A two-variate phenotype-targeted test for detection of phenotypic biomarkers on breast cancer

Jin-Xiong Lv, Shikui Tu, Lei Xu
Department of Computer Science and Engineering,
Center for Cognitive Machines and Computational Health (CMaCH),
School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai, China
{lvjinxiong, tushikui, leixu}@sjtu.edu.cn

***Abstract*—Traditional pipeline for the task of detecting phenotypic biomarkers is a two-stage implementation, i.e., differentially expressed candidates are identified by NT tests, and then a subset of the candidates are further detected by phenotype-targeted tests (PT test) for significant phenotypic features, where N is short for Normal data and T is for Treatment/Trouble data. Such a two-stage procedure has low detection power as they do not make full use of the information contained in the (T, N). In this paper, we apply the two-variate PT test which jointly considers tumor-adjacent data and tumor data for improving the detection power. We investigate its performance by experiments on real-world datasets of breast cancer, considering phenotypes including BMI, overall survival time, pathologic stage, and tumor size. The results show that the method has high detection power and is more reliable, and the tumor-adjacent normal data plays an important role in the detection of phenotypic biomarkers. Finally, we obtain a new finding that the gene *TCTEX1D2* is significantly related to tumor size in breast cancer.**

***Index Terms*—two-variate phenotype-targeted test, phenotypic biomarkers, breast cancer**

I. Introduction

The detection of phenotypic biomarkers plays a crucial role in medical research for breast cancer which ranks first in the incidence rate and mortality rate for females among cancers [1]. There are many valued clinical features for cancers including pathologic stages, number of lymph nodes, tumor size, survival time and so on, and they can be regarded as phenotypes for patients. The comprehensive understanding of the relationship between these features and cancer biomarkers (functional molecules e.g., genes or ncRNAs) helps to reveal the roles for biomarkers in the progression of cancers [15], [24].

Candidate cancer biomarkers can be searched by differential expression NT test on cases and controls, where T indicates abnormal or trouble conditions such as Tumor or Treatment data, and N represents normal conditions or Tumor-adjacent data which comes from histologically normal tissue adjacent to the tumor (up to 1 cm from the margins of the tumor) [38]. In traditional pipelines, the phenotypic biomarkers are either identified directly by uni-variate phenotype-targeted test (PT test) [38], or further selected from a candidate set of cancer biomarkers in a two-stage implementation [21], [39]. For instance, stromal genes expression is utilized to predict clinical outcome in breast cancer as the stromal compartment plays an important role in cancer initiation and progression [6]. The above traditional pipelines have low detection power owing to two limitations. As described in the Fig.7d(2) [38], the first limitation shows that different phenotypes may not be differentiated if considering T data only, but actually can be well separated when both T and N are jointly considered. Therefore, it is essential to pay attention to the important role of Normal conditions played in the PT test. The second limitation, resulted from the two-stage implementation, is that the differential expression between T vs N may not well cope with the distinctions between phenotypes [38]. Therefore, the number of biomarkers detected by both the NT test and PT test is very few.

In this paper, both T and N are jointly considered into the two-variate phenotype-targeted test (two-variate PT test) as suggested in Table 4 of Ref. [38]. The conditions (T and N), phenotype and expression profiles form a data cubic. The two-variate PT test analyzes the data cubic directly and overcomes the two limitations discussed above. We evaluate the involved methods on the real-world datasets of breast cancer in considering different data types and different phenotypes. The RNA-seq gene expression profiles come from The Cancer Genome Atlas (TCGA) database while the Microarray gene expression profiles from the Gene Expression Omnibus (GEO) database, on which we apply two kinds of implementation of two-variate PT test. One implements the two-variate PT test by the Hotelling test and the other is based on Fisher discrimination analysis (FDA).

We take four phenotypes including BMI, overall survival (OS) time, pathologic stage, and tumor size into account for studies of breast cancer. Based on the literature searches, we collected three biomarkers for BMI. Two out of three are associated with BMI and the rest one is not. As for the OS time, we first annotate the top 10 protein-coding genes ordered by two-variate PT test and they are classified into three classes according to whether they are associated with survival of patients with breast cancer. Subsequently, we collected 20 biomarkers which are regarded as survival-related genes to show whether the two-variate PT test can detect them or not. As a result, 22 biomarkers are utilized to evaluate the detection power of related methods while the rest one is regarded as a false negative case on BMI and OS time. Moreover, the

978-1-7281-1867-3/19/$31.00 ©2019 IEEE
978-1-7281-1867-3/19/$31.00 ©2019 IEEE

proportion of genes in class A can also evaluate that. The FDA-based two-variate PT test can detect all of biomarkers which are detected by other methods, and we show that on the studies of OS time and pathologic stage. Finally, we detect one susceptibility gene for tumor size in breast cancer.

