Mathematical Biosciences 311 (2019) 39–48

Contents lists available at ScienceDirect

Mathematical Biosciences

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

A novel matched-pairs feature selection method considering with tumor purity for differential gene expression analyses

Liang Sen[a,b,c], Yang Sen[a], Liang Dayang[d], Ma Jiechao[c], Tian Yuan[a,e], Zhao Jing[f], Zhang Xu[g], Xu Ying[a,b,g], Wang Yan[a,b,*]

[a] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, and College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] Cancer Systems Biology Center, China-Japan Union Hospital, Jilin University, Changchun 130033, China
[c] Advanced Institute, Infervision, Beijing 100000, China
[d] School of Mechatronics Engineering, Nanchang University, Nanchang 330031, China
[e] School of Artificial Intelligence, Jilin University, Changchun, 130012, China
[f] Sanford Research, Sioux Falls, SD 57104, USA
[g] Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Keywords:
Feature selection
Tumor purity
Test statistic
Gene expression analyses, Matched case-control design

ABSTRACT

Tissue-based gene expression data analyses, while most powerful, represent a significantly more challenging problem compared to cell-based gene expression data analyses, even for the simplest differential gene expression analyses. The result in determining if a gene is differentially expressed in tumor vs. non-tumorous control tissues does not only depend on the two expression values but also on the percentage of the tissue cells being tumor cells, i.e., the tumor purity. We developed a novel matched-pairs feature selection method, which takes into full consideration of the tumor purity when deciding if a gene is differentially expressed in tumor vs. control experiments, which is simple, effective, and accurate. To evaluate the validity and performance of the method, we have compared it with four published methods using both simulated datasets and actual cancer tissue datasets and found that our method achieved better performance with higher sensitivity and specificity than the other methods. Our method was the a matched-pairs feature selection method on gene expression analysis under matched case-control design which takes into consideration the tumor purity information, which can set a foundation for further development of other gene expression analysis needs.

1. Introduction

Compared to the traditional cell-based gene expression data collected under laboratory conditions, tissue-based gene expression data analyses enable researchers to study cancer evolution in the actual cancer-forming environment directly. In addition, direct collection of tissue-level gene expression data without separating tissues into distinct cell types, followed by single-cell sequencing provides a considerably more efficient and economically more feasible approach for large-scale tumor data analyses. However, the approach poses a significant challenge to bioinformatics researchers since the collected data are compositions of gene expressions from multiple cell types. At the forefront of the analysis of such data is the issue of tumor purity, i.e., the percentage of tissue cells being tumor cells as the meaning of observed gene expression data changes with the different percentage of the tissue cells being tumor cells.

The tumor purity issue has been well recognized as a technical issue that needs to be solved before reliable information can be derived from tissue-based expression data [1,2]. Aran et al. recently gave a systematic pan-cancer analysis on tumor purity [3] and found that some immunotherapy gene signatures were not detected by traditional differential expression analysis, but became detectable when tumor purity was taken into consideration. Different types of information have been employed in the published methods for tumor purity estimation, including gene expression data (ESTIMATE [4]), somatic copy-number variation data (ABSOLUTE [5], THetA [6] and others [7]), somatic

* Corresponding author at: Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, and College of Computer Science and Technology, Jilin University, Changchun 130012, China.
E-mail addresses: hawkcoder@gmail.com (S. Liang), yangsen18@mails.jlu.edu.cn (S. Yang), 5910116336@email.ncu.edu.cn (D. Liang), mjiechao@infervision.com (J. Ma), tianyuan12@mails.jlu.edu.cn (Y. Tian), zj1228@gmail.com (J. Zhao), xuhzhang@outlook.com (X. Zhang), xyn@uga.edu (Y. Xu), wy6868@jlu.edu.cn (Y. Wang).

https://doi.org/10.1016/j.mbs.2019.02.007
Received 17 January 2019; Received in revised form 21 February 2019; Accepted 22 February 2019
Available online 27 February 2019
0025-5564/ © 2019 Elsevier Inc. All rights reserved.

mutation data (PurityEst [8] and Accurity [9]), and DNA methylation data (InfiniumPurify [10], MethylPurify [11] and others [12,13]). Some of these methods are reviewed in detail in Wang et al. [14]. While numerous methods have been published to estimate tumor purity, the issue largely remains a challenge regarding how to use such purity information in gene expression data analyses in a more informed manner.

Here we present a matched-pairs feature selection method to tackle this issue. The basic idea of the method is a paired *t*-test widely used in matched case-control design research, which has been used extensively in gene expression data analyses [27]. Liang et al. [15] has recently reviewed these methods which can be categorized into three groups: (1) test statistic, followed by a classification approach [16,17]; (2) conditional logistic regression, which considers dependence among features [18,19]; and (3) boosting strategy, which combine multiple weak classifiers [20]. Our method falls into the first category. The main contribution of our study here is to integrate the tumor purity information into this framework. Initially, paired *t*-test was widely used in gene expression analysis to screen out informative marker genes [21], Then, Tan et al. developed a modified paired *t*-test statistic to identify a subset of relevant features that served as a basis for classification via support vector machines [17]. Recently, Cao et al. proposed another modified version of paired *t*-test statistic using fold-change value with the hypothesis that different paired data have different experimental environments and conditions [16]. And Cao et al. believed that the measurement of the difference between case sample and control sample in originally paired *t*-test is unstable and lack of enough generalization ability. Although they put forward many improvements on paired *t*-test method to make it more accurate and effective in the gene expression analysis, they overlooked the purity information of tumor tissue. Moreover, there were some recently works considerating with tumor purity. Wang et al. developed an unsupervised deconvolution method (UNDO) to dissect mixed gene expression in heterogeneous tumor samples, which can be used as a purity levels prediction method [22]. Shen et al. proposed contamDE method account for the contamination of tumor samples with the matched and unmatched sample situation which did not require any extra information [23]. Zhang et al. developed a novel differential gene method (DECtp) by integrating tumor purity information into a generalized least square procedure and following with a Wald test, whose purity information was known estimated by existing methods [24].

