets. One interesting application is to search biological pathways or gene sets that are associated with certain phenotypes [17]. In addition to the aforementioned edge-count test, some other tests for equality of distributions may also be considered, including Kolmogorov-Smirnov (KS) test [19] and traditional graph-based test [20, 21]. However, these methods have practical limitations in real applications. For instance, KS test is known to be very conservative, i.e., the null hypothesis is too often not rejected [22, 23] (see our simulation study in S1 File for illustrating the conservativeness of KS test). Moreover when the feature is multi-dimensional, the implementation of KS test can be prohibitively computationally intensive. Graph-based tests such as the traditional edge-count tests are easy to implement but they could be problematic under certain location and scale alternatives. As reported recently [16], the traditional edge-count test works well for location alternative under low dimension, however, it becomes problematic for scale alternative (or location+scale alternative, i.e., the two distributions are different in both location and scale), especially when the dimension is moderate to high. This is caused by the fact that the number of within-sample edges in the inner layer would be larger than its null expectation, while the number of within-sample edges in the outer layer would be less than its null expectation, making the edge-count test have low or even no power [16].

Multiple testing with dependence-adjustment

As we discussed in the previous sections, the prevailing Benjamini-Hochberg procedure may fail to control the false discovery rate in the presence of moderate or strong feature dependence. In the feature screening problem, the test statistics $\{S_1, \ldots, S_p\}$ are correlated under feature dependencies, therefore the BH procedure is not appropriate. To overcome the issue, we adapted a dependence-adjusted multiple testing procedure suggested by Efron (2007) [24]. Unlike the BH procedure, Efron's procedure does not rely on the independence assumption and generally applies to any dependency structure. It has been extensively studied and widely applied by the statistic community. For instance, Liu (2013, 2017) employed this procedure as a key step to control false discovery rate in the Gaussian graphical model estimation and differential network estimation [25, 26]. To implement Efron's method, we first transformed the test statistics $\{S_1, \ldots, S_p\}$ into z-values by quantile normalization

$$z_i = \Phi^{-1}(\mathrm{P}(\chi^2_{df=J} \leq S_i)), \quad i = 1, \ldots, p,$$

where $\Phi^{-1}(\cdot)$ represents the inverse cumulative distribution function of $N(0, 1)$. Following the notations in Efron (2007), let $A = (P_0 - \hat{P}_0)/Q_0$, where $P_0 = 2\Phi(1) - 1$, $\hat{P}_0 = \sum_{i=1}^{p} I\{|z_i| \leq 1\}/p$, $Q_0 = 1/\sqrt{\pi e}$. In addition, we let

$$A(z) = \left\{1 + |A| \frac{|z|\phi(z)}{\sqrt{2}(1 - \Phi(z))}\right\}^{-1},$$

where $\phi(\cdot)$ represents the probability density function of $N(0, 1)$. Here, $A(z)$ is used to control the influence of correlation between test statistics (under independence and sparsity, $A(z)$ is close to 1, thus the procedure is same as BH procedure). The critical value can be obtained as follows:

$$z_0 = \inf\left\{-\infty < z < \infty, 1 - \Phi(z) \leq \frac{\alpha A(z)}{p} \max\,(1, \sum_{i=1}^{p} I\{z_i \geq z\})\right\}.$$

To control the FDR at the level of α (e.g., $\alpha = 0.05$ or $\alpha = 0.10$), one can solve for the cutoff z_0 and reject H^*_{i0} if $z_i > z_0$. This testing procedure asymptotically controls the FDR at the desired level under some mild regularity conditions (though it might be slightly conservative for some cases) and it works well under all settings in our simulation study. The detailed proof and regularity conditions for Gaussian case can be found in Liu (2017) ([26], see Theorems 3.1 and 3.3).

Results

Simulation studies

The simulation studies in this part examined the performance of the proposed procedure under several different settings. Without loss of generality, we considered a binary outcome variable $Y \in \{0, 1\}$ (i.e., $J = 2$) and p continuous features $\{X_1, \ldots, X_p\}$ with sample size N ($p \geq N$). Four high-dimensional settings (each setting refers to a combination of model and feature

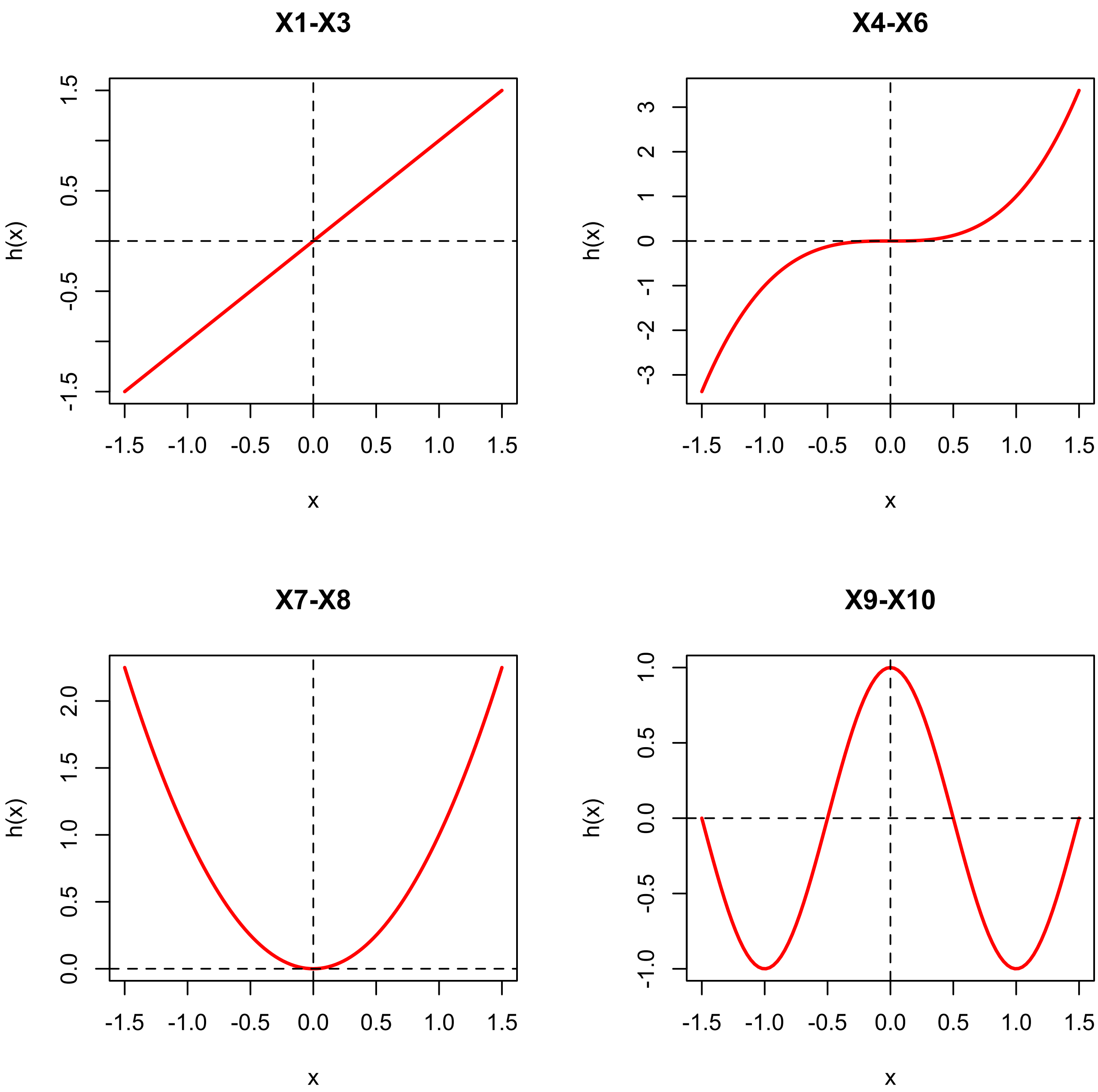

Fig 1. Four transformation functions in the simulation study.