II. Methods

A. Motivations

The PT test utilizes the one-dimension (T or N) data of one biomarker to test whether it can differentiate phenotypes or not as illustrated in Fig.7b-d [38]. For example, the significant SNPs can be detected by χ^2 test in GWASs [14]. After the classification of clinical outcome (good survival or poor survival), we can identify survival-related genes by univariate t-test [11]. Both of them employ one condition and belong to the PT test. Two toy examples are given in Fig.1. The Fig.1(A) shows that the PT test works while phenotypes cannot be differentiated depending on tumor data in Fig.1(B). If the normal data is also involved, there is an obvious boundary between two phenotypes. Therefore, the tumor-adjacent normal data has a great effect on the performance of the PT test, which motivates us to jointly utilize both normal and tumor data for phenotype differentiation.

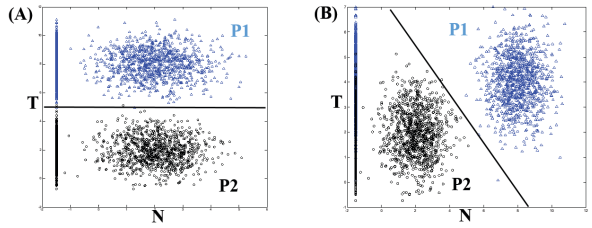

Fig. 1. Two toy examples

The traditional pipeline for the detection of phenotypic biomarkers consists of the NT test and PT test. The NT test provides significant differentiation expression biomarkers and then they are tested to confirm whether they can also differentiate phenotypes. From the perspective of the lattice taxonomy of tests for the multivariate test, the pipeline considers the strongest collegiality leading to lower detection power [38]. The significant differentiation phenotype biomarker set is a subset of a significant differentiation expression biomarker set. In Fig.2, most of the significantly differentiated expression biomarkers are not phenotypic biomarkers as the differentiation expression may not well cope with the distinctions between phenotypes. Altogether, the sufficient number of identified biomarkers cannot be guaranteed, which motivates us to integrate both the NT test and PT test rather than utilize them like a factory assembly line.

B. The two-variate phenotype-targeted test

In considering the two motivations, we integrate expression profiles, tumor and adjacent-tumor normal conditions and phenotypes into a data cubic, on which the recently proposed

Fig. 2. Graph description of the traditional pipeline

two-variate PT test is performed [37], [38] to detect phenotypic biomarkers. In Fig.1(B), two phenotypes can be separated by a line in the scatter plot. Only considering the tumor data, i.e., the projections on y axis in the Fig.1, the PT test can not distinguish two phenotypes, while the two-variate PT test offers opportunities to detect more biomarkers for distinguishing Phenotype 1 (PT1) and Phenotype 2 (PT2) by jointly considering the (T, N). In sequel, we perform two types of implementation, as suggested in Table 4 of Ref. [38].

The first implements the two-variate PT test by the Hotelling test, which generalizes the Student's t-test into the multivariate statistical method, and its corresponding statistics are defined as:

$$\begin{aligned} T^2 &= \frac{N_1 N_2}{N_1+N_2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \\ \mu_w &= \sum_i^{N_w} \mathbf{x}_i^w,\ w=1,2 \\ \Sigma &= \Sigma_1 = \Sigma_2, \end{aligned} \tag{1}$$

and

$$\frac{N_1+N_2-p-1}{(N_1+N_2-2)p}T^2 \sim F_{p,N_1+N_2-p-1}, \tag{2}$$

where the N_1 and N_2 are the number of case and control populations, the F follows F distribution with parameters p and N_1+N_2-p-1 and p is the dimension of $\mathbf{x}$.

The second is based on FDA, which aims to seek a linear projection $y=\theta'\mathbf{x}$ so that samples can be separated into two populations. We first assume the mean vector of two classes C_1 and C_2 for $\mathbf{x}$:

$$\mu_w = \frac{1}{N_w}\sum_{i\in C_w} x_i\ w=1,2 \tag{3}$$

After projection, we obtain:

$$\begin{aligned} \mu_w^y &= \frac{1}{N_w}\sum_{i\in C_w} y_i^w \\ y_i^w &= \theta' x_i^w \\ {\sigma_w^y}^2 &= \frac{\sum_i^{N_w}(y_i^w-\mu_w^y)}{N_w},\ w=1,2 \end{aligned} \tag{4}$$

And the objective function can be expressed as:

$$\begin{aligned} \max_\theta J(\theta) &= \frac{(\mu_1^y-\mu_2^y)^2}{\alpha_1{\sigma_1^y}^2+\alpha_2{\sigma_2^y}^2} \\ &= \frac{\theta'(\mu_1-\mu_2)(\mu_1-\mu_2)'\theta}{\alpha_1\theta'\sigma_1^2\theta+\alpha_2\theta'\sigma_2^2\theta}, \end{aligned} \tag{5}$$

where $\alpha_1=\frac{N_1}{N_1+N_2}$ and $\alpha_2=\frac{N_2}{N_1+N_2}$. After obtain θ, we use the FDA projection $\theta\mathbf{x}$ to replace the original expression value,

and conduct the two-sample t-test in which the null hypothesis is:

$$H_0 : \mu_1^y = \mu_2^y, \tag{6}$$

where μ_1^y and μ_1^y are mean of projection values for two phenotypes. When the y is one-dimensional, it shows that $T^2 = \frac{N_1 N_2}{N_1+N_2} J(\theta)$ and the FDA aims to seek a direction θ along which two populations differ mostly [37].