The major novelty of the proposed method in this study is as follows. We first infer the true expression level of each gene in the case samples, which under matched case-control design experiments, taking estimated tumor purity into consideration based on the observed expression level, which is followed by a revised paired *t*-test methods. To evaluate the performance of the proposed method, we compared it with four popular methods on both simulated and actual tumor tissue datasets. Our method achieved a better performance than the other methods on both datasets. The rest of this study is organized as follows: Section 2 introduces our method as well as the evaluation metrics. Section 3 demonstrates the results of comparing our method with four other methods using both test and real datasets. Section 4 discusses the advantages and disadvantages of the proposed method and points out the future research directions. Finally, a conclusion is given in Section 5.

2. Materials and methods

2.1. Matched-pair feature selection description

Considering n paired data samples $X = \{X_i | i = 1, 2, ..., n\}$ under 1:m case-control design [15]. For each sample X_i, it has p case experiments and $q = mp$ control experiments, where m represents the ratio between the number of control and case samples. Let $Z_{i,j}, j = 1, \ldots, p + q$, denote the case/control status of the jth experiment of the ith sample $X_i = \{X_{i,j} | j = 1, 2, ..., p + q\}$, while $Z_{i,j} = 1$ representing it was a case experiment and 0 a control experiment. Each sample has K features, denoted as $X_{i,j} = \{x_{i,j,k} | k = 1, 2, ..., K\}$. The aim of our paired feature selection method was to find the largest subset of features from all the K features with the considerison of the paired case/control status. In the setting of gene expression data analysis, each sample has one case experiment and one control experiment.

2.2. The standard paired t-test for feature selection

The paired *t*-test statistic method [21] because of its simplicity and effectiveness, has been widely used in identifying biomarker genes in differential gene expression analyses. The main idea of the approach is to estimate the statistical significance *p*-value for each feature, assuming that it follows a Student's *t*-distribution. We then select features with *p*-values below a specified threshold. For each paired sample under 1:1 case-control design, the test statistic t_k for the kth feature is given by

$$t_k = \bar{d}_k \ / \ s_k \tag{1}$$

where $\bar{d}_k$ and s_k is the mean difference and the standard error of the kth feature across n paired samples, denoted as

$$\bar{d}_k = (1/n) \sum_{i=1}^{n} d_{i,k} \tag{2}$$

$$s_k = \sqrt{\sum_{i=1}^{n} (d_{i,k} - \bar{d}_k)^2/(n-1)} \tag{3}$$

where $d_{i,k}$ is the differential between the paired case and control for the kth feature of the ith sample, given as

$$d_{i,k} = X_{i,1,k} - X_{i,2,k} \tag{4}$$

2.3. Our method using the tumor purity information

Here, we assume that each control was a pure control, hence no need for purity-based correction. For each case sample, we predict the expression level of each gene in a pure tumor sample based on the estimated purity level and the observed gene expression level in the sample. Then we conduct a paired *t*-test as outlined above with the following modification, i.e., add a positive constant to the denominator of *t*-test statistic to improve the generalizability of our method, as done in [17]. And the workflow of our method is shown in Fig. 1 and described below.

Let the purity of the experiment j of sample i be $pr_{i,j}$, which is estimated by existing methods, for a case experiment in this study, and let $pr_{i,j} = 1$ for a control experiment. For the 1:1 case-control setting, each sample i contains one case experiment and one paired control experiment, where $Z_{i,1} = 1$, $Z_{i,2} = 0$, and $p = q = 1$. We use $X^{pure}_{i,case,k}$ to estimate

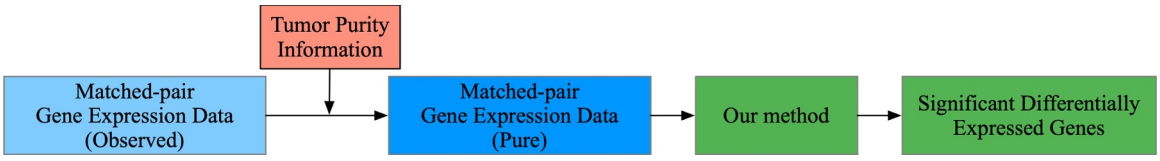

Fig. 1. The workflow of the proposed method. The inputs are observed matched-pair gene expression data and tumor purity information which can be estimated by known methods, and the outputs are selected significant differentially expressed genes. We first conduct purity-based correction to product pure matched-pair gene expression data, and then make gene selection with our novel modified paired *t*-test method.