https://doi.org/10.1371/journal.pone.0217463.g001

dependency structure) were used to generate the data. To be precise, let $k \in \{1, 2, \ldots, N\}$ be the index of sample, and $i \in \{1, 2, \ldots, p\}$ be the index of feature, where we set $p = 500$ and $N = 50$, 100, 200, 500 respectively. In addition, we assumed that only the first 10 features, $\{X_1, \ldots, X_{10}\}$, were associated with Y and the other 490 features were redundant. The transformation functions $\{h_i(X_{ik}), 1 \leq i \leq 10\}$ were set as $h_i(X_{ik}) = X_{ik}$ for $1 \leq i \leq 3$ (linear transformation), $h_i(X_{ik}) = X_{ik}^3$ for $4 \leq i \leq 6$ (nonlinear monotonic transformation), $h_i(X_{ik}) = X_{ik}^2$ for $7 \leq i \leq 8$ (nonlinear non-monotonic transformation) and $h_i(X_{ik}) = \sin(2\pi X_{ik}/3)$ for $9 \leq i \leq 10$ (nonlinear non-monotonic transformation), representing a combination of linear effects and nonlinear effects. The four transformation curves were shown in Fig 1.

To establish the relation between Y and $\{X_1, \ldots, X_{10}\}$, we considered two different models:

- Logistic regression model: $Y_k \sim Bernoulli(\pi_k)$, $\log\{\pi_k/(1-\pi_k)\} = \sum_{i=1}^{10} \beta_i h_i(X_{ik})$, $\beta_1 = \beta_2 = \beta_4 = \beta_6 = \beta_7 = \beta_9 = 0.5$, and $\beta_3 = \beta_5 = \beta_8 = \beta_{10} = -0.5$
- Latent variable model: $Y_k = I\{Y_k^* > 0\}$, where $Y_k^* = \sum_{i=1}^{10} \beta_i h_i(X_{ik}) + \epsilon_k$, $\epsilon_k \sim N(0, 0.5^2)$, $\beta_1 = \beta_2 = \beta_4 = \beta_6 = \beta_7 = \beta_9 = 0.5$, and $\beta_3 = \beta_5 = \beta_8 = \beta_{10} = -0.5$

Furthermore, to evaluate the effect of feature dependencies on statistical power and FDR control, we generated the data by two methods:

- Independent features: $X_{ik} \sim \text{Unif}(-1.5, 1.5)$ for $1 \le i \le 500$.
- Dependent features: $X_{ik} = \sqrt{2}Z_{ik}$, where $\{Z_{ik}\}_{1 \le i \le 500} \sim \boldsymbol{N}_{500}(\boldsymbol{0}, \Sigma)$ and Σ is a random correlation matrix containing both positive and negative elements (generated by R package *clusterGeneration*). In addition, we conducted an interval truncation (between -1.5 and 1.5) for the samples to avoid extreme values.

The following six testing procedures were applied to each combination of model and feature dependency structure above, namely logistics model with independent features, logistic model with dependent features, latent variable model with independent features and latent variable model with dependent features:

- Edge-count test with Efron's multiple testing procedure
- Edge-count test with Benjamini-Hochberg procedure
- Welch's t test with Efron's multiple testing procedure
- Welch's t test with Benjamini-Hochberg procedure
- Mutual information z-test with Efron's multiple testing procedure
- Mutual information z-test with Benjamini-Hochberg procedure

In the edge-count test, a 3-MST was constructed as the similarity graph for better approximation of p-values [17]. To implement Welch's t test with dependence-adjusted multiple testing, we first calculated and transformed the test statistics into z values via quantile normalization:

$$z_i = \Phi^{-1}(\mathrm{P}(t_{df=v_i} \le t_i)), \quad i = 1, \ldots, p,$$

where the degree of freedom v_i was approximated by Welch-Satterthwaite equation and the test statistics t_i was calculated by the standard formula for t test with unequal variances:

$$v_i = \frac{\left(\frac{s_{i1}^2}{n_1} + \frac{s_{i0}^2}{n_0}\right)^2}{\frac{s_{i1}^4}{n_1^2(n_1-1)} + \frac{s_{i0}^4}{n_0^2(n_0-1)}}, \quad t_i = \frac{\overline{X}_{i1} - \overline{X}_{i0}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i0}^2}{n_0}}},$$

where $\{n_1, n_0\}$ stand for the sample sizes for $Y = 1$ and $Y = 0$, $\{\overline{X}_{i1}, \overline{X}_{i0}\}$ and $\{s_{i1}^2, s_{i0}^2\}$ represent the sample means and sample standard deviations of X_i in two groups, respectively.

To test whether the mutual information is zero, we used the following Fisher-z transformation:

$$z_i^{MI} = \frac{1}{2} \log \frac{1 + \hat{I}^*(Y, X_i)}{1 - \hat{I}^*(Y, X_i)}, \tag{3}$$

where $\hat{I}^{*}(Y, X_i)$ represents the normalized sample mutual information between response Y and X_i, and it can be computed as $\hat{I}^{*}(Y, X_i) = \hat{I}(Y, X_i)/(\hat{H}(Y) + \hat{H}(X_i))$, where $\hat{I}(Y, X_i)$ stands for the sample mutual information between Y and X_i, and $\{\hat{H}(Y), \hat{H}(X_i)\}$ stand for the sample entropies of Y and X_i. By the classical decision theory, $z_i^{MI} \sim N(0, 1/\sqrt{N-3})$ under the null hypothesis [27, 28]. The sample mutual information and sample entropy were obtained by R package *infotheo* (https://cran.r-project.org/web/packages/infotheo), where the continuous X_i was discretized into $N^{1/3}$ bins.

The targeted FDR was chosen to be $\alpha = 0.10$. Figs 2 and 3 summarized the empirical statistical power and false discovery proportion by six procedures based on 100 replications. It can be seen that the edge-count test was superior to Welch's t test and mutual information test in both false discovery rate control and statistical power under all settings. Notably, the edge-count test showed a substantial power gain (ranging from 0.17 $\sim$ 0.44) over other tests. For independent features, the BH procedure and Efron's procedure performs very similar in FDR control. However, under feature dependence, the BH procedure is slightly worse than Efron's procedure for all tests.