III. Description of datasets and preprocessing

The RNA-seq dataset comes from the TCGA database and was collected in GSE62944 from GEO database, while Microarray dataset was collected in GSE70951. The RNA-seq dataset measures the level of transcription by FPKM value of RNA-seq [27]. And it contains 113 samples in which the TCGA-A7-A0DC has no tumor data. After removing male samples, the sample size of that is 111. The Microarray data were background-corrected and quantile normalized using limma in R [32]. To reduce the error introduced by different platforms, we chose the cohort with 148 samples, and the corresponding platform is GPL13607. 20 samples do not have BMI and the sample size is 128. As for the study of tumor size, five samples have no tumor size data so that the sample size is 143.

The OS time is defined as the length of time from either the date of diagnosis or the start of treatment for a disease to death of one patient. Here, OS time is utilized to classify good and poor survival. We filtered out censored data which are not taken into account in the phenotype-targeted test including the simplest case and two-variate cases. After removing the censored samples, the sample size of breast cancer is 87. There is one missing value for the study of the pathologic stage and the sample size is 110. We define stage I and II as low pathologic stage and the rest belongs to the high one. As for the BMI, we obtain binary phenotypes by comparing with its mean value. The tumor size is often classified into two groups according to whether the diameter is greater than or equal to 2 cm [36].

IV. Results

A. Body Mass Index in breast cancer

The expression profiles of tumor-adjacent normal for both the macrophage scavenger receptor (*MSR1*) and the leptin (*LEP*) are associated with BMI except for the adipocyte-derived hormone adiponectin (*ADIPOQ*) [26]. In other words, the gene *MSR1* and *LEP* can be regarded as benchmarkers for evaluation of detection power of the two-variate PT test, while the gene *ADIPOQ* can be made use of evaluation of the reliability. Three probes of gene *MSR1* have no missing value while those probes of both gene *MSR1* and *LEP* have missing value. We filtered out samples with missing values rather than imputed them for reduction of errors introduced by imputation methods. After the transformation of continuous values of BMI into binary values compared with its mean value, four methods including PT test, NT test and two implementations of two-variate PT test were conducted for the three genes and related results were shown in Table I.

The gene *MSR1* can be detected by the PT test and two implementations of two-variate PT test. However, gene *LEP* can be detected by two implementations of two-variate PT while the PT test cannot. The gene *ADIPOQ* cannot be detected by two implementations of two-variate PT test, which is consistent with existing evidence. The Fig.3 shows the scatter plots of three genes. Both the gene *MSR1* and *LEP* can be classified into two classes by the black line and the tumor data makes contributions to the detection. As for gene *ADIPOQ*, there is no boundary to divide them into high-BMI group and low-BMI group. Therefore, tumor data does not result in false positive discoveries. Altogether, two-variate PT test is more powerful and reliable.

B. Survival outcome in breast cancer

The survival outcome can be classified into good survival and poor survival based on survival time. Here, we divided samples into the good-survival group or poor-survival group according to five-year overall survival time. If the survival time of one sample is greater than or equal to five years and it is not censored, we classify that into the good-survival group. We obtained 4047 significant genes by FDA-based two-variate PT test and 1260 by Hotelling test. The PT test can identify 1154 genes and 716 genes can be detected by both PT test and NT test ($\alpha = 0.05$). The number of biomarkers detected by four methods is described by the Venn plot Fig.4(A). The FDA-based two-variate PT test can detect all genes which were detected by the PT test or Hotelling test. Moreover, the Hotelling test detected more biomarkers than PT test, while the biomarker set detected by the PT test is not a subset of the biomarker set of Hotelling test. As a result, the FDA-based two-variate PT test achieves more detection power.

To evaluate the performance of two-variate PT test, we annotated top 10 protein genes ordered by FDA-based two-variate PT test and classified them into three classes. Class A contains genes which are associated with survival. Genes in class B are related to breast cancer but there is no obvious evidence for the existence of relations between them and survival while the rest genes make up the class C. The results were shown in Table II. There are four genes in the class A including gene *FXYD1*, *ZNF425* , *IRF2* and *ANG*. The highest conservation between *FXYD* family members lies in the region of the transmembrane domain of Dysadherin which plays a role in cancer progression [19]. The Zinc Finger Protein 425 (*ZNF425*) occupies significant copy number aberrations distinguishing primary breast tumors associated with and without lymph node metastases [4]. The Interferon Regulatory Factor 2 (*IRF2*) can lead to growth inhibition of human breast carcinoma cell lines [41]. The estradiol-induced Angiogenin (*ANG*) derived from cancer cells significantly increases endothelial cell proliferation [22]. The class B contains two genes including gene *SGSH* [16] and *EIF3J* [9], and class C has four genes. The ten genes also can be detected by the Hotelling test. As a result, the proportion of genes in class A is 40%, which indicates that the two-variate PT test is more powerful.