the pure expression level of gene k in sample i with purity information as

$$X_{i,case,k}^{pure} = \frac{X_{i,1,k} - (1 - pr_{i,1})X_{i,2,k}}{pr_{i,1}} \tag{5}$$

For 1: m case-control settings, each sample i contains p case experiments and q paired control experiments, where $Z_{i,j} = 1, j = 1, 2, ...,p$ and $Z_{i,j} = 0, j = p + 1, p + 2, ...,p + q$. We use $X_{i,\ control,\ k}$ estimate the average experssion level of control experiments, and $X_{i,case,k}^{pure}$ estimate the pure expression level of case experiments, for gene k in sample i within q control experiments and p case experiments, as

$$X_{i,control,k} = \frac{1}{q}\sum_{j=p+1}^{p+q} X_{i,j,k}. \tag{6}$$

$$X_{i,case,k}^{pure} = \frac{1}{p}\sum_{j=1}^{p} \frac{X_{i,j,k} - (1 - pr_{i,j})X_{i,control,k}}{pr_{i,j}} \tag{7}$$

In both 1:1 and 1:m case-control settings, we define the revised fold-change $d'_{i,k}$ to measure the difference between paired case and control data as

$$d'_{i,k} = \begin{cases} FC_{i,k} - 1; & if \quad FC_{i,k} \geq 1 \\ 1 - 1/FC_{i,k}; & if \quad FC_{i,k} < 1 \end{cases} \tag{8}$$

$$FC_{i,k} = \frac{X_{i,case,k}^{pure}}{X_{i,control,k}} \tag{9}$$

We then add a positive constant s_0 to the dominantor of t-test statistic, so that it can smooth the t-test statistic when some genes have small standard deviations, where we set s_0 to be the median of all the standard deviations across all K features. The new t-test statistic t'_k is as

$$t'_k = \frac{\bar{d}'_k}{s_0 + s'_k} \tag{10}$$

where the modified $\bar{d}'_k$ and s'_k with fold-change and purity information are calculated as

$$\bar{d}'_k = \frac{1}{n}\sum_{i=1}^{n} d'_{i,k} \tag{11}$$

$$s'_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (d'_{i,k} - \bar{d}'_k)^2} \tag{12}$$

2.4. Evaluation criteria and measurement

In this section, we applied seven metrics to evaluate the performance of our model: specificity, sensitivity (recall), false positive rate, precision and Precision–Recall (P–R) curve. In this study, we denoted true differentially expression genes (tDEGs) as these genes that really have a difference between normal samples and tumor samples, and prediction differentially expression gene (pDEGs) as these genes that were predicted by feature selection method, which contain tDEGs and false tDEGs.

Assuming each sample have N genes, and this dataset have G tDEGs. If one feature selection method predicted M pDEGs, which F genes were tDEGs and $M - F$genes was not tDEGs. We defined true positive (TP) genes as the quantity of the given label that is positive, and the predicted result is also a positive, where $TP = F$; we defined true negative (TN) as the quantity of the given label that is negative, and the predicted result is also a negative, where $TN = G - F$; we defined false positive (FP) as the quantity of the given label that is negative, but the predicted result is positive, where $FP = M - F$; and defined false negative (FN) as the quantity of the given label that is positive, but the predicted result is negative, where $FN = N + F - G - M$.

Specificity was the true negative rate that is correctly identified, as in Eq. (13). Sensitivity, also known as recall, was the positive predictions that were actually positive, as in Eq. (14). False positive rate was the fraction of false positive across all prediction result, as in Eq. (16). Precision was the fraction of true positive across all prediction result, as in Eq. (15). In the experiments, we also applied P–R curve to show the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

$$\text{specificity} = \frac{TN}{TN + FP} \tag{13}$$

$$\text{sensitivity} = \text{recall} = \frac{TP}{TP + FN} \tag{14}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{false positive rate} = \frac{FP}{TP + FP} \tag{16}$$

3. Results

3.1. Performance on simulated data

To evaluate the performance of our method, we simulated four scenarios with different paired samples and genes. For each simulated scenario, we assumed that the expression distribution of each gene across all samples follows a normal distribution. Consider we have N paired samples with K genes with M being differentially expressed. When simulating the distribution of each gene, we first generated pure case data and control data, followed by generation of mixed case data which were the mixture of pure case and control data. We applied the following procedure for generation of the simulated data:

Simulating the purity for each sample. We assumed the purity pr of each sample follows a normal distribution as $pr_i = N(\mu_{pr}, \sigma^2_{pr})$, $i = 1, 2, ...,N$, where in this study we set $\mu_{pr} = 0.6$ and $\sigma_{pr} = 0.1$. Simulating M differentially expressed genes.

For each gene, we simulated its ith control data for N control samples as $X_{i,2,k} = N(\mu 2_k, \sigma 2^2_k)$, $i = 1, 2, ...,N$, where $\mu 2_k$ was the mean expression of gene k as $\mu 2_k = N(100, 10^2)$, $k = 1, 2, ...,K$, and σ^2_k was their variance as $\sigma_k = \mu 2_k \times \beta_k$ and $\beta_k = N(0.1, 0.01^2)$, $k = 1, 2, ...,K$. Then, we simulated the matched pure case data was 5 fold-change larger or smaller as $X_{i,1,k}^{pure} = N(\mu 2_k + \mu d_k, (\sigma 2_k + \sigma d_k)^2)$, $i = 1, 2, ...,N$, where $\mu d_k = \mu 2_k \times \lambda_k$, $\sigma d_k = \sigma 2_k \times \beta_k$, and λ_k following a uniform distribution as $uniform(-0.9, 4)$.