Fig 4 presented an illustrative example where feature X_7 was missed by Welch's t test and mutual information test but captured by the edge-count test in our simulation. The reason is that feature X_7 has a quadratic effect ($h_7(X_7) = X_7^2$) on Y, and the difference between two sample means (vertical dashed lines) becomes subtle and undetectable. However, feature X_7 showed very different patterns in two groups (a clear bimodal shape in $Y = 1$ and much weaker bimodal shape in $Y = 0$) which was detected by the edge-count test. Fig 5 showed an example of false negative where the feature was missed by all methods due to a small difference in both sample mean and sample distributions.

Application to three cancer genomic datasets

We first applied the new procedure to a colon cancer dataset [29] to search genes that differentiate cancer and normal subjects. The data contained expression level of 2,000 genes in 40 tumor and 22 colon tumor samples, probed by oligonucleotide arrays. To reduce variance and remove potential effects, the data for each subject were first log-transformed and then normalized by the trimmed mean and trimmed standard deviation (the lowest and highest 5% data were excluded). Two procedures were compared in selecting differentially expressed genes in two groups, including the edge-count method with Efron's multiple testing procedure (a 3-MST was used as the similarity graph) and Welch's t test with Benjamini-Hochberg procedure, both with targeted FDR $\alpha = 0.10$. As can be seen from our simulation results (Figs 2 and 3), when the sample sizes are relatively small ($N = 50$), the mutual information z-test exhibited extremely low power, therefore we did not consider this method for real data analysis.

Out of 2,000 genes, 36 and 26 genes were selected by the two methods and Fig 6 showed a Venn diagram summarizing the agreement between two selections. As shown in Fig 6, most of the 26 genes by Welch's t test were also captured by the edge-count test, but a list of 11 genes that were identified by edge-count test were missed by the Welch's t test, which included genes *Hsa.3180, Hsa.1804, Hsa.40177, Hsa.4937, Hsa.2157, Hsa.44676, Hsa.2847, Hsa.3026, Hsa.108, Hsa.11632, Hsa.27716*. Figs 7 and 8 presented the expression levels of two such genes, including *Hsa.108* and *Hsa.2157*, where the sample distributions in normal and tumor groups were significantly different from each other but both skewed. Our edge-count test successfully detected this difference, while the Welch's t test failed to detect it due to close sample means (indicated by the two vertical dashed lines). Similar results were observed for the other nine genes (see Figs B-J in S1 File for details).

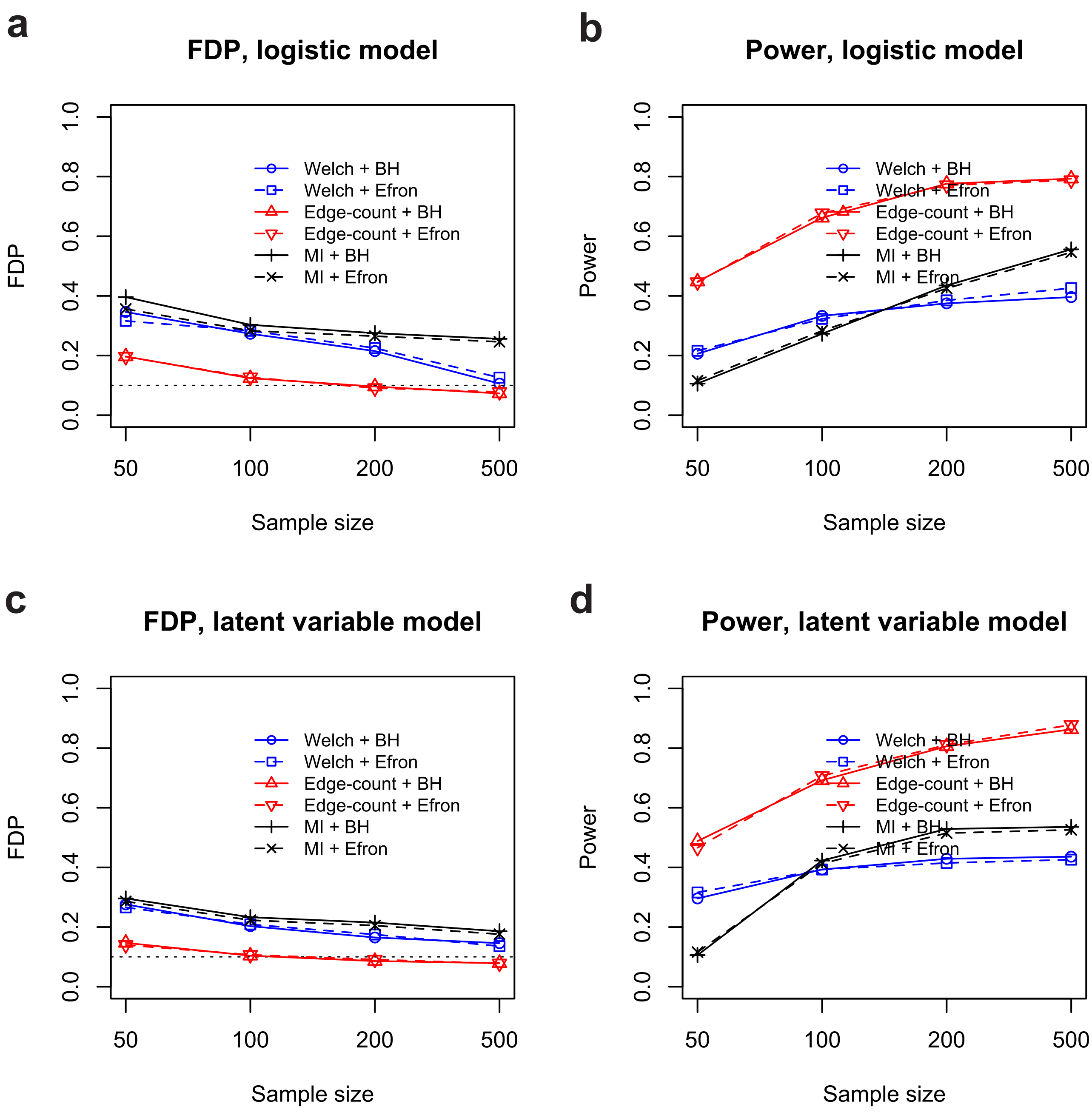

Fig 2. False discovery proportions and empirical statistical powers by six different procedures under independent features: (a) false discovery proportion for logistic model; (b) statistical power for logistic model; (c) false discovery proportion for latent variable model; (d) statistical power for latent variable model. All results were based on 100 replications.