TABLE I
RESULTS OF THREE GENES FOR BMI STUDY IN BREAST CANCER

Gene Symbol	Probe ID	P_{FDA}	$P_{Hotelling}$	P_{NT}	P_{PT}	Transcription ID	Postion
MSR1	A_23_P392942	5.14E-06	3.20E-05	1.39E-13	8.74E-06	NM_002445	chr8:015998022-015997963
	A_33_P3379326	3.67E-06	2.32E-05	1.05E-02	4.17E-06	NM_138715	chr8:015967653-015967594
	A_24_P372223	8.23E-06	5.42E-05	3.28E-03	1.19E-05	NM_138715	chr8:015965525-015965466
LEP	A_24_P397817	4.23E-05	2.37E-04	7.09E-22	2.03E-01	NM_000230	chr7:127897079-127897138
ADIPOQ	A_23_P369237	7.79E-01	9.61E-01	1.87E-21	7.81E-01	NM_004797	chr3:186575311-186575370

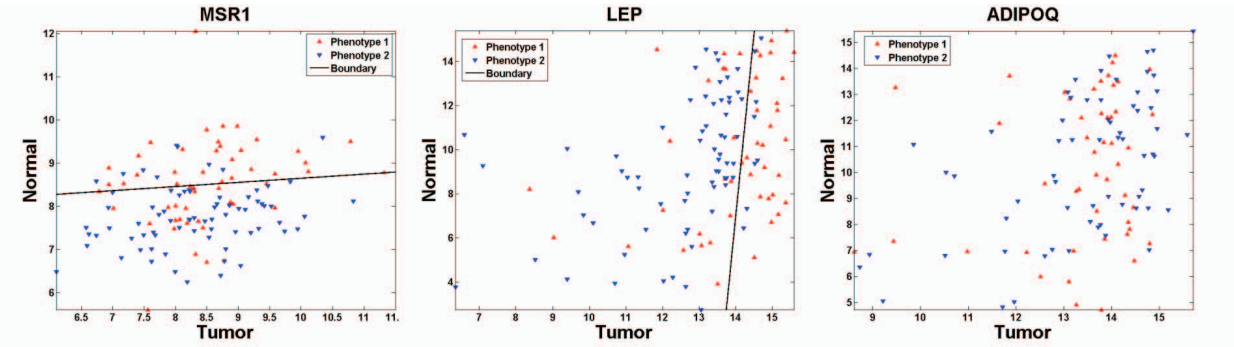

Fig. 3. Scatter plots of three genes

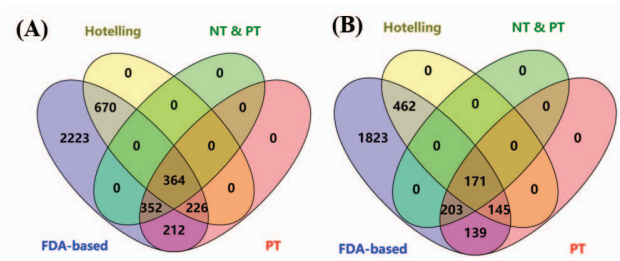

Fig. 4. Venn plots for survival-related and stages-related genes

TABLE II
RESULTS OF TOP 10 GENES ORDERED BY TWO-VARIATE PT TEST

Gene	P_{FDA}	$P_{Hotelling}$	P_{NT}	P_{PT}	Class
FXYD1	7.17E-06	4.58E-05	5.11E-30	2.51E-01	A
ZNF425	7.64E-05	4.27E-04	3.96E-02	4.52E-04	A
IRF2	8.15E-05	4.54E-04	4.75E-04	4.28E-04	A
ANG	1.21E-04	6.56E-04	2.00E-12	2.64E-02	A
SGSH	6.17E-05	3.49E-04	1.90E-01	6.90E-05	B
EIF3J	7.78E-05	4.34E-04	1.39E-02	2.29E-02	B
MARCH2	6.47E-06	4.15E-04	1.92E-04	2.80E-05	C
PCYOX1L	1.12E-04	6.12E-04	1.22E-03	1.87E-02	C
UROD	1.28E-04	6.94E-04	4.56E-06	7.57E-04	C
C5orf22	1.29E-04	6.97E-04	9.36E-07	1.41E-03	C

