Computational Biology and Chemistry 87 (2020) 107285

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

A graph-based multi-sample test for identifying pathways associated with cancer progression

Qingyang Zhang[a,*,1], Ghadeer Mahdi[a,b,1], Jian Tinker[a,1], Hao Chen[c,*,1]

[a] Department of Mathematical Sciences, University of Arkansas, USA
[b] Department of Mathematics, College of Education, Baghdad University, Iraq
[c] Department of Statistics, University of California at Davis, USA

ARTICLE INFO

Keywords:
Edge-count test
Tumorigenesis
Serous ovarian cancer
Pathway analysis
The Cancer Genome Atlas

ABSTRACT

Cancer is in general not a result of an abnormality of a single gene but a consequence of changes in many genes, it is therefore of great importance to understand the roles of different oncogenic and tumor suppressor pathways in tumorigenesis. In recent years, there have been many computational models developed to study the genetic alterations of different pathways in the evolutionary process of cancer. However, most of the methods are knowledge-based enrichment analyses and inflexible to analyze user-defined pathways or gene sets. In this paper, we develop a nonparametric and data-driven approach to testing for the dynamic changes of pathways over the cancer progression. Our method is based on an expansion and refinement of the pathway being studied, followed by a graph-based multivariate test, which is very easy to implement in practice. The new test is applied to the rich Cancer Genome Atlas data to study the (epi)genetic alterations of 186 KEGG pathways in the development of serous ovarian cancer. To make use of the comprehensive data, we incorporate three data types in the analysis representing gene expression level, copy number and DNA methylation level. Our analysis suggests a list of nine pathways that are closely associated with serous ovarian cancer progression, including cell cycle, ERBB, JAK-STAT signaling and p53 signaling pathways. By pairwise tests, we found that most of the identified pathways contribute only to a particular transition step. For instance, the cell cycle and ERBB pathways play key roles in the early-stage transition, while the ECM receptor and apoptosis pathways contribute to the progression from stage III to stage IV. The proposed computational pipeline is powerful in detecting important pathways and gene sets that drive cancers at certain stage(s). It offers new insights into the understanding of molecular mechanism of cancer initiation and progression.

1. Background

Cancer is a heterogeneous disease driven by abnormality of multiple genetic or epigenetic factors, such as gene expression level, DNA methylation level, somatic mutation and copy number variation. Over the past decades, many individual genes have been discovered to govern important functions in different cancer types, for instance, genes *BRCA1*, *BRCA2*, *PIK3C* and *GATA3* for breast cancer (The Cancer Genome Atlas Research and Network, 2008), genes *MYC*, *RIT1*, *ECFR* and *ERBB2* for prostate cancer (The Cancer Genome Atlas Research and Network, 2014). These individual genes were usually identified through univariate two-sample comparison between normal and tumor groups. Although the single-gene analyses are powerful in detecting important oncogenes and suppressors as potential treatment targets, they provided limited insights into the molecular mechanisms of tumorigenesis, as it omitted all regulatory relations between genes (Ivshina et al., 2006; Ma et al., 2003; Hoek et al., 2006; Talantov et al., 2005; Smith et al., 2005; Tomlins et al., 2007; Pancione et al., 2012; Subramanian et al., 2005). From the view of systems biology, a simultaneous analysis of a gene set or a genetic pathway might provide more clear functional insight into cause of the phenotypic changes. There have been many recently developed tools for gene set analysis, including EnrichNet (Glaab et al., 2012), GAGE (Luo et al., 2009), CSEA (Al-Shahrour et al., 2007), PAGE (Kim and Volsky, 2005), MEGO (Tu et al., 2005), Catmap (Breslin et al., 2004), ErmineJ (Lee et al., 2005), and GeneTrail (Backes et al., 2007). Most of these approaches, however, are knowledge-based

Abbreviations: TCGA, The Cancer Genome Atlas; MST, minimum spanning tree; CNV, copy number variation; BH, Benjamini-Hochberg; MDS, multidimensional scaling; BN, Bayesian network

* Corresponding authors.
E-mail addresses: qz008@uark.edu (Q. Zhang), hxchen@ucdavis.edu (H. Chen).

[1] All authors contributed equally to this work.

https://doi.org/10.1016/j.compbiolchem.2020.107285
Received 6 February 2020; Accepted 9 May 2020
Available online 26 May 2020
1476-9271/

and inflexible to user-defined pathways. One exception is Edelman et al. (2008), which introduced a data-driven approach to modeling cancer progression via pathway dependencies and flexible to any user-defined pathways.

Unlike the knowledge-based enrichment analysis, Edelman et al.'s method is essentially a hierarchical analysis that targets pathways relevant to particular transition between cancer stages, for instance, from normal to primary tumor, and from primary tumor to metastasis. However, Edelman's approach relied on several computationally intensive steps, such as regularized multi-task learning, inverse regression, learning gradients as well as leave-one-out cross validation, which greatly limited its application to large-scale data. To this end, we proposed an efficient and powerful test based on inter-point distance to identify cancer-driving pathways. Similar as Edelman's method, our method is purely data-driven and capable of dealing with new pathways adapted to certain tissue/disease. In the meanwhile, the test is model-free and very easy to implement as it only requires few simple calculations, such as minimum spanning tree and chi-square test.

The rest of the paper is structured as follows: In Section 2, we introduce the new test and establish its asymptotic distribution. The finite sample performance is evaluated via simulations. In Section 3, we apply the new test to the rich Cancer Genome Atlas data to identify important pathways that drive serous ovarian cancer. We discuss the strengths and shortcomings of our approach in Section 4 and conclude this paper in Section 5. Technical proof about the asymptotic distribution of the edge-count test is provided in the Appendix.

2. Methods

2.1. Problem formulation

Detecting cancer-driving pathways is essentially detecting differentially acted pathways between cancer stages. Here, we formulate it as a statistical problem of testing the equality of two or multiple joint distributions, where each random variable represents the expression level of one gene. To be precise, we let $i \in \{1, 2, ..., p\}$ be the index for cancer stages, and $(X_1^{(i)}, ..., X_d^{(i)})$ be the expression levels of d genes in the pathway being studied, with a joint distribution $\boldsymbol{F}^{(i)}$. Given n_i i.i.d. observations in stage i, $(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_{n_i}^{(i)})$, where $\mathbf{x}_k^{(i)} = (x_{k1}^{(i)}, x_{k2}^{(i)}, ..., x_{kd}^{(i)})$, we concern the following hypothesis testing:

$$H_0: \ \boldsymbol{F}^{(1)} = \boldsymbol{F}^{(2)} = \cdots = \boldsymbol{F}^{(p)},$$
$$H_a: \ \boldsymbol{F}^{(j)} \neq \boldsymbol{F}^{(j')}, \ \text{forsome} \ \ j, j', 1 \leq j, j' \leq p.$$

For a particular transition step, from stage i to stage $i+1$ $(i = 1, ..., p-1)$, we could conduct the following pairwise test, which is a special case when $p = 2$:

$$H_0: \ \boldsymbol{F}^{(i)} = \boldsymbol{F}^{(i+1)},$$
$$H_a: \ \boldsymbol{F}^{(i)} \neq \boldsymbol{F}^{(i+1)}.$$