https://doi.org/10.1371/journal.pone.0217463.g002

As previously reported in the literature, several of these 11 genes are associated with human cancers. To name a few, gene *Hsa.1804 (SFN)* promotes lung adenocarcinoma progression at an early stage [30]. Gene *Hsa.4937 (CREBBP)* acts as a potent tumor suppressor in small cell lung cancer, and inactivation of *CREBBP* enhances responses to a targeted therapy [31]. Gene *Hsa.44676 (VAV1)* promotes cancer growth by instigating tumor-microenvironment cross-talk via growth factor secretion [32]. Gene *Hsa.108 (POSTN)*, a matricellular protein-coding

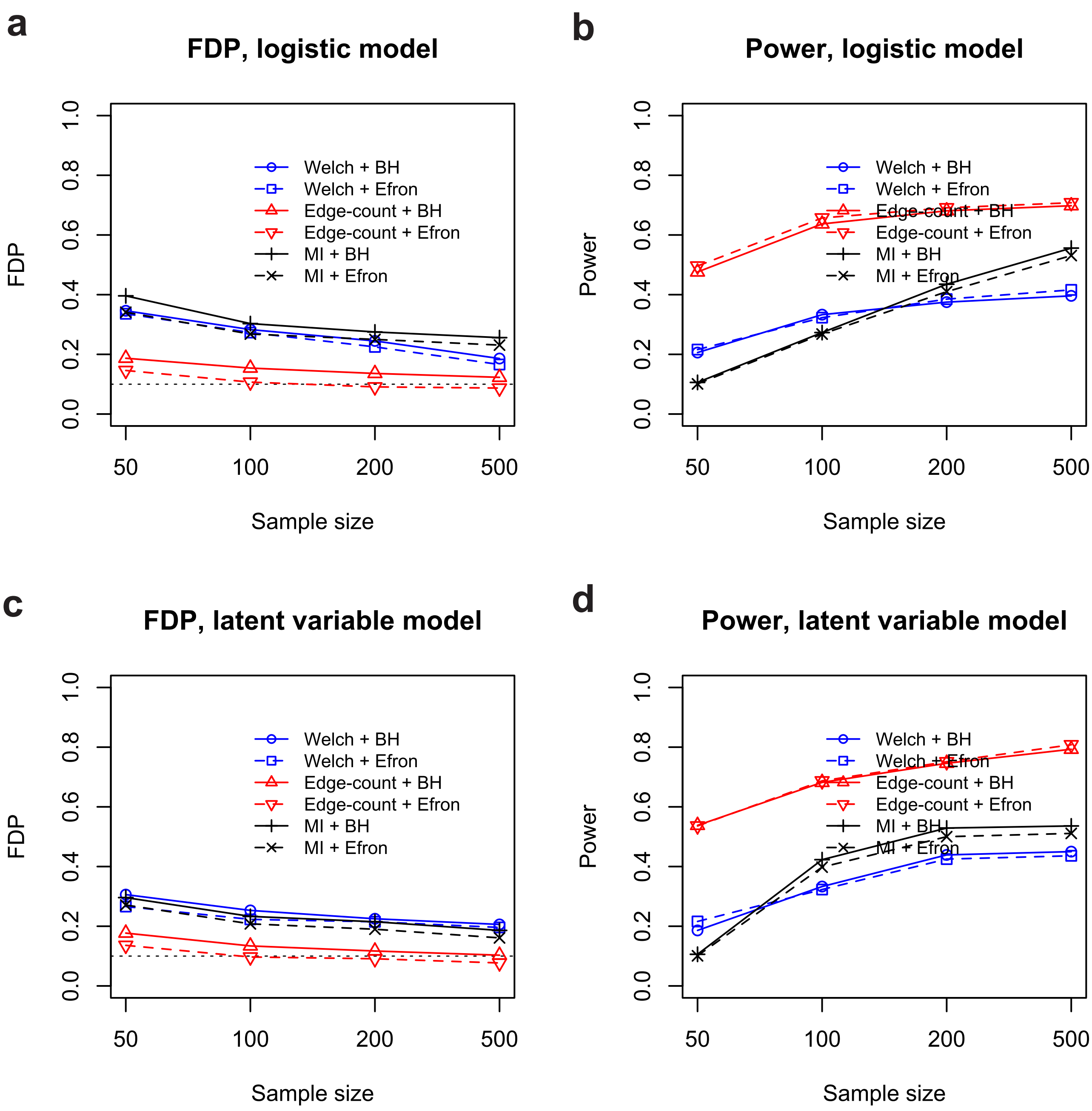

Fig 3. False discovery proportions and empirical statistical powers by six different procedures under dependent features: (a) false discovery proportion for logistic model; (b) statistical power for logistic model; (c) false discovery proportion for latent variable model; (d) statistical power for latent variable model. All results were based on 100 replications.

https://doi.org/10.1371/journal.pone.0217463.g003

gene, has been shown to regulate key aspects of tumor biology, including proliferation, invasion, matrix remodeling, and dissemination to pre-metastatic niches in distant organs [33]. Gene *Hsa.11632 (RYR1)*, together with *RYR2* stimulates apoptosis of prostate cancer cells [34].

The results from colon cancer data well confirmed our findings from simulation study, i.e., the edge-count test can not only detect the mean difference, but also detect distributional differences, thus it is more sensitive to nonlinear change compared to normal-based tests such as