We also collected 20 overall survival-related genes as benchmarkers for the evaluation and related results were shown in Table III. Among them, nine genes can be identified by at least one of four methods including PT test, Logrank test and two implementations of two-variate PT test ($\alpha = 0.05$). For calculation of *P* value of Logrank test, we dichotomized the expression profiles of those genes into binary values according to their mean value. Five out of the nine genes can be identified by FDA-based two-variate PT test including gene *TP53*, *TIMP2*, *KLF5*, *VDR* and *FOXP3*. The FDA-based two-variate PT test can detect biomarkers that are detected by the Hotelling test. Both PT test and FDA-based two-variate PT test can detect gene *ITGB1* and *MTDH* while Logrank test cannot. Furthermore, gene *ERBB2* and *CRP* cannot be detected by both PT test and FDA-based two-variate PT test while the Logrank test can. The reason might be that Logrank test focuses on the difference of survival time distribution between the high-expression group and low-expression group while both two-variate PT test and PT test focus on the difference of expression profiles of one gene on good-survival and poor-survival group. In a word, the FDA-based two-variate PT test is more powerful for detection of survival-related biomarkers in breast cancer.

Finally, we also intend to show the role of tumor-adjacent normal data played in two-variate PT test. Top 30 genes ordered by *P* values of two-variate PT test were chosen, and the heatmap of tumor data and projection values of them was offered in Fig.5. The top panel shows projection values of 30 genes and the bottom panel shows the tumor data. The difference between the two groups is more distinct after projection, and the projection values are more helpful for detection of phenotypic genes.

C. Pathologic stage in breast cancer

The pathologic stage is a crucial phenotype for cancers. With the development of pathologic stage, the survival outcomes of patients are poorer and poorer [30]. As shown in Fig.4(B), the genes which are detected by other three methods can also be detected by the FDA-based two-variate PT test.

D. Tumor size in breast cancer

There are limited researches on the relationship between genes and tumor size in breast cancer. While the tumor size is associated with the survival rate in breast cancer [2]. In

TABLE III
RESULTS OF 20 SURVIVAL-RELATED BIOBIOMARKERS

Gene Symbol	Description	P_{FDA}	$P_{Hotelling}$	P_{NT}	P_{PT}	$P_{Logrank}$	Ref
TP53	Tumor Protein P53	4.16E-03	1.70E-02	3.64E-01	3.70E-01	4.95E-01	[7]
ITGB1	Integrin Subunit Beta 1	1.40E-02	5.00E-02	2.58E-04	3.14E-02	7.50E-02	[40]
TIMP2	TIMP Metallopeptidase Inhibitor 2	2.11E-02	7.15E-02	2.93E-03	9.29E-01	9.39E-01	[18]
KLF5	Kruppel Like Factor 5	2.20E-02	7.40E-02	1.78E-05	5.04E-01	9.75E-01	[35]
MTDH	Metadherin	2.65E-02	8.70E-02	3.52E-07	2.87E-02	9.49E-01	[13]
VDR	Vitamin D Receptor	3.72E-02	1.16E-01	6.72E-03	3.54E-01	9.78E-01	[5]
FOXP3	Forkhead Box P3	4.36E-02	1.33E-01	7.89E-11	1.11E-01	3.42E-01	[17]
BCL2	BCL2 Apoptosis Regulator	6.21E-02	1.78E-01	1.27E-04	5.95E-01	9.53E-01	[29]
PARP1	Poly(ADP-Ribose) Polymerase 1	8.99E-02	2.40E-01	2.66E-20	2.42E-01	4.31E-01	[28]
EPCAM	Epithelial Cell Adhesion Molecule	1.31E-01	3.22E-01	1.34E-13	7.25E-01	4.85E-01	[8]
EGFR	Epidermal Growth Factor Receptor	1.07E-01	2.75E-01	1.00E-02	4.79E-01	7.08E-01	[20]
PTGS2	Prostaglandin-Endoperoxide Synthase 2	1.08E-01	2.77E-01	1.85E-08	2.27E-01	3.02E-01	[3]
ERBB2	Erb-B2 Receptor Tyrosine Kinase 2	2.56E-01	5.27E-01	2.89E-03	9.35E-01	2.36E-02	[31]
HERC5	Hect Domain And RLD 5	2.69E-01	5.46E-01	3.42E-01	5.46E-01	1.13E-01	[12]
GLI1	GLI Family Zinc Finger 1	2.88E-01	5.71E-01	5.24E-02	3.10E-01	4.04E-01	[34]
CRP	C-Reactive Protein	3.03E-01	5.90E-01	7.12E-01	3.82E-01	7.58E-03	[25]
ESR1	Estrogen Receptor 1	3.89E-01	6.93E-01	1.33E-04	4.19E-01	9.39E-01	[23]
CTSD	Cathepsin D	4.35E-01	7.40E-01	5.07E-08	4.97E-01	2.69E-01	[33]
HDAC6	Histone Deacetylase 6	4.85E-01	7.85E-01	2.62E-03	5.68E-01	5.37E-01	[42]
AR	Androgen Receptor	5.48E-01	8.36E-01	1.92E-04	6.36E-01	6.00E-01	[10]