2.2. Graph-based multivariate test

The two-sample multivariate tests in the statistics literature can be roughly classified into two categories, namely the parametric and nonparametric multivariate tests. Hotelling's T^2 test is a simple and widely used test for comparing the mean vectors of two multivariate Gaussian distributed populations, which generally works well for low-dimension case. There are also numerous tests recently developed for high dimensions (Srivastava et al., 2013; Cai et al., 2014; Gregory et al., 2015), which have mostly focused on the equality of mean vectors. Unlike the parametric tests, the nonparametric tests directly test the equality of two multivariate distributions, for instance, Kolmogorov-Smirnov (KS) test (Lopes et al., 2008) and edge-count test (Friedman and Rafsky, 1979; Rosenbaum, 2005), which are two popular families of such test, but both have practical limitations in real world applications. The KS test for two-sample comparison quantifies a distance between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution. KS test is known to be very conservative, i.e., the null hypothesis is too often not rejected. Moreover, by the brute force algorithm, the implementation of multi-dimensional KS test can be prohibitively computationally intensive. Edge-count tests are easy to implement but they could be problematic under certain location and scale alternatives, as pointed out by Chen and Friedman (2017). The rationale of the edge count test is that if two groups have different distributions, samples would be preferentially closer to others from the same group than those from the other group in a similarity graph, therefore the edges in the graph would be more likely to connect samples from the same group. The test rejects the null if the number of between-group edges is significantly less than expected. Chen and Friedman (2017) developed a modified edge-count test and established its asymptotic distribution under two samples. This test works properly under different alternatives and exhibits substantial power gains over existing edge-count tests. Similar as other edge-count tests, their test is based upon a similarity graph constructed over the pooled samples from different groups. For instance, one could use a minimum spanning tree (MST, Cheriton and Tarjan, 2006) based on euclidean distance as the similarity graph.

In order to test the hypotheses in Section 2.1, we extended Chen and Friedman's test to a multi-sample case and derived the asymptotic result for p-value approximation. We began with pooling samples from all p groups and indexing them by $1, 2, ..., N = \sum_{i=1}^{p} n_i$. A similarity graph, denoted by G, was then constructed on the pooled observations. Let R_i be the number of edges in the graph that connect observations within sample i. We worked under the permutation null distribution, which places $1/\binom{N}{n_1, n_2, ..., n_p}$ probability on each of the $\binom{N}{n_1, n_2, ..., n_p}$ choices of n_i out of the total N observations for group i, $i = 1, ..., p$, with each observation being chosen once. When there is no further specification, we denote by P_P, E_P, Var_P, Cov_P probability, expectation, variance, and covariance, respectively, under the permutation null distribution.

It is not hard to show that:

$$\mathsf{E}_\mathsf{P}(R_i) = |G|\frac{n_i(n_i-1)}{N(N-1)} \triangleq \mu_i, \tag{1}$$

$$\mathsf{Var}_\mathsf{P}(R_i) = \mu_i(1-\mu_i) + 2C\frac{n_i(n_i-1)(n_i-2)}{N(N-1)(N-2)} + (|G|(|G|-1) - 2C)\frac{n_i(n_i-1)(n_i-2)(n_i-3)}{N(N-1)(N-2)(N-3)} \triangleq \sigma_i^2, \tag{2}$$

$$\mathsf{Cov}_\mathsf{P}(R_i, R_j) = (|G|(|G|-1) - 2C)\frac{n_i n_j (n_i-1)(n_j-1)}{N(N-1)(N-2)(N_3)} - \mu_i\mu_j, \quad \text{for} i \neq j, \tag{3}$$

where $|G|$ is the number of edges in graph G, and $C = \frac{1}{2}\sum_{k=1}^{N} |G_k|^2 - |G|$ with G_k being the subgraph in G that includes all edge(s) that connect to node k, so C is the number of edge pairs that share a common node in G. We consider the following test statistic

$$S: = (R_1 - \mu_1, R_2 - \mu_2, ..., R_p - \mu_p)\Sigma^{-1}\begin{pmatrix} R_1 - \mu_1 \\ R_2 - \mu_2 \\ \cdots \\ R_p - \mu_p \end{pmatrix}, \tag{4}$$

Theorem 1. *For an edge $e \in G$, let*

$$A_e = \{e\} \cup \{e' \in G: e' \text{and } e \text{ share a node}\},$$
$$B_e = A_e \cup \{e'' \in G: \exists \ \ e' \in A_e, \text{ suchthat} e'' \text{and } e' \text{share a node}\}.$$

$$A_e = \{e\} \cup \{e' \in G: e' \text{and } e \text{ share a node}\},$$
$$B_e = A_e \cup \{e'' \in G: \exists \ \ e' \in A_e, \text{ suchthat} e'' \text{and } e' \text{share a node}\}.$$

If $|G| = O(N)$, $\sum_{k=1}^{N} |G_k|^2 - 4|G|^2/N = O(N)$, $\sum_{e\in G} |A_e||B_e|$, *then* $S \to \chi^2_p$ $= o(N^{1.5})$, $\lim_{N\to\infty} n_i/N = \lambda_i \in (0, 1)$ *under the permutation null.*

2.3. Simulation study I: Accuracy of p-value approximation

We conducted a simulation study under $p = 4$ to evaluate the finite performance of the p-value approximation under moderate sample sizes. In particular, we compared the approximate p-value with a permutation p-value from 10,000 permutations, under different dimensions ($d = 10, 100$) and sample sizes ($N = 80$, $n_1 = n_2 = n_3 = n_4 = 20$, $N = 140$, $n_1 = n_2 = 20$, $n_3 = n_4 = 50$ and $N = 200$, $n_1 = n_2 = n_3 = n_4 = 50$). The data were generated from p-dimension Gaussian distribution with zero mean vector and identical covariance matrix. A k-MST ($k = 1, 3, 5$) was then constructed over the pooled observations, based on which the test statistics S could be calculated. The k-MST is defined as the union of the first k disjoint minimum spanning trees and it can be obtained by several fast algorithms such as Chazelle's algorithm (Chazelle, 2000). The approximate p-value can be obtained as follows:

$$p-\text{value} = \Pr(\chi^2_{\text{df}=4} > S),$$

which is to be compared with permutation p-value from 10,000 permutations.

Figs. 1 and 2 summarized the accuracy of the approximate p-values (approximate p-value minus permutation p-value) under different dimensions, sample sizes, and similarity graphs. It can be seen that under all conditions, the approximate p-values tend to be slightly conservative and increasing dimension leads to a slightly decreasing accuracy. In addition, using a denser similarity graph, e.g., 3-MST or 5-MST can slightly improve the p-value approximation. A sample size such that $\min_i n_i > 20$ seems to be enough in practice, in order to approximate p-value with chi-square distribution.

2.4. Simulation study II: Power comparison with other tests

Our second simulation study compared the empirical statistical power of the proposed multi-sample edge-count test with three similar tests studied in Chen and Friedman (2017):

- $T_1 = \sum_{i=1}^{4} |R_i - \mu_i|$
- $T_2 = \sum_{i=1}^{4} |R_i - \mu_i|/\sqrt{\Sigma_{\text{ii}}}$
- $T_3 = \sum_{i=1}^{4} (R_i - \mu_i)^2$,

where R_i and μ_i represent the number of edges connecting samples within group i and its expectation. We consider four phenotypes and a simulated pathway with 100 nodes. Assuming independent nodes, we generated the data from a multivariate Gaussian distribution of dimension 100 for each group, where the mean vector of group i is $\boldsymbol{c}_i$ and the covariance matrix is an identity matrix. We set $\boldsymbol{c}_1 = 0$, $\boldsymbol{c}_2 = 0.1$, $\boldsymbol{c}_3 = 0.2$, $\boldsymbol{c}_4 = -0.3$, and varied the sample size from 100 to 500. Fig. 3 showed the comparison of the four tests in terms of the empirical statistical power, where it can be seen that our proposed test outperforms the other three test statistics.