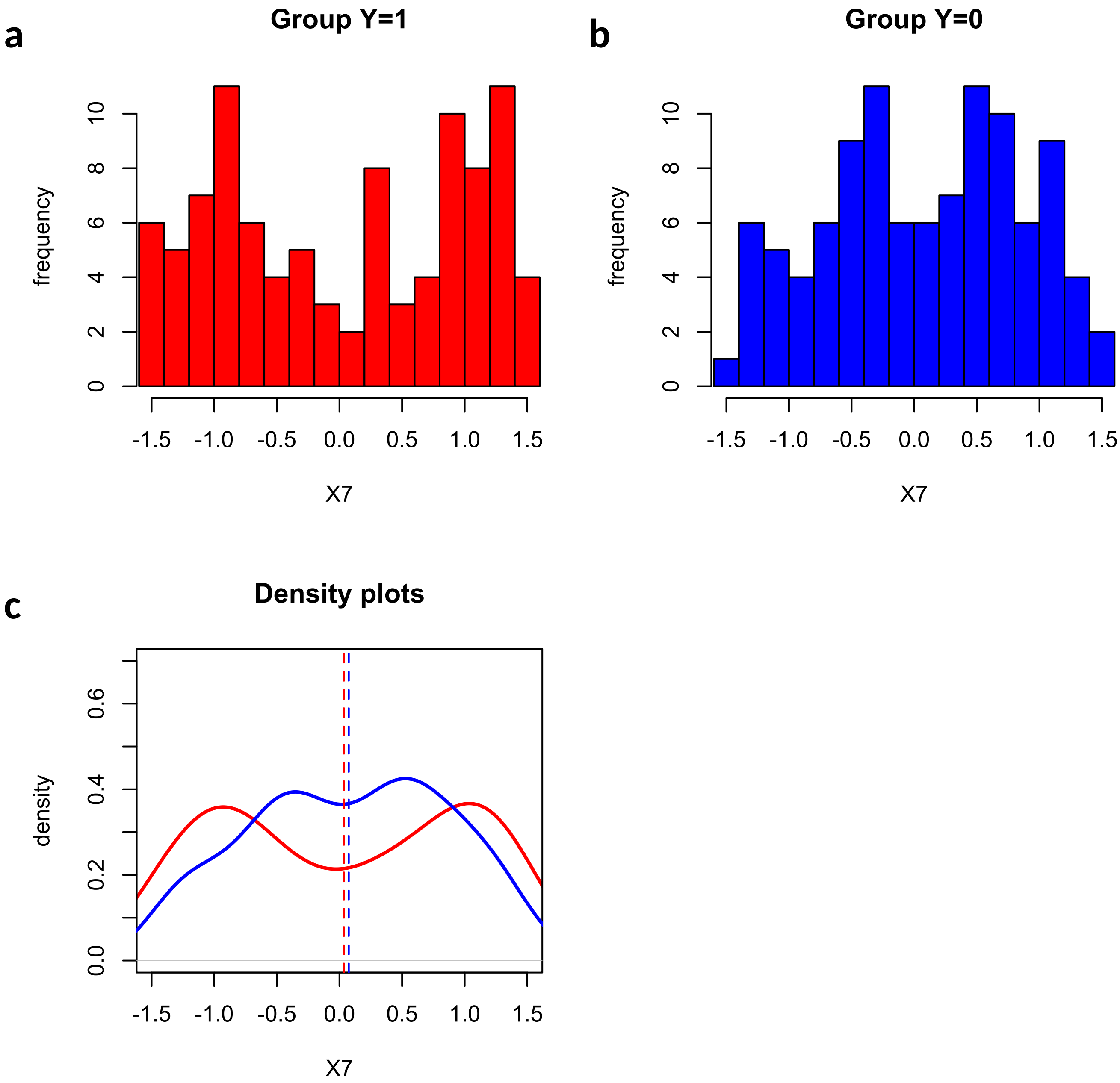

Fig 4. An example that feature X_7 was captured by edge-count test but missed by Welch's t test: (a) histogram of X_7 in group $Y = 1$; (b) histogram of X_7 in group $Y = 0$; (c) comparison of two fitted density curves, where the vertical dashed lines indicate the sample means in two groups.

https://doi.org/10.1371/journal.pone.0217463.g004

t test, F test and Hotelling's t test. Additionally, we conducted feature selection using p-values from a simple logistic regression (implemented by R function *glm()*), followed by a Benjamini-Hochberg procedure with $\alpha = 0.10$. We detected a total of 28 significant genes, and 26 of them were consistent with the selection by Welch's t test. However, this model fails to detect any of the 11 genes with nonlinear effects. The logistic regression model was further modified by adding a quadratic term in order to capture the nonlinear relations, however, this modification did not lead to any improvement.

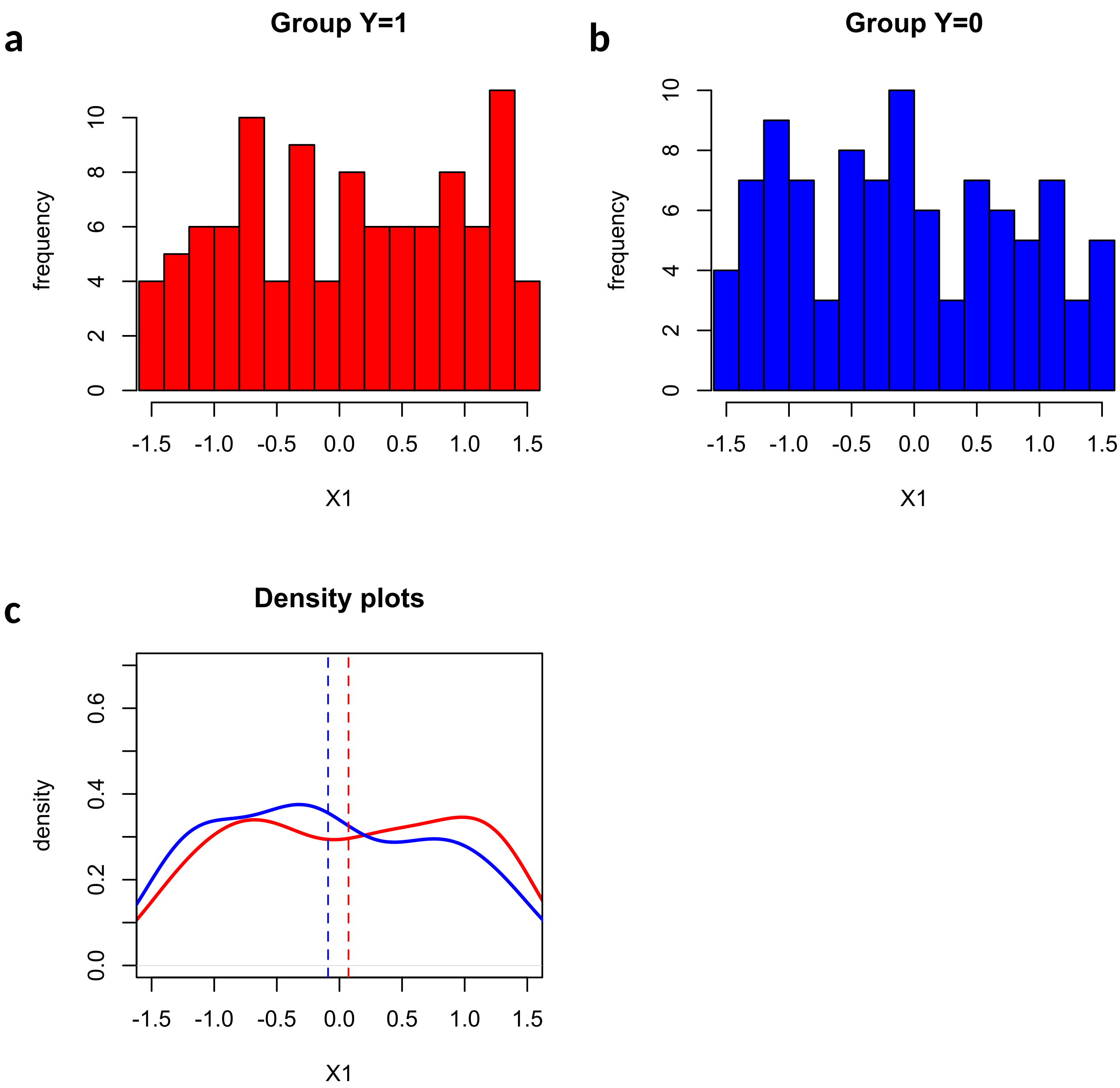

Fig 5. An example that feature X_7 was missed by both edge-count test and Welch's t test: (a) histogram of X_7 in group $Y = 1$; (b) histogram of X_7 in group $Y = 0$; (c) comparison of two fitted density curves, where the vertical dashed lines indicate the sample means in two groups.

https://doi.org/10.1371/journal.pone.0217463.g005

The new method was further tested on two additional cancer genomic datasets, including the RNA-seq data for cervical cancer [35] and the microarray data for prostate cancer [36] (see S1 File for details about data analysis). Similar to the results from the colon cancer data, the edge-count test consistently detected more genes than the Welch's t test (in the cervical cancer data, the new method identified 16 more genes and in the prostate cancer, the new method identified 12 more genes). All the newly discovered genes have close sample means but significantly different distributions in normal and tumor groups. The details of nine such genes were shown in S1 File, see Figs K-S in S1 File.