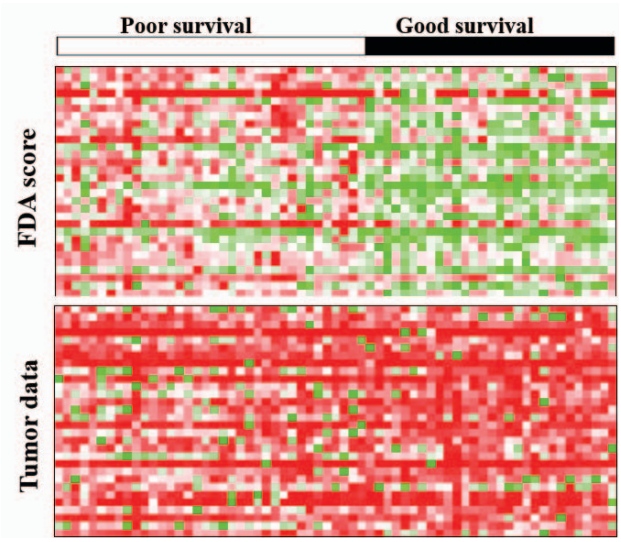

Fig. 5. Heatmap of top 30 genes

this section, we apply the FDA-based two-variate PT test to identify tumor size-related genes. In the dataset GSE70951, there are two cohorts from two different platforms including GPL4133 and GPL13607. For the cohort with GPL4133, the sample size is 46 as one sample has no tumor size data while the sample size is 143 for the other cohort in which five samples have missing value. We remove all probes with missing values. The number of detected genes on GPL4133 is 86 and 175 for GPL13607 by two-variate PT test ($\alpha = 0.01$). Only gene *TCTEX1D2* can be identified on two platforms ($P_{FDA}^{GPL4133} = 1.90\text{E-}03$ and $P_{FDA}^{GPL13607} = 9.99\text{E-}03$) and we consider to be a potential biomarker about tumor size.

V. CONCLUDING REMARKS

We applied the two-variate PT test for detection of phenotypic biomarkers in clinical investigation on breast cancer. Different from traditional methods, tumor-adjacent normal data or non-treatment data are also taken into consideration, resulting in higher detection power. Applied on the expression profiles of breast cancer, we observed a high detection power and reliability. Also, it is shown that the tumor-adjacent normal data does help to detect phenotypic biomarkers, and that the FDA-based implementation is more powerful than the Hotelling test. Finally, we found that the gene *TCTEX1D2* is associated with tumor size of breast cancer.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC 61802256), and the Zhi-Yuan Chair Professorship Start-up Grant (WF220103010), and Startup Fund (WF220403029) for Youngman Research, from Shanghai Jiao Tong University.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] C. L. Carter, C. Allen, and D. E. Henson, "Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases," *Cancer*, vol. 63, no. 1, pp. 181–187, 1989.

[3] C. Denkert, K.-J. Winzer, B.-M. Müller, W. Weichert, S. Pest, M. Köbel, G. Kristiansen, A. Reles, A. Siegert, H. Guski *et al.*, "Elevated expression of cyclooxygenase-2 is a negative prognostic factor for disease free survival and overall survival in patients with breast carcinoma," *Cancer*, vol. 97, no. 12, pp. 2978–2987, 2003.

[4] M. M. Desouki, S. Liao, H. Huang, J. Conroy, N. J. Nowak, L. Shepherd, D. P. Gaile, and J. Geradts, "Identification of metastasis-associated breast cancer genes using a high-resolution whole genome profiling approach," *Journal of cancer research and clinical oncology*, vol. 137, no. 5, pp. 795–809, 2011.

[5] N. Ditsch, B. Toth, D. Mayr, M. Lenhard, J. Gallwas, T. Weissenbacher, C. Dannecker, K. Friese, and U. Jeschke, "The association between vitamin D receptor expression and prolonged overall survival in breast cancer," *Journal of Histochemistry & Cytochemistry*, vol. 60, no. 2, pp. 121–129, 2012.

[6] G. Finak, N. Bertos, F. Pepin, S. Sadekova, M. Souleimanova, H. Zhao, H. Chen, G. Omeroglu, S. Meterissian, A. Omeroglu *et al.*, "Stromal gene expression predicts clinical outcome in breast cancer," *Nature medicine*, vol. 14, no. 5, pp. 518–527, 2008.

[7] K. Friedrichs, S. Gluba, H. Eidtmann, and W. Jonat, "Overexpression of p53 and prognosis in breast cancer," *Cancer*, vol. 72, no. 12, pp. 3641–3647, 1993.

[8] G. Gastl, G. Spizzo, P. Obrist, M. Dünser, and G. Mikuz, "Ep-cam overexpression in breast cancer as a predictor of survival," *The Lancet*, vol. 356, no. 9246, pp. 1981–1982, 2000.

[9] J. W. Hershey, "Regulation of protein synthesis and the role of eif3 in cancer," *Brazilian Journal of Medical and Biological Research*, vol. 43, no. 10, pp. 920–930, 2010.