3. Results

In this section, we applied the multi-sample edge-count test to the rich TCGA data (The Cancer Genome Atlas Research and Network, 2008, 2014) and studied the roles of 186 KEGG pathways (http://www.genome.jp/keg) in the progression of serous ovarian cancer. In TCGA, each subject is represented by multiple molecular data types including gene expression, exon expression, genotype (SNP), MicroRNA expression, copy number variation (CNV), somatic mutation, DNA methylation, as well as a complete clinical record including tumor stage, survival, age, race, outcomes of debulking surgery and chemotherapy. To take advantage of this comprehensive data, we considered three important data types in our analysis including gene expression level, CNV and DNA methylation. Other data types could also be incorporated, without causing too much calculation.

3.1. Data preprocessing

The TCGA ovarian cancer data contains information of 17,813 genes on 565 subjects. Based on the clinical classification, four groups (stages I, II, III, IV) contain 16, 27, 438 and 84 subjects, respectively. The transcriptome profiling data, CNV data as well as the methylation data were downloaded through Genomic Data Commons (GDC) portal in January 2017. Out of 17,813 genes, 12,831 had methylation level measured for each CpG island located in their promoter regions. For genes containing more than one CpG islands, we took the average as the methylation level as suggested by The Cancer Genome Atlas Research and Network (2008). The copy number was measured on each chromosome segment by circular binary segmentation (CBS). A gene was assigned the "seg.mean" value of the segment that it falls in. If a gene spans two chromosomal segments, we took the average of the "seg.mean" value as its overall copy number. We normalized each data type for each gene by subtracting the median and dividing by the standard deviation to avoid possible dominance by any of these three data types.

The expression level of each gene was quantified by the count of reads mapped to the gene. The quantifications were done by software *HTSeq* of version 0.9.1 (Anders et al., 2015) and the count data were log-transformed for further processing. In addition, we removed the effects due to different age groups and batches using a median-matching and variance-matching strategy (Hsu et al., 2012; Zhang et al., 2014). For example, the batch effect can be removed in the following way:

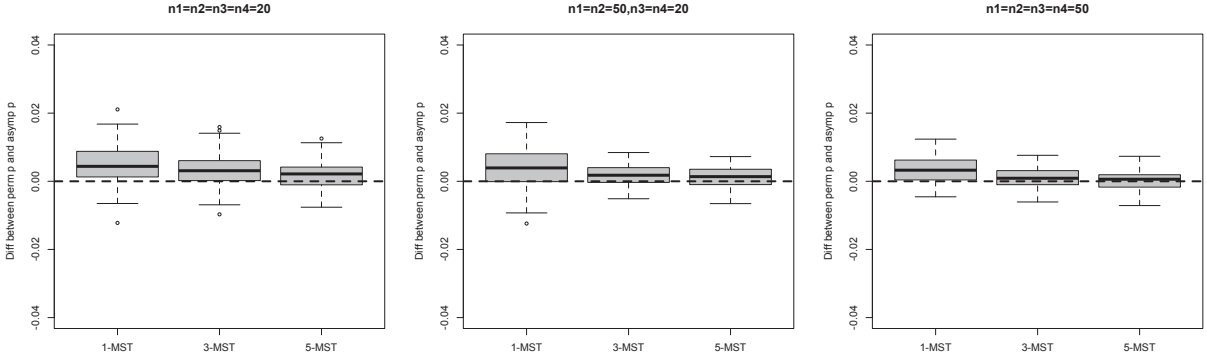


Fig. 1. Boxplots of the difference between approximate p-value and permutation p-value (approximate p minus permutation p), under different sample sizes and similarity graph construction (1-MST, 3-MST and 5-MST). Dimension d is 10.

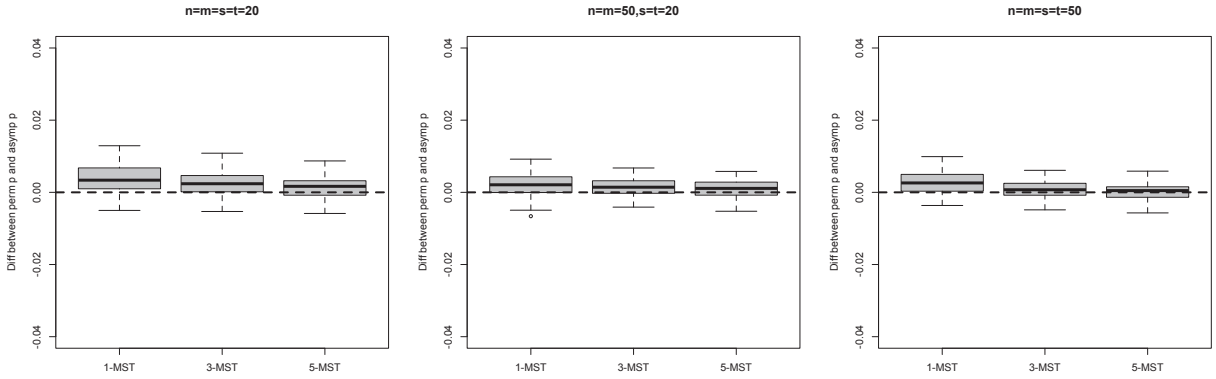


Fig. 2. Boxplots of the difference between approximate p-value and permutation p-value (approximate p minus permutation p), under different sample sizes and similarity graph construction (1-MST, 3-MST and 5-MST). Dimension d is 100.

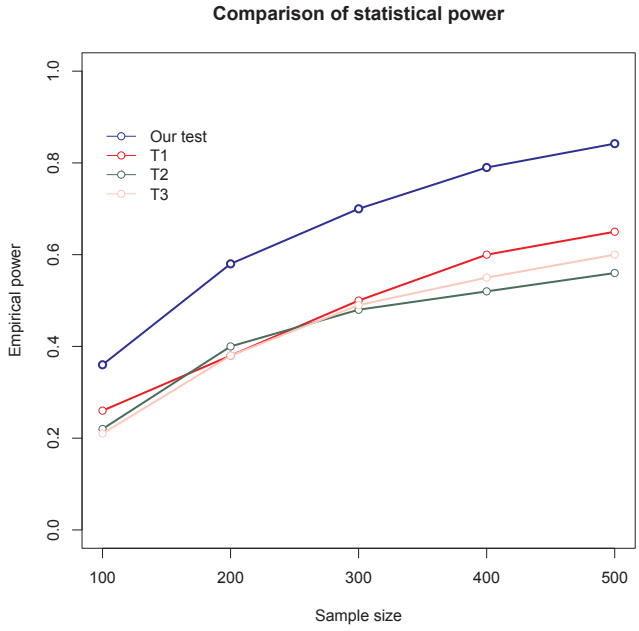


Fig. 3. Comparison of empirical statistical power under different sample sizes and similarity graph 3-MST. Dimension d is 100.

$$g^*_{ijk} = M_i + (g_{ijk} - M_{ij})\frac{\hat{\sigma}_{g_i}}{\hat{\sigma}_{g_{ij}}},$$

where g_{ijk} refers to the expression value for gene i from sample k in batch j ($j = 1, 2, ..., J$; $k = 1, 2, ..., n_j$), M_{ij} represents the median of $g_{ij} = (g_{ij1}, ..., g_{ijn})$, M_i refers to the median of $g_i = (g_{i1}, ..., g_{iJ})$, $\hat{\sigma}_{g_i}$ and $\hat{\sigma}_{g_{ij}}$ stand for the standard deviations of g_i and g_{ij}, respectively. Same strategy was used for DNA methylation data and CNV data.