For each differentially expressed genes, we gave a differentially expression rank with a difference level Δ as

$$\Delta = \left| \frac{\sum_{i=1}^{N} X_{i,1,k}^{pure} - \sum_{i=1}^{N} X_{i,2,k}}{\sum_{i=1}^{N} X_{i,2,k}} \right|, \tag{17}$$

where the bigger Δ, the larger the difference between case and control sample, and the rank of this gene more top-ranked genes.

Simulating the remaining genes. For each gene, as similarly like step b), we simulated its pure control data for N samples as $X_{i,2,k} = N(\mu 2_k, \sigma 2^2_k)$, $i = 1, 2, ...,N$, and its paired pure case also as $X_{i,1,k}^{pure} = N(\mu 2_k, \sigma 2^2_k)$, $i = 1, 2, ...,N$, which have the same mean and standard error with control samples.

Then we simulated the mixed case data for each gene k in sample i as $X_{i,1,k} = pr_i X_{i,1,k}^{pure} + (1 - pr_i)X_{i,2,k}$, $i = 1, 2, ...,N$, which was the mixture of pure case data and control data with the purity pr_i.

Overall, our simulated data are as follows:

a) Scenarios 1: Paired samples N = 100, genes K = 1000, differential genes M = 100
b) Scenarios 2: Paired samples N = 100, genes K = 5000, differential genes M = 500