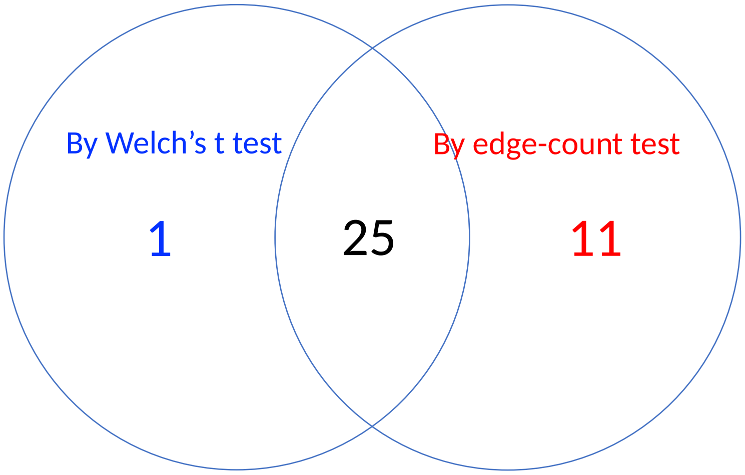

Fig 6. A Venn diagram showing the agreement between two selections by Welch's t test (with BH procedure) and edge-count test (with Efron's multiple testing procedure).

https://doi.org/10.1371/journal.pone.0217463.g006

Discussion

Genomic studies with high-dimensional data often rely on feature screening. In this work, we developed and validated a model-free feature screening method which reliably selects continuous features associated with a categorical outcome under high dimension. The new method tackles two major challenges in feature screening and feature selection, namely nonlinear effect detection and false discovery rate control under feature dependencies. The edge-count test is based on some simple calculations such as MST construction and Chi-square test, therefore it is easy-to-implement and feasible for large-scale data sets such as cancer genomic data and brain mapping data. For instance, in the colon cancer example with 2,000 genes, the computation took less than 10 seconds by R implementation on single CPU (2.5 GHz Intel Core i7).

There are several possible extensions of the proposed selector. For instance, in addition to feature screening, our method can also be used to select feature sets. One appealing property of the edge-count test is that it only requires a similarity graph constructed on the samples. In practice, one could simply build a MST or m-MST based on Euclidean distance as the similarity graph, and the main result $S_i \to \chi^2_{df=J}$ holds regardless of the sizes of feature sets. This extension can be used to search important pathways associated with certain disease, which is biologically more interesting than single gene based selection as the pathway-level analysis provides more functional insights into the mechanism underlying the phenotype change.

Efron's multiple testing procedure was used in our method to control FDR under feature dependencies, but it might be replaced by other recently developed procedures. For instance, when the test statistics are positively dependent, one may also use Benjamini-Hochberg-Yekutieli (BHY) procedure to control FDR [37]. Fan et al. (2012) introduced a new multiple testing based on principal factor approximation, which adjusts the feature dependencies of arbitrary structure [38]. However, Fan et al.'s method relied on the true covariance matrix of the test statistics, which is unknown in most cases. To obtain a good sample covariance matrix of the test statistics $\{S_1, \ldots, S_p\}$ in our framework, a subsampling without replacement might be needed in order to get independent samples of $\{S_1, \ldots, S_p\}$, however, the estimation may require relatively large sample size, e.g., $N > 1,000$.

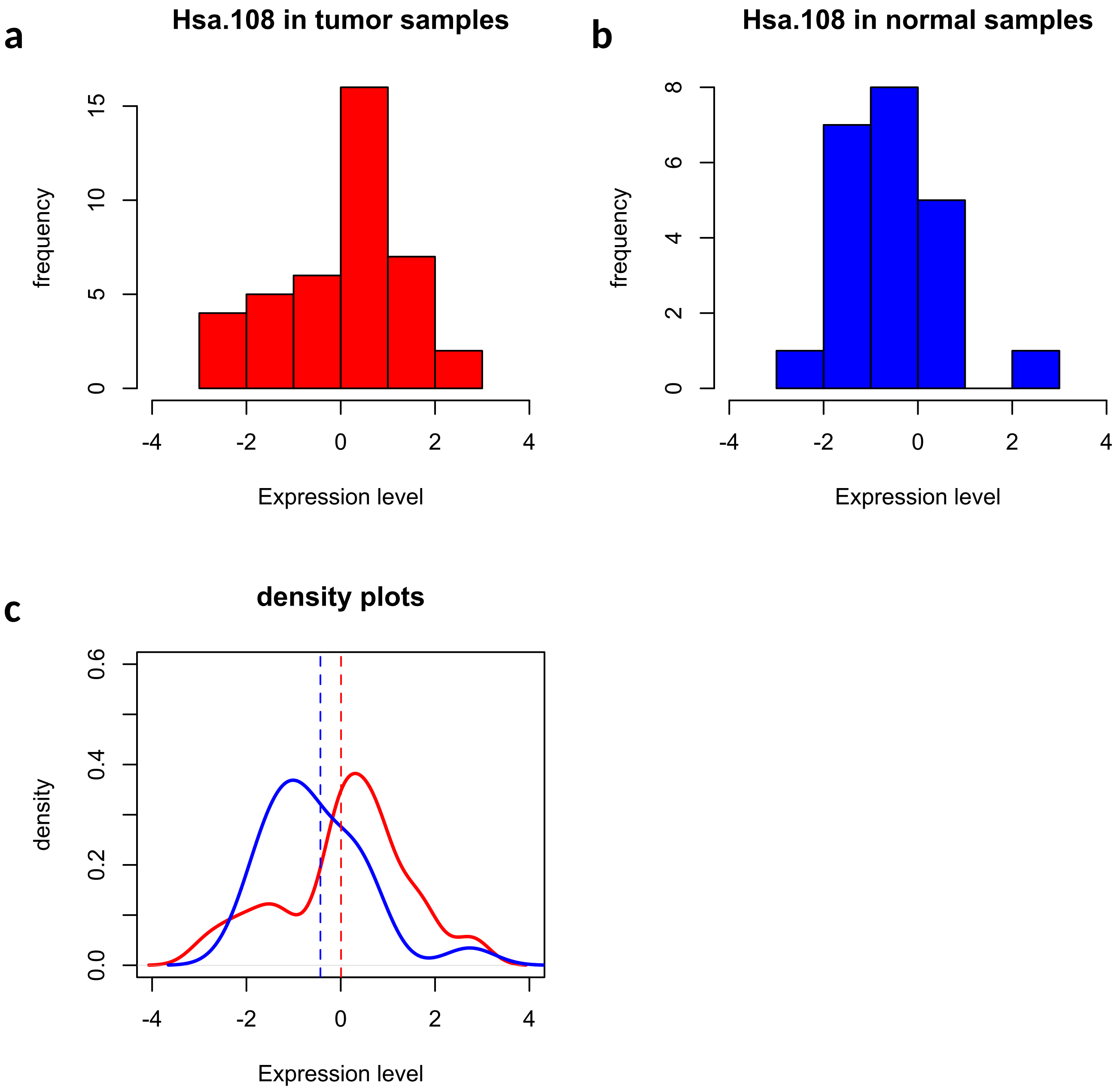

Fig 7. An example that gene *Hsa.108* was selected by edge-count test but missed by Welch's t test: (a) histogram of gene *Hsa.108* in tumor samples; (b) histogram of gene *Hsa.108* in normal samples; (c) comparison of two fitted density curves, where the vertical dashed lines indicate the sample means in two phenotypic groups.