[10] R. Hu, S. Dawood, M. D. Holmes, L. C. Collins, S. J. Schnitt, K. Cole, J. D. Marotti, S. E. Hankinson, G. A. Colditz, and R. M. Tamimi, "Androgen receptor expression and breast cancer survival in postmenopausal women," *Clinical cancer research*, vol. 17, no. 7, pp. 1867–1874, 2011.

[11] H. Huang, X. Wen, S. Tu, J. Ji, R. Chen, and L. Xu, "An Enviro-Geno-Pheno state analysis framework for biomarker study," in *International Conference on Intelligent Science and Big Data Engineering*. Springer, 2018, pp. 663–671.

[12] K. Keyomarsi, S. L. Tucker, T. A. Buchholz, M. Callister, Y. Ding, G. N. Hortobagyi, I. Bedrosian, C. Knickerbocker, W. Toyofuku, M. Lowe *et al.*, "Cyclin E and survival in patients with breast cancer," *New England Journal of Medicine*, vol. 347, no. 20, pp. 1566–1575, 2002.

[13] J. Li, N. Zhang, L.-B. Song, W.-T. Liao, L.-L. Jiang, L.-Y. Gong, J. Wu, J. Yuan, H.-Z. Zhang, M.-S. Zeng *et al.*, "Astrocyte elevated gene-1 is a novel prognostic marker for breast cancer progression and overall patient survival," *Clinical Cancer Research*, vol. 14, no. 11, pp. 3319–3326, 2008.

[14] J.-X. Lv, H.-C. Huang, R.-S. Chen, and L. Xu, "A comparison study on multivariate methods for joint-snvs association analysis," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2016, pp. 1771–1776.

[15] X.-J. Ma, R. Salunga, J. T. Tuggle, J. Gaudet, E. Enright, P. McQuary, T. Payette, M. Pistone, K. Stecker, B. M. Zhang *et al.*, "Gene expression profiles of human breast cancer progression," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5974–5979, 2003.

[16] S. F. Mahmood, N. Gruel, E. Chapeaublanc, A. Lescure, T. Jones, F. Reyal, A. Vincent-Salomon, V. Raynal, G. Pierron, F. Perez *et al.*, "A siRNA screen identifies RAD21, EIF3H, CHRAC1 and TANC2 as driver genes within the 8q23, 8q24. 3 and 17q23 amplicons in breast cancer with effects on cell growth, survival and transformation," *Carcinogenesis*, vol. 35, no. 3, pp. 670–682, 2013.

[17] A. Merlo, P. Casalini, M. L. Carcangiu, C. Malventano, T. Triulzi, S. Menard, E. Tagliabue, and A. Balsari, "FOXP3 expression and overall survival in breast cancer," *J Clin Oncol*, vol. 27, no. 11, pp. 1746–1752, 2009.

[18] L. Nakopoulou, I. Tsirmpa, P. Alexandrou, A. Louvrou, C. Ampela, S. Markaki, and P. S. Davaris, "MMP-2 protein in invasive breast cancer and the impact of MMP-2/TIMP-2 phenotype on overall survival," *Breast cancer research and treatment*, vol. 77, no. 2, pp. 145–155, 2003.

[19] J.-S. Nam, S. Hirohashi, and L. M. Wakefield, "Dysadherin: a new player in cancer progression," *Cancer letters*, vol. 255, no. 2, pp. 161–169, 2007.

[20] R. Nicholson, J. Gee, and M. . Harper, "EGFR and cancer prognosis," *European journal of cancer*, vol. 37, pp. 9–15, 2001.

[21] J. M. Nigro, A. Misra, L. Zhang, I. Smirnov, H. Colman, C. Griffin, N. Ozburn, M. Chen, E. Pan, D. Koul *et al.*, "Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma," *Cancer research*, vol. 65, no. 5, pp. 1678–1686, 2005.

[22] U. W. Nilsson, A. Abrahamsson, and C. Dabrosin, "Angiogenin regulation by estradiol in breast tissue: tamoxifen inhibits angiogenin nuclear translocation and antiangiogenin therapy reduces breast cancer growth in vivo," *Clinical Cancer Research*, vol. 16, no. 14, pp. 3659–3669, 2010.

[23] A. A. Onitilo, J. M. Engel, R. T. Greenlee, and B. N. Mukesh, "Breast cancer subtypes based on er/pr and her2 expression: comparison of clinicopathologic features and survival," *Clinical medicine & research*, vol. 7, no. 1-2, pp. 4–13, 2009.

[24] B. Pardini, D. De Maria, A. Francavilla, C. Di Gaetano, G. Ronco, and A. Naccarati, "MicroRNAs as markers of progression in cervical cancer: a systematic review," *BMC cancer*, vol. 18, no. 1, pp. 696–712, 2018.