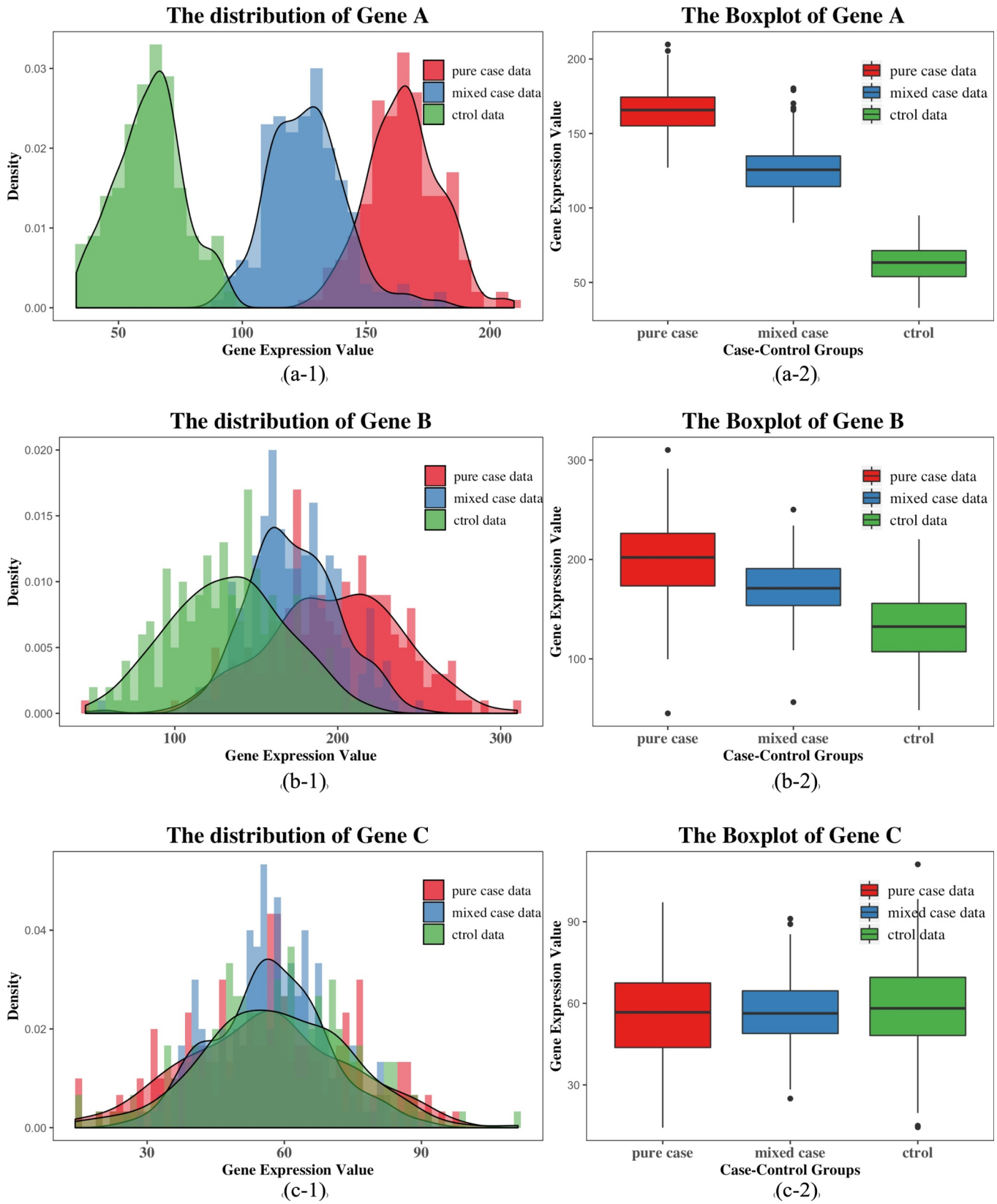

Fig. 2. The distribution of three representational genes of Scenarios 1. The simulation settings of Scenarios 1 was N = 100 samples with K = 1000 genes and M = 100 differential genes. The left column was the distribution of Gene A, B, and C with Figure (a-1), (b-1) and (c-1), respectively. The *X*-axis was the gene expression value and *Y*-axis was the distribution density. The right column was the corresponding boxplot of Gene A, B, and C with Figure (a-2), (b-2) and (c-2), respectively. The colors represented pure case data(red), mixed case samples (blue), and control samples (yellow) respectively. The *X*-axis was the type of different cases and *Y*-axis was the gene expression value.

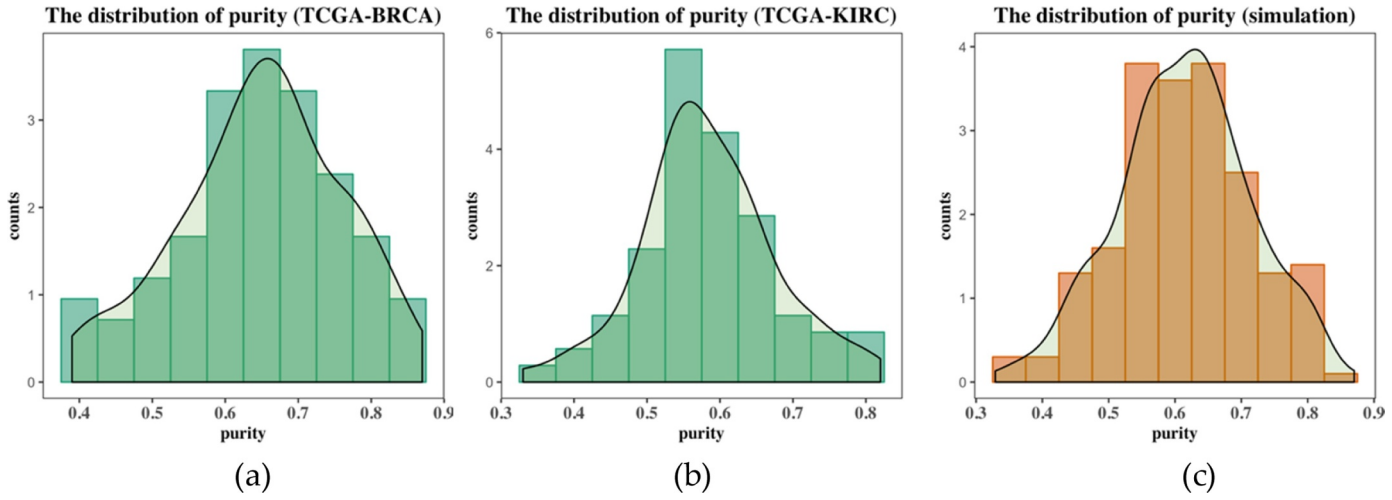

Fig. 3. The distribution of real data purity and simulated purity. Figure (a,b) shows the distribution of purity of the dataset of BRCA and KIRC, which both were the real data from TCGA dataset. Figure (c) shows the distribution of purity of our simulation dataset.

c) Scenarios 3: Paired samples N = 1000, genes K = 5000, differential genes M = 500
d) Scenarios 4: Paired samples N = 1000, genes K = 10,000, differential genes M = 1000

To illustrate our simulation result of genes expression data, we showed three typical genes of Scenarios 1 at Fig. 2. Gene A and B were significant differentially expressed genes, but Gene C was not. However, Gene B had a small difference between control experiments and its mixed case experiments, while having a larger difference between control experiments and its pure case experiments. These feature selection methods without purity information were easy to select out genes like Gene A, but it was difficult to work on genes like Gene B. That was the reason why we proposed this novel matched-pairs feature selection with purity information. To make our simulation purity more practical, we let the simulated distribution of purity more like the distribution of Breast Invasive Carcinoma (BRCA) and Kidney Renal Clear Cell Carcinoma (KIRC) dataset (we will describe them at Section 3.2.1) from the Cancer Genome Atlas (TCGA) dataset [25,26]. Their distributions was illustrated in Fig. 3.

To evaluate the performance of our method, we compared it with four popular methods: 1) a paired *t*-test [21], 2) Tan et al.'s method [17], 3) Cao et al.'s method [16], 4) The joint approach of Tan et al.'s and Cao et al.'s method [16,21], which denoted as Tan + Cao's method. These methods were typical test statistic matched-pairs feature selection methods considering case-control design, which was presented in the Liang et al's survey [15]. We used a P–R curve to measure the prediction performance. Firstly, we predicted a gene to be differentially expressed as its expression mean values in case vs. control samples are different with p-value < 0.05, referred as a pDEGs. Then we ranked these pDEGs in the descending order of their its p-values. Secondly, we considered these M differentially expressed genes which were pre-defined in each simulation scenarios as tDEGs, which was ranked descendingly by Δ. Finally, we drew the P–R curve with the paired precision and recall points for each top m genes, where $m = 1,2,\ldots,M$. Please note that the P–R curve was fragmentary and the maximum recall of each method was less than 1.0, as we only considered the top M differentially genes, where M was 100 (10% of total genes), 500 (10%), 500 (10%) and 1000 (10%) for four scenarios, respectively. The P–R curves of four simulation scenarios were shown at Fig. 4.