3.2. Pathway expansion and refinement

Each of the 186 pathways was expanded by integrating all the three data types. For instance, if gene i was in the pathway being studied, then its methylation level and CNV level were added to the pathways as two new nodes. As it is well known that the expression level of a gene could be greatly affected by genetic or epigenetic changes such as copy number variation and DNA methylation, the inclusion of these two factors may provide insights about the upstream cause of abnormal expression.

To adapt the KEGG pathways to our context, we further refined the pathways by removing genes that are irrelevant to phenotypic changes. Similar as in Edelman's et al. (2008), an F-test was used to calculate the p-value of each single gene. A gene was excluded if its p-value exceeds some predefined threshold. In this analysis, we set the threshold to be 0.1, and the same procedure was applied to the methylation and CNV data. By the refinement, the sizes of 186 KEGG pathways were greatly reduced, for instance, the ERBB pathway was reduced from 87 genes to 29 genes, and 165 out of 174 methylation/CNV nodes were excluded.

3.3. Cancer-driving pathways

For each of the 186 pathways, we first conducted the multi-sample test to obtain the p-value for the whole process of development. By a Benjamini-Hochberg (BH) procedure with level 0.05, a set of nine pathways were identified, including some well-studied cancer-related pathways such as cell cycle, ERBB, p53 signaling, and JAK-STAT signaling pathways. For each of the identified pathways, a two-sample test was further conducted in order to investigate the roles of these pathways in each particular transition step (I→II, II→III, III→IV). Table 1 lists the nine cancer-driving pathways with p-values for the entire process and each particular transition. Interestingly, we found that most of these pathways contributed only to a particular step (the ERBB pathway is an exception, which has significant p-values in both transition I→II and transition II→III). For instance, five pathways were found to contribute only to the early-stage transition, three pathways contribute only to the high-grade transition. Six pathways, including ERBB, cell cycle, prostate cancer, TGF β signaling, pancreatic cancer and p53 signaling pathways, were found to significantly contribute to the transition I→II, which confirmed some existing studies. For instance, the ERBB pathway contains important proto-oncogenes and tumor suppressors such as *PIK3C*, *KRAS* and *STAT5*. It is known that the ERBB pathway is closely related to the development of a wide variety of cancers. Especially, the excessive signaling of growth factor receptors *ERBB1* and *ERBB2* are critical factors in the malignancy of solid tumor, and several studies have reported the critical role of ERBB in the early

Table 1
Pathways that drive ovarian cancer progression.

Pathway	p (overall)	p (I→II)	p (II→III)	p (III→IV)
1. ERBB	4.4×10^{-7}	7.2×10^{-5}	3.8×10^{-2}	2.0×10^{-1}
2. Cell cycle	9.6×10^{-7}	4.0×10^{-6}	5.3×10^{-6}	9.4×10^{-2}
3. Prostate cancer	4.1×10^{-6}	2.5×10^{-4}	1.1×10^{-1}	3.0×10^{-1}
4. ECM receptor	1.0×10^{-5}	1.7×10^{-1}	7.7×10^{-1}	6.2×10^{-5}
5. TGF β signaling	3.9×10^{-4}	2.7×10^{-3}	6.1×10^{-1}	7.9×10^{-1}
6. Apoptosis	6.2×10^{-4}	1.8×10^{-1}	8.5×10^{-1}	1.9×10^{-3}
7. Pancreatic cancer	6.9×10^{-4}	3.3×10^{-3}	7.2×10^{-2}	6.0×10^{-1}
8. P53 signaling	7.5×10^{-4}	9.1×10^{-4}	4.0×10^{-1}	5.5×10^{-4}
9. JAK-STAT signaling	1.4×10^{-3}	4.9×10^{-1}	8.2×10^{-1}	1.6×10^{-2}

Presented in the table is the list of pathways identified by the new test. The columns represent pathway name, overall p-value and p-value for each particular transition step, i.e., I→II, II→III, III→IV.

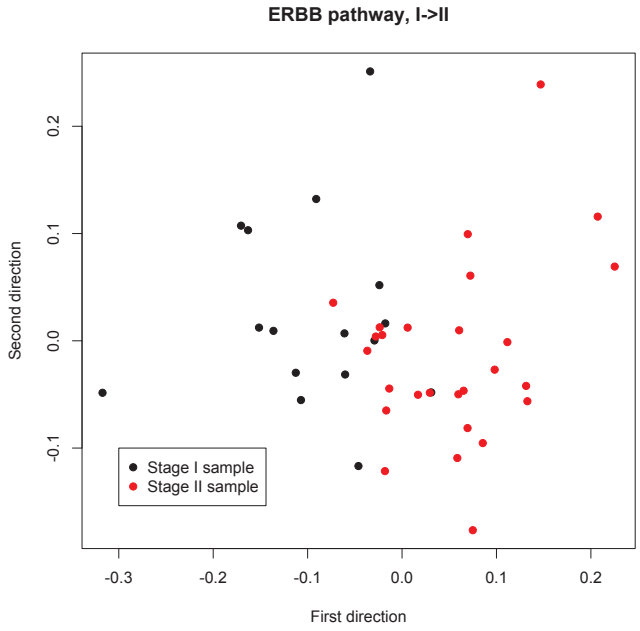


Fig. 4. Multidimensional scaling (MDS) plot for subjects at stage I and stage II, based on the refined gene set of ERBB pathway.

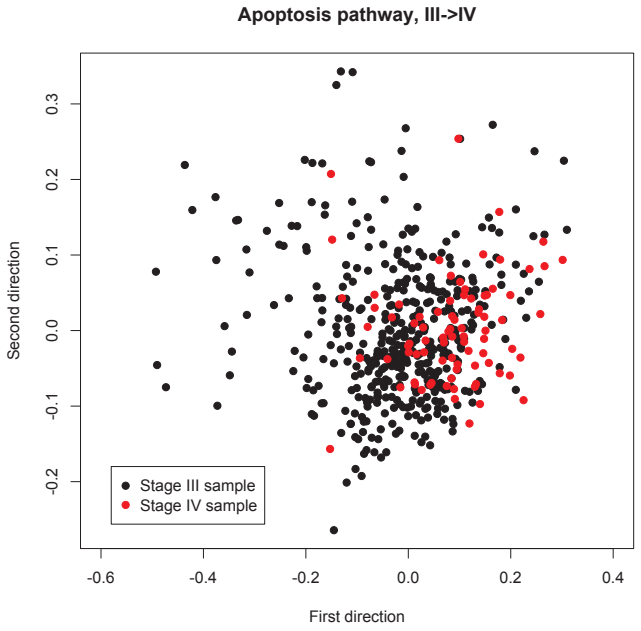


Fig. 6. Multidimensional scaling (MDS) plot for subjects at stage III and stage IV, based on the refined gene set of apoptosis pathway.

progression of ovarian cancer and breast cancer (The Cancer Genome Atlas Research and Network, 2008; Appert-Collin et al., 2015; Eccles, 2001; Stern, 2000). The cell cycle pathway contains many genes that co-regulate cell proliferation, including *ATM*, *RB1*, *CCNE1* and *MYC*. Abnormal regulation among these genes may cause the over proliferation of cells and an accumulation of tumor cell numbers (The Cancer Genome Atlas Research and Network, 2008). One novel finding from our analysis is the critical roles of ECM receptor pathway in the late-stage transition (III→IV). The extracellular matrix (ECM) is a major component of the local microenvironment in a cancer cell, which play important roles in cancer development (Lu et al., 2012).