[25] B. L. Pierce, R. Ballard-Barbash, L. Bernstein, R. N. Baumgartner, M. L. Neuhouser, M. H. Wener, K. B. Baumgartner, F. D. Gilliland, B. E. Sorensen, A. McTiernan *et al.*, "Elevated biomarkers of inflammation are associated with reduced survival among breast cancer patients," *Journal of Clinical Oncology*, vol. 27, no. 21, pp. 3437–3444, 2009.

[26] D. A. Quigley, A. Tahiri, T. Lüders, M. H. Riis, A. Balmain, A.-L. Børresen-Dale, I. Bukholm, and V. Kristensen, "Age, estrogen, and immune response in breast adenocarcinoma and adjacent normal tissue," *Oncoimmunology*, vol. 6, no. 11, p. e1356142, 2017.

[27] M. Rahman, L. K. Jackson, W. E. Johnson, D. Y. Li, A. H. Bild, and S. R. Piccolo, "Alternative preprocessing of rna-sequencing data in The Cancer Genome Atlas leads to improved analysis results," *Bioinformatics*, vol. 31, no. 22, pp. 3666–3672, 2015.

[28] F. Rojo, J. Garcia-Parra, S. Zazo, I. Tusquets, J. Ferrer-Lozano, S. Menendez, P. Eroles, C. Chamizo, S. Servitja, N. Ramirez-Merino *et al.*, "Nuclear PARP-1 protein overexpression is associated with poor overall survival in early breast cancer," *Annals of oncology*, vol. 23, no. 5, pp. 1156–1164, 2011.

[29] R. Silvestrini, S. Veneroni, M. G. Daidone, E. Benini, P. Boracchi, M. Mezzetti, G. Di Fronzo, F. Rilke, and U. Veronesi, "The Bcl-2 protein: a prognostic indicator strongly related to p53 protein in lymph node-negative breast cancer patients," *JNCI: Journal of the National Cancer Institute*, vol. 86, no. 7, pp. 499–504, 1994.

[30] S. E. Singletary and J. L. Connolly, "Breast cancer staging: working with the sixth edition of the ajcc cancer staging manual," *CA: a cancer journal for clinicians*, vol. 56, no. 1, pp. 37–47, 2006.

[31] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire, "Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene," *science*, vol. 235, no. 4785, pp. 177–182, 1987.

[32] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 397–420.

[33] A. K. Tandon, G. M. Clark, G. C. Chamness, J. M. Chirgwin, and W. L. McGuire, "Cathepsin D and prognosis in breast cancer," *New England Journal of Medicine*, vol. 322, no. 5, pp. 297–302, 1990.

[34] A. ten Haaf, N. Bektas, S. von Serenyi, I. Losen, E. C. Arweiler, A. Hartmann, R. Knüchel, and E. Dahl, "Expression of the glioma-associated oncogene homolog (GLI) 1 in human breast cancer is associated with unfavourable overall survival," *BMC cancer*, vol. 9, no. 1, pp. 298–309, 2009.

[35] D. Tong, K. Czerwenka, G. Heinze, M. Ryffel, E. Schuster, A. Witt, S. Leodolter, and R. Zeillinger, "Expression of klf5 is a prognostic factor for disease-free survival and overall survival in patients with breast cancer," *Clinical Cancer Research*, vol. 12, no. 8, pp. 2442–2448, 2006.

[36] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[37] L. Xu, "Bi-linear matrix-variate analyses, integrative hypothesis tests, and case-control studies," in *Applied Informatics*, vol. 2, no. 1. SpringerOpen, 2015, pp. 1–39.

[38] ——, "A new multivariate test formulation: theory, implementation, and applications to genome-scale sequencing and expression," in *Applied Informatics*, vol. 3, no. 1. Springer, 2016, pp. 1–23.

[39] L.-X. Yan, X.-F. Huang, Q. Shao, M.-Y. Huang, L. Deng, Q.-L. Wu, Y.-X. Zeng, and J.-Y. Shao, "Microrna miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis," *Rna*, vol. 14, no. 11, pp. 2348–2360, 2008.

[40] E. S. Yao, H. Zhang, Y.-Y. Chen, B. Lee, K. Chew, D. Moore, and C. Park, "Increased β1 integrin is associated with decreased survival in invasive breast cancer," *Cancer Research*, vol. 67, no. 2, pp. 659–664, 2007.

[41] J. H. Yim, S. H. Ro, J. K. Lowney, S. J. Wu, J. Connett, and G. M. Doherty, "The role of interferon regulatory factor-1 and interferon regulatory factor-2 in IFN-γ growth inhibition of human breast carcinoma cell lines," *Journal of interferon & cytokine research*, vol. 23, no. 9, pp. 501–511, 2003.

[42] Z. Zhang, H. Yamashita, T. Toyama, H. Sugiura, Y. Omoto, Y. Ando, K. Mita, M. Hamaguchi, S.-i. Hayashi, and H. Iwase, "HDAC6 expression is correlated with better survival in breast cancer," *Clinical Cancer Research*, vol. 10, no. 20, pp. 6962–6968, 2004.