https://doi.org/10.1371/journal.pone.0217463.g007

Conclusions

Identification of disease-related biomarkers from large-scale data is essential in many genomic studies. However, existence of nonlinear effects and strong feature dependencies make existing methods inappropriate and unreliable. In this work, we presented a model-free feature screening method which is sensitive to both linear and nonlinear effects. In addition, the dependence-adjusted multiple testing procedure can well control the false discovery rate under feature dependencies. On a whole, we put forward a simple yet effective testing procedure that reliably captures different types of effects. Although we used gene expression data for

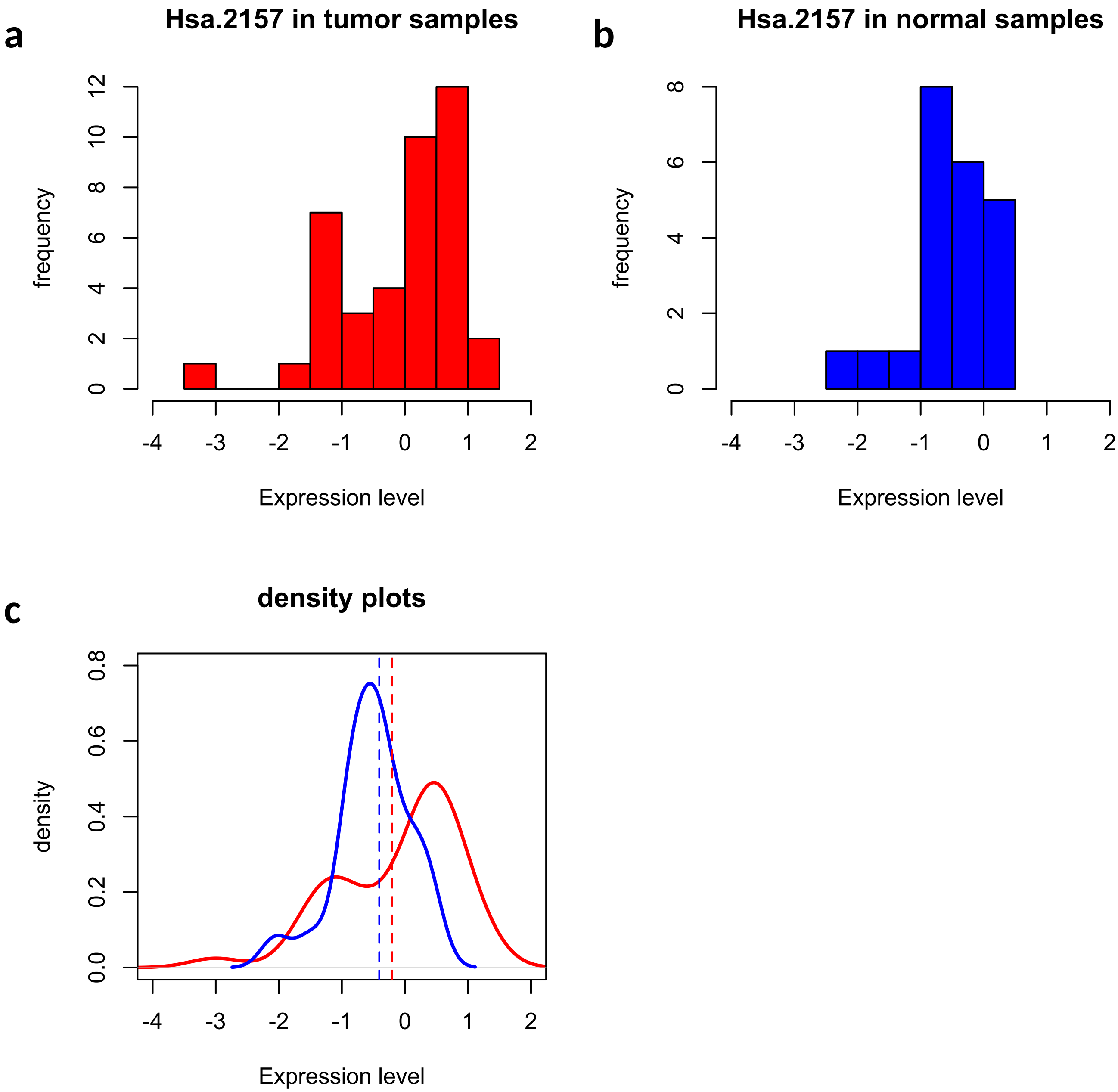

Fig 8. An example that gene *Hsa.2157* was selected by edge-count test but missed by Welch's t test: (a) histogram of gene *Hsa.2157* in tumor samples; (b) histogram of gene *Hsa.2157* in normal samples; (c) comparison of two fitted density curves, where the vertical dashed lines indicate the sample means in two phenotypic groups.

https://doi.org/10.1371/journal.pone.0217463.g008

illustration in the paper, the proposed test can be readily applied to other data types and problems, such as DNA methylation data and protein expression data and pathway selection.

Supporting information

S1 File. Additional analyses. This file contains additional simulation studies and real data applications, as well as the technical report by Zhang, Mahdi and Chen.
(PDF)

Author Contributions

Conceptualization: Qingyang Zhang.

Methodology: Qingyang Zhang, Yuchun Du.

Validation: Qingyang Zhang.

Writing – original draft: Qingyang Zhang.

Writing – review & editing: Qingyang Zhang, Yuchun Du.