Figs. 4 and 5 show the multidimensional scaling (MDS) plots (with Manhattan distance) using stage I and stage II samples, based on the

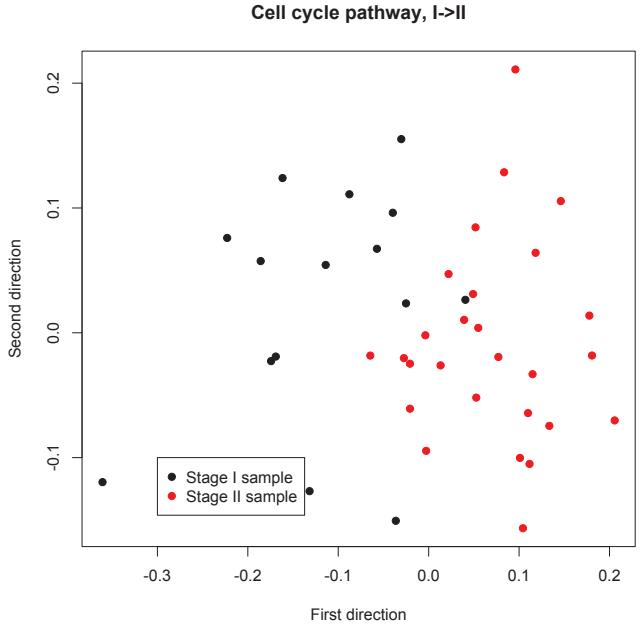


Fig. 5. Multidimensional scaling (MDS) plot for subjects at stage I and stage II, based on the refined gene set of cell cycle pathway.

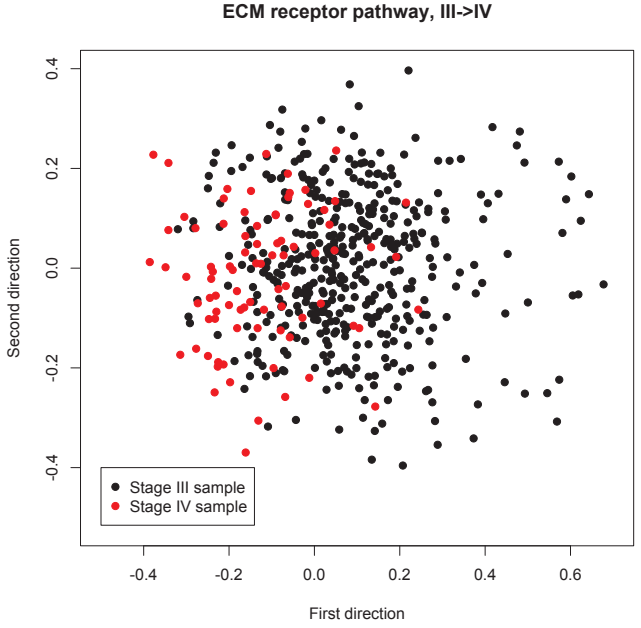


Fig. 7. Multidimensional scaling (MDS) plot for subjects at stage III and stage IV, based on the refined gene set of ECM receptor pathway.

refined gene sets of ERBB pathway (p-value = 7.2×10^{-5}) and cell cycle pathway (p-value = 4.0×10^{-6}), respectively. It can be seen that both pathways well differentiate the two groups, indicating their substantial involvement in this particular step of transition. Figs. 6 and 7 display the MDS plots for stage III and stage IV samples, based on the refined gene sets of ECM receptor pathway (p-value = 6.2×10^{-5}) and apoptosis pathway (p-value = 1.9×10^{-3}), respectively. As compared to III→IV transition, the transition from stage I to stage II is more substantial, which partially confirm some existing studies and clinical statistics (The Cancer Genome Atlas Research and Network, 2011; Lheureux et al., 2019; Chien and Poole, 2017). For instance, the 5-year survival rate for stage-I ovarian cancer patients is above 90%, while the rate for stage-II cancer patients drop below 30% (The Cancer Genome

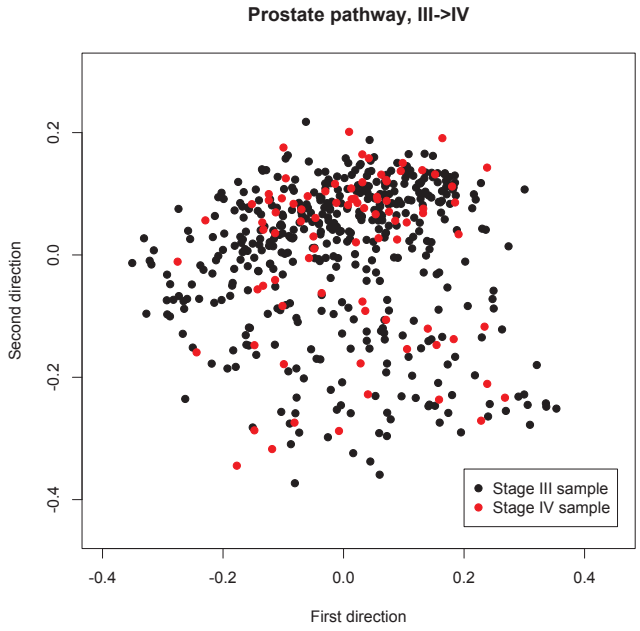


Fig. 8. Multidimensional scaling (MDS) plot for subjects at stage III and stage IV, based on the refined gene set of prostate cancer pathway.

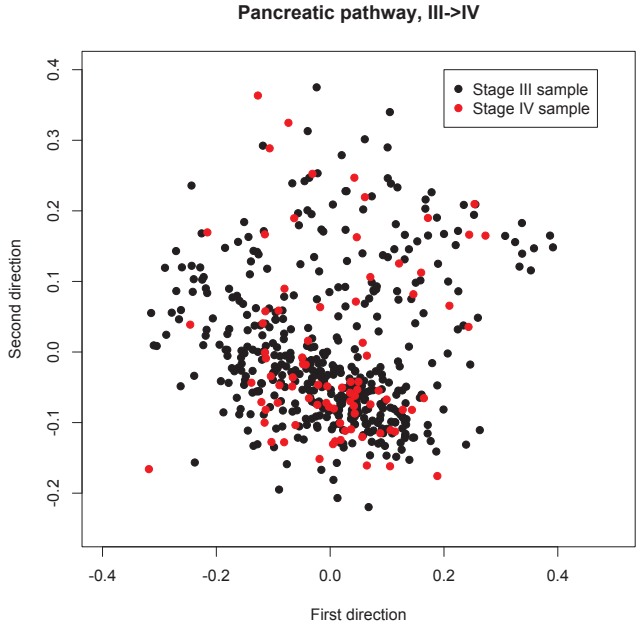


Fig. 9. Multidimensional scaling (MDS) plot for subjects at stage III and stage IV, based on the refined gene set of pancreatic cancer pathway.

Atlas Research and Network, 2011), suggesting an intrinsic genetic/epigenetic difference between these two groups. As a contrast, we also presented two MDS plots with insignificant p-values in Fig. 8 (p-value = 0.3, prostate cancer pathway) and Fig. 9 (p-value = 0.6, pancreatic cancer pathway).