To check the ability of identifying tDEGs of each method, we calculated the precisions where 25%, 50% and 75% of pDEGs were involved respectively. The comparing results of scenarios 4 were show in Table 1, where the result of scenarios 1, 2 and 3 were shown in Table A.1–A.3 of the supplementary files, respectively.

The P–R curves of the four simulation scenarios (Fig. 4) showed that our method yielded the largest area under the P–R curve compared to the other four methods, with the detailed numbers given in Table 1.

3.2. Performance on actual tumor-tissue data

3.2.1. Datasets and pre-processing

We used gene expression data of Breast Invasive Carcinoma (BRCA) dataset and Kidney Renal Clear Cell Carcinoma (KIRC) dataset from The Cancer Genome Atlas (TCGA) database [25,26]. We first selected paired samples for analyses of performance. We found 85 paired BRCA samples and 70 paired KIRC samples. Each gene expression profile of the two datasets was in Fragments Per Kilobase Million (FPKM) value, and pre-analyzed by the following procedures: (i) transforming the gene expression by log2(FPKM + min), where min was set to 1 by default; and (ii) filtering genes by p-value < 0.005 (t-test), variance > 0.1, and the fold-change > 1.5 or $< 1/1.5$ between case and control data.

We calculated the purity of each tumor sample by InfiniumPurify [10,13], which predicts the purity level based on an observation that the number of probes with intermediate methylation level is significantly greater in tumor samples than that in normal samples. All the tumor purity data of TCGA was calculated and provided by the authors in https://zenodo.org/record/253193. We acquired these purity data for the assessment. We also conducted experiments in a later section to show the robustness of our method comparing to others purity estimated methods, like ABSOLUTED [5] and ESTIMATE [4] methods.

3.2.2. Comparing methods for differentially expressed gene analysis

We predicted pDEGs by applying our method and the aforementioned four comparing methods with a significance level of 0.05, and ranked these genes in a descending order by their p-values. Unlike the analyses on simulated data with pre-defined tDEGs, we calculated approximate tDEGs as follows: estimating the expression level of each gene based on the observed expression and the estimated purity information; then calculate the difference level Δ of each gene between pure case vs. control samples; next consider those genes whose difference larger than a predefined threshold δ with $\Delta > \delta$ as differentially expressed genes. In our experiment, we let $\delta = 1, 2, 3$. For $\delta = 1$, BRCA and KIRC datasets had 1257 and 1810 tDEGs, respectively. Table 2 summarized the predictions by the five methods.. Our method found 1247 tDEGs in the BRCA datasets, and achieved the best sensitivity of 0.992 and specificity of 0.967 with the lowest false positive rate of 0.223. It also found 1810 tDEGs in KIRC, and achieved the best