References

1. Guo C, Yang H, Lv J. Robust variable selection for generalized linear models with a diverging number of parameters. Comm Stat—Theo & Meth. 2017 Oct; 46(6):2967–2981. https://doi.org/10.1080/03610926.2015.1053940
2. Li Z, Wang S, Lin X. Variable selection and estimation in generalized linear models with the seamless L0 penalty. Canadian J Stat. 2012 Jan; 40(4): 745–769. https://doi.org/10.1002/cjs.11165
3. Gertheiss J, Maity A, Staicu A. Variable selection in generalized functional linear models. Stat. 2013; 2(1): 86–101. https://doi.org/10.1002/sta4.20 PMID: 25132690
4. Tsagris M, Lagani V, Tsamardinos I. Feature selection for high-dimensional temporal data. BMC Bioinformatics. 2018 June; 19(17) https://doi.org/10.1186/s12859-018-2023-7 PMID: 29357817
5. Li G, Peng H, Zhang J, Zhu L. Robust rank correlation based screening. Ann Stat. 2012; 40(3): 1846–1877 https://doi.org/10.1214/12-AOS1024
6. Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Amer Stat Assoc. 2012; 107(499) https://doi.org/10.1080/01621459.2012.695654
7. Zhang Q, Burdette J, Wang J. Integrative network analysis of TCGA data for ovarian cancer. BMC Syst Biol. 2014; 8(1338): 1–18.
8. Szekely G, Rizzo M, Bakirov N. Measuring and testing dependence by correlation distances. Ann Stat. 2007; 35: 2769–2794 https://doi.org/10.1214/009053607000000505
9. Szekely G, Rizzo M. Brownian distance covariance. Ann Appl Stat. 2009; 3: 1233–1303 https://doi.org/10.1214/09-AOAS312REJ
10. Szekely G, Rizzo M. The distance correlation t-test of independence in high dimension. J Mult Anal. 2013; 117: 193–213 https://doi.org/10.1016/j.jmva.2013.02.012
11. Zhou N, Wang L. A Modified T-test Feature Selection Method and Its Application on the HapMap Genotype Data. Genot, Proteo & Bioinf. 2007; 5(3): 242–9 https://doi.org/10.1016/S1672-0229(08)60011-X
12. Lu Y, Liu P, Xiao P, Deng H. Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays. Bioinformatics. 2005; 21(14): 3105–3113 https://doi.org/10.1093/bioinformatics/bti496 PMID: 15905280
13. Zhang Q. A powerful nonparametric method for detecting differentially co-expressed genes: distance correlation screening and edge-count test. BMC Syst Biol. 2018; 12(58): 1–16
14. Benjamini Y, Hochberg L. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Proc Nat Acad Sci. 1995; 96: 6745–6750
15. Agresti A. An introduction to categorical data analysis Wiley-Interscience.: 2006
16. Chen H, Friedman J. A new graph-based two-sample test for multivariate and object data. J Amer Stat Assoc. 2017; 112: 397–409. https://doi.org/10.1080/01621459.2016.1147356
17. Zhang Q, Mahdi G, Chen H. A graph-based multi-sample test for identifying pathways associated with cancer progression. Technical Report. 2017
18. Cheriton D, Tarjan R. Finding minimum spanning trees. SIAM J Comp. 2006; 5(4): 724–742. https://doi.org/10.1137/0205051
19. Lopes R, Hobson P, Reid I. Computationally efficient algorithms for the two-dimensional Kolmogorov-Smirnov test. J Phys: Conf Series. 2008; 19(4)
20. Friedman J, Rafsky L. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann Stat. 1979; 7(4): 697–717 https://doi.org/10.1214/aos/1176344722
21. Rosenbaum P. An exact distribution-free test comparing two multivariate distributions based on adjacency. J Royal Stat Soc B. 2005; 67(4): 515–530 https://doi.org/10.1111/j.1467-9868.2005.00513.x

22. Steinskog D, Tjostheim D, Kvamsto N. A Cautionary Note on the Use of the Kolmogorov-Smirnov Test for Normality. Monthly Weather Rev. 2007; 135(3): 1151–1157 https://doi.org/10.1175/MWR3326.1

23. Crutcher H. A Note on the Possible Misuse of the Kolmogorov-Smirnov Test. J Appl Met. 1975; 14(8): 1600–1603 https://doi.org/10.1175/1520-0450(1975)014%3C1600:ANOTPM%3E2.0.CO;2

24. Efron B. Correlation and large-scale simultaneous significance testing. J Amer Stat Assoc. 2007; 102: 93–103 https://doi.org/10.1198/016214506000001211

25. Liu W. Gaussian graphical model estimation with false discovery rate control. Ann Stat. 2013; 41(6): 2948–2978 https://doi.org/10.1214/13-AOS1169

26. Liu W. Structural similarity and difference testing on multiple sparse Gaussian graphical models. Ann Stat. 2017; 45(6): 2680–2707 https://doi.org/10.1214/17-AOS1539

27. Kalisch M, Buhlmann P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Lear Res. 2007; 8: 613–636

28. Zhang X, Zhao X, He K, Lu L, Cao Y, Liu J. et al. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics. 2012; 28(1): 98–104 https://doi.org/10.1093/bioinformatics/btr626 PMID: 22088843

29. Alon U, Barkai N, Notterman D, Gish K, Ybarra S, Mack D, Levine A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Nat Acad Sci. 1999 June; 96(12): 6745–6750 https://doi.org/10.1073/pnas.96.12.6745 PMID: 10359783

30. Shiba-Ishii A, Kim Y, Shiozawa T, Iyama S, Satomi K, Kano J. et al. Stratifin accelerates progression of lung adenocarcinoma at an early stage. Mol Cancer. 2015 July; 14(142): 1–6

31. Jia D, Augert A, Kim D, Eastwood E, Wu N, Ibrahim A. et al. Crebbp loss drives small cell lung cancer and increases sensitivity to HDAC inhibition. Cancer Disc. 2018 May; 8(11) https://doi.org/10.1158/2159-8290.CD-18-0385

32. Sebban S, Farago M, Rabinovich S, Lazer G, Idelchuck Y, Ilan L. et al. Vav1 promotes lung cancer growth by instigating tumor-microenvironment cross-talk via growth factor secretion. Oncotarget. 2014; 5(19): 9214–9226 https://doi.org/10.18632/oncotarget.2400 PMID: 25313137

33. Gonzalez-Gonzalez L, Alonso J. Periostin: A Matricellular Protein With Multiple Functions in Cancer Development and Progression. Frontiers in Oncology. 2018; 8(225) https://doi.org/10.3389/fonc.2018.00225 PMID: 29946533

34. Mariot P, Prevarskaya N, Roudbaraki M, Le Bourhis X, Van Coppenolle F, Vanoverberghe K. et al. Evidence of functional ryanodine receptor involved in apoptosis of prostate cancer (LNCaP) cells. Prostate. 2000; 43(3): 205–214 https://doi.org/10.1002/(SICI)1097-0045(20000515)43:3%3C205::AID-PROS6%3E3.0.CO;2-M PMID: 10797495

35. Witten D, Tibshirani R, Gu S, Fire A, Lui W. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC Biology. 2010; 8(58) https://doi.org/10.1186/1741-7007-8-58 PMID: 20459774

36. Lapointe J, Li C, Higgins J, van de Rijn M, Bair E, Montegomery K, et al. (2004), Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Nat Acad Sci. 2004; 101(3): 811–816. https://doi.org/10.1073/pnas.0304146101 PMID: 14711987

37. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001 May; 29(4), 1165–1188

38. Fan J, Han X, Gu W. Estimating false discovery proportion under arbitrary covariance dependence. J Amer Stat Assoc. 2012 Jan; 40(4): 745–769.