After identifying these pathways, we further adopted a logistic Bayesian network model (Zhang et al., 2014) to predict the regulatory relations of the genes. A coordinate descent algorithm was employed to solve the penalized likelihood function and the penalizing constant λ was tuned by a likelihood method suggested by Zhang et al. (2014). Figs. 10 and 11 show the predicted Bayesian networks for the cell cycle pathway and ERBB pathway. It is noteworthy that in this analysis, each gene may have up to three measures including expression level, methylation level and CNV, and we treated them as three separate variables indicating by different colors. For instance, in Fig. 9, there is an edge from SFN (green) to SFN (yellow), which indicates that the methylation level of gene SFN directly affects its expression level. These directed networks may provide more clues about how the cancer-driving genes regulate each other, and together change the phenotype. For instance, in the cell cycle pathway (Fig. 9), we found gene *CDC20* is a driver gene (high outdegree and high centrality), regulating eight other cancer-driving genes in the network. Intervention on this gene may result in a global change of this network.

4. Discussion

Gene set analyses hold great promise in elucidating the molecular basis of complex diseases such as cancers. Nevertheless, most of the existing gene set analyses are knowledge-based and relied on the enrichment analysis of the pathways defined in existing databases. The main advantage of the proposed method over most existing analyses is the feasibility to user-defined gene sets, because the existing methods are mostly enrichment based, where the enrichment score assigned to a pathway depends on existing knowledge database, e.g., the biological function of each gene/protein and the interactions between genes/proteins reported in the literature. It is infeasible to incorporate new gene sets, for instance, the gene sets with different data types such as DNA methylation, CNV or somatic mutation. The KEGG pathways that we used in this paper can be viewed as newly defined gene sets after pathway refinement and expansion.

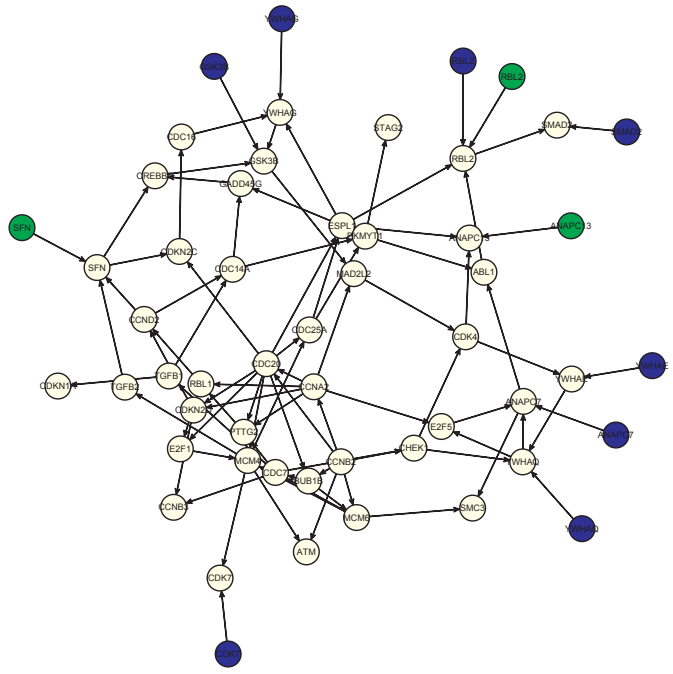
The new test is very easy to implement in practice as it only requires the calculation of a similarity graph over the observations. In practice, one can apply our test to large data set as an efficient searching tool for phenotype-changing pathways, as the computation is simple and efficient. Moreover, under the permutation null and some mild conditions, the test statistics converges to a chi-square distribution and sample size in tens for each group is generally good enough for p-value approximation, as we show in the simulation study.

The new test also has some limitations. First, the new test only targets for the difference of multiple joint distributions, but cannot detect the direction of changes. A follow-up test or a graphical visualization will be needed in order to tell the direction of changes, i.e., the up-regulation or down-regulation. For example, one could calculate an enrichment score for the pathway being studied as in existing gene set analyses or use a heatmap of the gene set.

Second, as we illustrated in the TCGA application, the performance of the new method relies on a refinement of the gene set, as the existence of irrelevant genes may greatly affect the identification of genes that differentiate phenotypes. In a case that the majority of a gene set are uncorrelated with phenotypic change, the test will fail to identify that gene set. One possible way to solve this problem is redefining distance metric, for example, one can consider to use some weighted distance measures or the ℓ_∞-norm measures that are more sensitive to a change in single gene or a small proportion of genes.

5. Conclusions

In this work, we introduced a simple but powerful multivariate test to identify (epi)genetic pathways that drive the cancer progression. Different from most existing gene set analyses, our test is data-driven and nonparametric, with allowance of user-defined gene sets or pathways. As we see in the method section, this graph-based test can be easily implemented for high dimensional problems. Besides the cancer progression application presented in this paper, the new test can be also applied to many other problems. For instance, one can test for pathways that differentiate cancer subtypes.



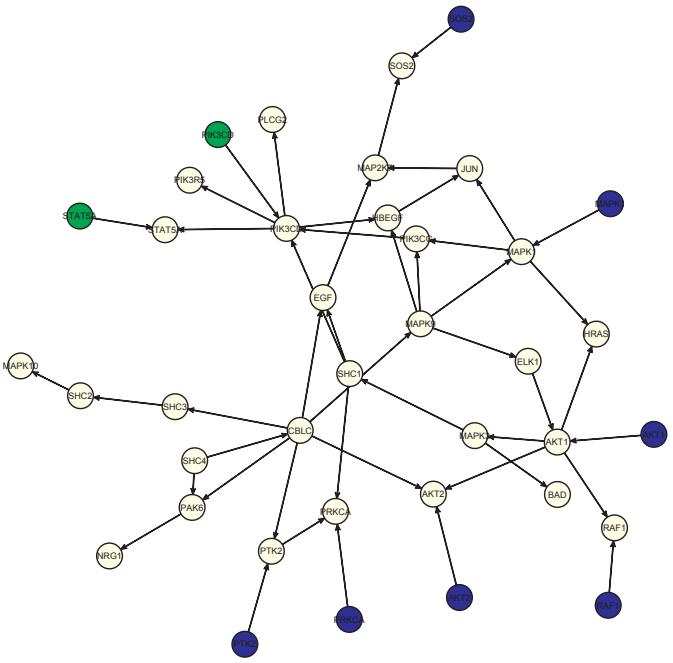


Fig. 11. Causal network for the refined ERBB pathway, predicted by the logistic Bayesian network model introduced in Zhang et al. (2014), where each yellow node represents the expression level of a gene, each green node represents the overall methylation level of a gene, and each blue node represents the copy number of a gene. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Competing interests

The authors have declared that no competing interests exist.

Fig. 10. Causal network for the refined cell cycle pathway, predicted by the logistic Bayesian network model introduced in Zhang et al. (2014), where each yellow node represents the expression level of a gene, each green node represents the overall methylation level of a gene, and each blue node represents the copy number of a gene. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Acknowledgements

QZ's research is supported in part by the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000. HC's research is supported in part by NSF award DMS-1513653.

Appendix: Detailed proof for asymptotic distribution of the test statistics

Proof. To prove the theorem, we take one step back to study the statistic under the bootstrap null distribution, which is defined as follows: For each observation, we assign it to be from sample i with probability n_i/N independently of other observations. Let V_i be the number of observations that are

assigned to be from sample i. Then, conditioning on $\{V_i = n_i\}_{i=1,\ldots,p}$, the bootstrap null distribution becomes the permutation null distribution. We use P_B, E_B, Var_B to denote the probability, expectation, and variance under the bootstrap null distribution, respectively.