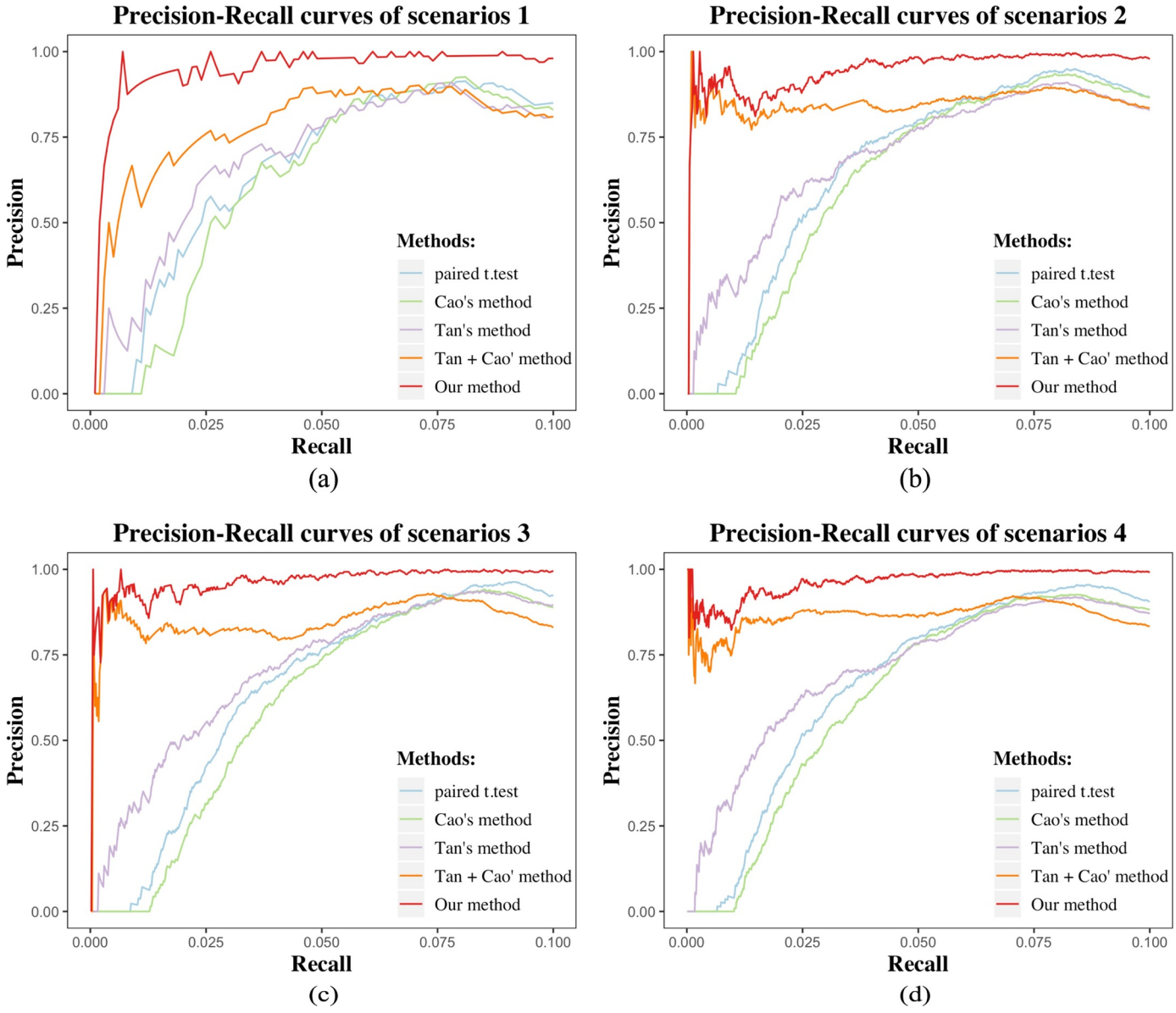

Fig. 4. The P–R curves of four simulation scenarios. Figure (a-d) were the P–R curves of scenarios 1,2,3 and 4, respectively. *X*-axis was the recall metric and *Y*-axis was the precision metric. Each method being compared was plotted by a specific color. Please note that the P–R curve was fragmentary and the maximum recall of each method was less than 1.0, as we only considered the top *M* differentially expressed genes, which were 100 (10% of total genes), 500 (10%), 500 (10%) and 1000 (10%) for four scenarios, regardingly.

sensitivity of 0.992 and specificity of 0.943 with the lowest false positive rate of 0.236. Predictions using $\delta = 2, 3$ were provided in Tables A.4 and A.5, respectively.

Like in our analyses over the simulated data, we considered these estimated tDEGs ($\Delta > 1$) as real tDEGs. With above pDEGs predicted by each method and tDEGs, we have calculated the P–R curves for the five methods as shown in Fig. 5. Clearly, our method has the largest area under the P–R curve compared to the other four methods. We then counted the True Positive genes and calculate the precision when having 25%, 50%, 75% pDEGs. The comparison results were shown in Table 3. Clearly, our method found the most tDEGs with precistion = 0.790, 0.920 and 0.926 when having 25%, 50%, 75% pDEGs in the BRCA dataset, and also found the most tDEGs with precistion = 0.690, 0.864 and 0.904 when involved 25%, 50%, 75% pDEGs in the KIRC dataset.

Moreover, to show our method was also robust when the tumor purity was estimated by other methods except for InfiniumPurify [10,13]. We calculated the P–R curves for TCGA-BRCA dataset with the tumor purity estimated by ABSOLUTED [5] and ESTIMATE [4] methods. The result as shown in Fig. 6, which our method also has the largest area under the P–R curve compared to the other four methods in both tumor purity estimated experiments.

3.2.3. *Enrichment analysis*

We also conducted gene set enrichment analysis against Gene Ontology(GO) [27,28], using the R package clusterProfile [29], and compared the enrichment results with Tan + Cao's method since it is the best among the four methods. GO database classifies functions along three aspects: molecular function (MF), molecular activities of gene products; cellular component (CC), where gene products are active; biological process(BP), pathways and larger processes made up of the activities of multiple gene products. We applied enrichment analysis on these three subtypes databases, respectively. In order to reduce the false discovery rate (FDR) and increase the chances of identifying all the differentially expressed genes, we used the Benjamini Hochberg [29] method to adjust *p*-values. We showed the results of comparing GO

Table 1
Compared the ability of finding tDEGs (scenarios 4).

Top N genes	Methods	TP	Precision
Top 25% **(tDEGs = 250)**	Paired *t* test	130	0.520
	Tan's method	157	0.628
	Cao's method	107	0.428
	Tan + Cao's method	220	0.880
	Our method	**243**	**0.972**
Top 50% **(tDEGs = 500)**	Paired *t* test	400	0.800
	Tan's method	391	0.782
	Cao's method	391	0.782
	Tan + Cao's method	431	0.862
	Our method	**489**	**0.978**
Top 75% **(tDEGs = 750)**	Paired *t* test	700	0.933
	Tan's method	677	0.903
	Cao's method	687	0.916
	Tan + Cao's method	688	0.917
	Our method	**746**	**0.995**

Note: tDEGs, the number of true differentially expressed genes. TP, the number of true positive genes, which was the number of genes which were found in the tDEGs. Precision was the rate of TP and tDEGs. The bold font represented the result of our method.

enrichment results on BP (only top 10 shown) in Tables 4 and 5 for BRCA and KIRC database, respectively, and the result on CC and MF were given in Tables A.6–A.9 of the supplementary materials.