We have

$$\mathsf{E}_\mathsf{B}(R_i) = \frac{n_i^2}{N^2}|G| \triangleq \mu_i^\mathsf{B},$$

$$\mathsf{Var}_\mathsf{B}(R_i) = \frac{n_i^2(N-n_i)^2}{N^4}|G| + \frac{n_i^3(N-n_i)}{N^4}\sum_{k=1}^{N}|G_k|^2 \triangleq (\sigma_i^\mathsf{B})^2.$$

Let

$$W_i^\mathsf{B} = \frac{R_i - \mu_i^\mathsf{B}}{\sigma_i^\mathsf{B}}, \quad W_i = \frac{R_i - \mu_i}{\sigma_i},$$

$$U_i = \frac{V_i - n_i}{\sqrt{N\lambda_{i,N}(1-\lambda_{i,N})}},$$

where $\lambda_{i,N} = n_i/N$. Under the conditions in the theorem, as $N \to \infty$, we can prove the following results:

(1) Under the bootstrap null, $(W_1^\mathsf{B}, W_2^\mathsf{B}, \ldots, W_p^\mathsf{B}, U_1, \ldots, U_{p-1})$ becomes multivariate normal distributed and the covariance matrix of $(U_1, \ldots, U_{p-1})$ is of positive definite.
(2) $\frac{\sigma_i^\mathsf{B}}{\sigma_i} \to c_i$, $\quad \frac{\mu_i^\mathsf{B} - \mu_i}{\sigma_i^\mathsf{B}} \to 0$, where c_i's are constants.
(3) $\mathrm{rank}(\Sigma) = p$.

Given (1), the conditional distribution of $(W_1^\mathsf{B}, W_2^\mathsf{B}, \ldots, W_p^\mathsf{B})'$ given $(U_1, \ldots, U_{p-1})$ becomes a multivariate Gaussian distribution under the boostrap null distribution as $N \to \infty$. Since the permutation null distribution is equivalent to the bootstrap null distribution given $U_i = 0$, $i = 1, \ldots, p-1$, $(W_1^\mathsf{B}, W_2^\mathsf{B}, \ldots, W_p^\mathsf{B})'$ becomes a multivariate Gaussian distribution under the permutation null distribution as $N \to \infty$. Since

$$W_i = \frac{\sigma_i^\mathsf{B}}{\sigma_i}\left(W_i^\mathsf{B} + \frac{\mu_i^\mathsf{B} - \mu_i}{\sigma_i^\mathsf{B}}\right),$$

given (2), we have $(W_1, W_2, \ldots, W_p)'$ becomes a multivariate Gaussian distribution under the permutation null distribution as $N \to \infty$. Together with (3), we have the conclusion in the theorem.

Assumption 1. Chen and Shao, 2005, p. 17

For each $i \in \mathcal{J}$ there exists $K_i \subset L_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{K_i^c}$ and ξ_{K_i} is independent of $\xi_{L_i^c}$.

We will use the following theorem in proving Theorem 1.

Theorem 2. Chen and Shao, 2005, Theorem 3.4*Under Assumption* 1*, we have*

$$\sup_{h \in \mathrm{Lip}(1)} |\mathsf{E}h(V) - \mathsf{E}h(Z)| \le \delta,$$

where $\mathrm{Lip}(1) = \{h: \mathbb{R} \to \mathbb{R};\ \|h'\| \le 1\}$*,* Z *has* $\mathcal{N}(0, 1)$ *distribution and*

$$\delta = 2\sum_{i \in \mathcal{J}}\left(\mathsf{E}|\xi_i\eta_i\theta_i| + |\mathsf{E}(\xi_i\eta_i)|\mathsf{E}|\theta_i|\right) + \sum_{i \in \mathcal{J}}\mathsf{E}|\xi_i\eta_i^2|$$

with $\eta_i = \sum_{j \in K_i}\xi_j$ *and* $\theta_i = \sum_{j \in L_i}\xi_j$*, where* K_i *and* L_i *are defined in Assumption* 1.For $e \in G$, let

$$\xi_e = \sum_{i=1}^{p} a_i \frac{I_{J_e = i} - \lambda_{i,N}}{\sigma_i^\mathsf{B}},$$

where $\{J_e = i\}$ means the edge connects two observations from sample i.For $k \in \{1, \ldots, N\}$, let

$$\xi_k = \sum_{i=1}^{p-1} b_i \frac{I_{g_k = i} - \lambda_{i,N}}{\sqrt{N\lambda_{i,N}(1-\lambda_{i,N})}},$$

where $\{g_k = i\}$ means node k is from sample i.For $e = (e_-, e_+) \in G$, let

$$K_e = A_e \cup \{e_-, e_+\},$$
$$L_e = B_e \cup \{\mathrm{nodesin}A_e\}.$$

Then K_e and L_e satisfy Assumption 1.For $k \in \{1, \ldots, N\}$, let

$$K_k = \{e \in G_k\} \cup \{i\},$$
$$L_k = \{e \in G_{k,2}\} \cup \{\mathrm{nodesin}G_i\}.$$

Then K_k and L_k satisfy Assumption 1.For $j \in \mathcal{J}$, let $\eta_j = \sum_{k \in K_j}\xi_k$, $\theta_j = \sum_{k \in L_j}\xi_k$. By Theorem 2, we have $\sup_{h \in \mathrm{Lip}(1)} |\mathsf{E}h(W) - \mathsf{E}h(Z)| \le \delta$ for $Z \sim \mathcal{N}(0, 1)$, where

$$\delta = \frac{1}{\sqrt{\mathrm{Var}_{\mathsf{B}}(W)}}\left(2\sum_{j\in\mathcal{J}}(\mathsf{E}_{\mathsf{B}}|\xi_j\eta_j\theta_j| + |\mathsf{E}_{\mathsf{B}}(\xi_j\eta_j)|\mathsf{E}_{\mathsf{B}}|\theta_j|) + \sum_{j\in\mathcal{J}}\mathsf{E}_{\mathsf{B}}|\xi_j\eta_j^2|\right)$$
$$\le \frac{1}{\sqrt{\mathrm{Var}_{\mathsf{B}}(W)}}\left(5\sum_{e\in G}\frac{p^3a^3}{\sigma^3}(|A_e|+2)(|B_e|+|A_e|+1) + 5\sum_{k=1}^{N}(|G_i|+1)(|G_{i,2}|+1)\right)$$
$$\le \frac{1}{\sqrt{\mathrm{Var}_{\mathsf{B}}(W)}}\frac{90p^3a^3}{\sigma^3}\sum_{e\in G}|A_e||B_e|$$

Since $\sigma = O(\sqrt{N})$, when $|A_e||B_e| = o(N^{1.5})$, we have $\delta \to 0$ as $N \to \infty$.We next check the covariance matrix of $(U_1, ..., U_{p-1})$. The diagonal elements of this matrix are all 1's and the off-diagonal element (i, j), $i \neq j$ is

$$-\sqrt{\frac{\lambda_{i,N}\lambda_{j,N}}{(1-\lambda_{i,N})(1-\lambda_{j,N})}}.$$

We next show result (2). Notice that σ_i^2 can be re-written as

$$\sigma_i^2 = \frac{n_i(n_i-1)(N-n_i)(N-n_i-1)}{N(N-1)(N-2)(N-3)}\left(|G| + \frac{n_i-2}{N-n_i-1}\left(\sum_{k=1}^{N}|G_i|^2 - \frac{4|G|^2}{N}\right) - \frac{2}{N(N-1)}|G|^2\right).$$

Since $|G|$, $\sum_{k=1}^{N}|G_k|^2 - 4|G|^2/N = O(N)$, let $\lim_{N\to\infty}|G|/N = a_0$ and $\lim_{N\to\infty}(\sum_{k=1}^{N}|G_k|^2 - 4|G|^2/N)/N = b_0$, then

$\lim_{N\to\infty}\sigma_i^2/N = \lambda_i^2(1-\lambda_i)^2(a_0 + b_0\lambda_i/(1-\lambda_i))$,

$\lim_{N\to\infty}(\sigma_i^{\mathsf{B}})^2/N = \lambda_i^2(1-\lambda_i)^2a_0 + \lambda_i^3(1-\lambda_i)(b_0+4a_0^2)$.

Hence

$$\lim_{N\to\infty}\frac{\sigma_i^{\mathsf{B}}}{\sigma_i} = \sqrt{1+\frac{4a_0^2\lambda_i}{a_0(1-\lambda_i)+b_0\lambda_i}}.$$

Also, since $\mu_i^{\mathsf{B}} - \mu_i = |G|\frac{n_i(N-n_i)}{N^2(N-1)}$, so

$$\lim_{N\to\infty}\frac{\mu_i^{\mathsf{B}}-\mu_i}{\sigma_i^{\mathsf{B}}} = \lim_{N\to\infty}\frac{a_0\lambda_i(1-\lambda_i)}{\sigma_i^{\mathsf{B}}} = 0.$$

We next show (3). The diagonal elements of Σ are σ_i^2's. The off-diagonal elements are, for $i \neq j$,

$$\Sigma[i,j] = \frac{n_in_j(n_i-1)(n_j-1)}{N(N-1)(N-2)(N-3)}\left(|G| - \left(\sum_{k=1}^{N}|G_k|^2 - \frac{4|G|^2}{N}\right) - \frac{2}{N(N-1)}|G|^2\right).$$

References

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Mingues, P., et al., 2007. From genes to functional classes in the study of biological systems. BMC Bioinform. 8, 114.

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq – a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169.

Appert-Collin, A., Hubert, P., Cremel, G., Bennasroune, A., 2015. Role of ErbB receptors in cancer cell migration and invasion. Front. Pharmacol. 6, 283–290.

Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., et al., 2007. GeneTrail-advanced gene set enrichment analysis. Nucleic Acid Res. 35, W186–W192.

Breslin, T., Edén, P., Krogh, M., 2004. Comparing functional annotation analyses with Catmap. BMC Bioinform. 5, 193.

Cai, T., Liu, W., Xia, Y., 2014. Two-sample test of high dimensional means under dependence. J. R. Stat. Soc. Ser. B 76 (2), 349–372.

Chazelle, B., 2000. A minimum spanning tree algorithm with inverse-Ackermann type complexity. J. ACM 47 (6), 1028–1047.

Chen, H.Y., Shao, Q., 2005. Stein's Method for Normal Approximation. Lecture Note Series, IMS. NUS.

Chen, H., Friedman, J.H., 2017. A new graph-based two-sample test for multivariate and object data. J. Am. Stat. Assoc. 112, 397–409.

Cheriton, D., Tarjan, R., 2006. Finding minimum spanning trees. SIAM J. Comput. 5 (4), 724–742.

Chien, J., Poole, E.M., 2017. Ovarian cancer prevention, screening and early detection: report from the 11th Biennial Ovarian Cancer Research Symposium. Int. J. Gynecol. Cancers 27 (9), 20–22.

Eccles, S.A., 2001. The role of c-erbB-2/HER2/neu in breast cancer progression and metastasis. J. Mammary Gland Biol. Neoplasia 6 (4), 393–406.

Edelman, E., Guinney, J., Chi, J.-T., Febbo, P.G., Mukherjee, S., 2008. Modeling cancer progression via pathway dependencies. PLoS Comput. Biol. 4 (2), e28.

Friedman, J.H., Rafsky, L.C., 1979. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann. Stat. 7 (4), 697–717.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., Valencia, A., 2012. EnrichNet: network-based gene set enrichment analysis. Bioinformatics 28 (18), i451–i457.

Gregory, K., Carroll, R.J., Baladandayuthapani, V., Lahiri, S.N., 2015. A two-sample test for equality of means in high dimension. J. Am. Stat. Assoc. 110 (510), 837–849.

Hoek, K.S., Schlegel, N.C., Brafford, P., Sucker, A., Ugurel, S., et al., 2006. Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. Pigment Cell Res. 19 (4), 290–302.

Hsu, F., Serpedin, E., Hsiao, T., Bishop, A., Dougherty, E., et al., 2012. Reducing confounding and suppression effects in TCGA data: an integrated analysis of chemotherapy response in ovarian cancer. BMC Genom. 13, S13.

Ivshina, A.V., George, J., Senko, O., Mow, B., Putti, T.C., et al., 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 66 (21), 10292–10301.

Kim, S., Volsky, D.J., 2005. PAGE: parametric analysis of gene set enrichment. BMC Bioinform. 6, 144.

Lee, H.K., Braynen, W., Kehsav, K., Pavlidis, P., 2005. ErmineJ: tool for functional analysis of gene expression data sets. BMC Bioinform. 6, 269.

Lheureux, S., Gourley, C., Vergote, I., Oza, A.M., 2019. Epithelia ovarian cancer. Lancet 10177, 1240–1253.

Lopes, R.H.C., Hobson, P.R., Reid, I.D., 2008. Computationally efficient algorithms for the two-dimensional Kolmogorov-Smirnov test. J. Phys.: Conf. Ser. 19 (4), 042019.

Lu, P., Weaver, V.M., Werb, Z., 2012. The extracellular matrix: a dynamic niche in cancer progression. J. Cell Biol. 196 (4), 395.

Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D., Woolf, P.J., 2009. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinform. 10, 161.

Ma, X., Salunga, R., Tuggle, J.T., Gaudet, J., Enright, E., et al., 2003. Gene expression profiles of human breast cancer progression. Proc. Natl. Acad. Sci. U. S. A. 100 (10), 5974–5979.

Pancione, M., Remo, A., Colantuoni, V., 2012. Genetic and epigenetic events generate multiple pathways in colorectal cancer progression. Pathol. Res. Int. 2012, 509348.

Rosenbaum, P.R., 2005. An exact distribution-free test comparing two multivariate distributions based on adjacency. J. R. Stat. Soc. Ser. B 67 (4), 515–530.

Smith, A.P., Hoek, K., Becker, D., 2005. Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanomas. Cancer Biol. Ther. 4 (9), 1018–1029.

Srivastava, M.S., Katayama, S., Kano, Y., 2013. A two sample test in high dimensional data. J. Multivar. Anal. 114, 349–358.

Stern, D.F., 2000. Tyrosine kinase signaling in breast cancer: ErbB family receptor tyrosine kinases. Breast Cancer Res. 2 (3), 176–183.
Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B.L., et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102 (43), 15545–15550.
Talantov, D., Mazumder, A., Yu, J., Briggs, T., Jiang, Y., et al., 2005. Novel genes associated with malignant melanoma but not benign melanocytic lesions. Clin. Cancer Res. 11 (20), 7234–7242.
The Cancer Genome Atlas Research and Network, 2008. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70.
The Cancer Genome Atlas Research and Network, 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550.
The Cancer Genome Atlas Research and Network, 2011. Integrated genomic analyses of ovarian carcinoma. Nature 474, 609–615.
Tomlins, S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., et al., 2007. Integrative molecular concept modeling of prostate cancer progression. Nat. Genet. 39 (1), 41–51.
Tu, K., Yu, H., Zhu, M., 2005. MEGO: gene functional module expression based on gene ontology. BioTechniques 38, 277–283.
Zhang, Q., Burdette, J.E., Wang, J.-P., 2014. Integrative network analysis of TCGA data for ovarian cancer. BMC Syst. Biol. 8 (1338), 1–18.