In the enrichment analysis result of the BRCA dataset (Table 4), 90% Go terms appeared in both methods, like sister chromatid segregation (our method adjust p-value = 3.04e−14, Tan + Cao's method adjust p-value = 1.74e−11), mitotic nuclear division (adjust p-value = 2.33e−13 / 3.41e−10), extracellular structure organization (adjust p-value = 2.33e−13 / 1.45e−12) and etc. We can see that these GO terms found by our method have higher significance with the lower adjust p-value than that of Tan + Cao's method. Similar findings were obtained in the enrichment analysis of the KIRC dataset (Table 5), where it has 60% Go terms identified by both methods, like T cell activation (adjust p-value = 1.23e−20 / 3.01e−14), leukocyte cell-cell adhesion (adjust p-value = 5.79e−20 / 1.86e−14), regulation of cell-cell adhesion (adjust p-value = 3.90e−17 / 1.17e−15).

4. Discussion

In this work, we developed a simple and effective matched-pairs feature selection method with purity information. To evaluate the performance of our method with known differentially expressed genes,

Table 2
Comparing the performance of feature selection (Δ > 1).

Dataset	methods	pDEGs	TP	FP	FN	SN	SP	FPR
BRCA **(tDEGs = 1257)**	Paired *t* test	6958	1247	5711	10	0.992	0.645	0.821
	Tan's method	1507	1062	445	195	0.845	0.960	0.295
	Cao's method	7063	1247	5816	10	0.992	0.641	0.823
	Tan + Cao's method	1508	1128	380	129	0.897	0.965	0.252
	Our method	**1605**	**1247**	**358**	**10**	**0.992**	**0.967**	**0.223**
KIRC **(tDEGs = 1810)**	Paired *t* test	7522	1787	5735	23	0.987	0.614	0.762
	Tan's method	2347	1640	797	170	0.906	0.921	0.327
	Cao's method	7711	1788	5923	22	0.988	0.607	0.768
	Tan + Cao's method	2141	1628	513	182	0.899	0.948	0.240
	Our method	**2349**	**1795**	**554**	**15**	**0.992**	**0.943**	**0.236**

Note: tDEGs, the number of true differentially expression genes. pDEGs, the number of prediction differentially expression genes which predicted by each method with p-value < 0.05. TP, true positive. FP, false positive. FN, false negative. SN, sensitivity. SP, specificity. FPR, false positive rate.

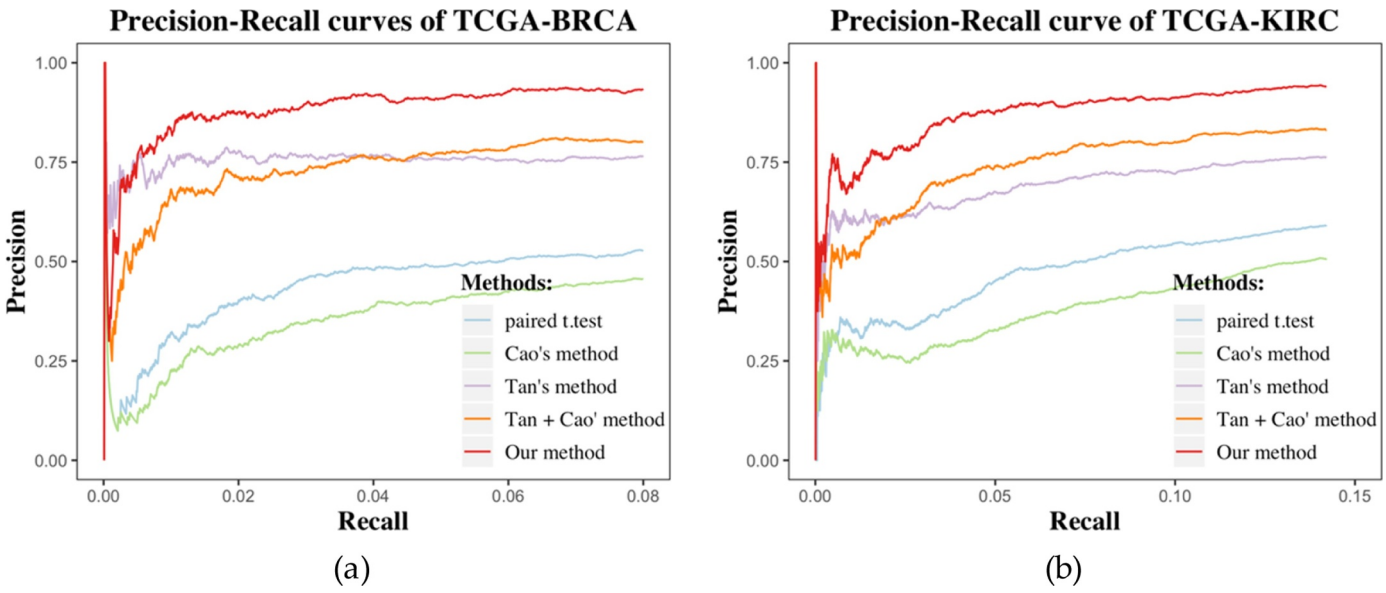

Fig. 5. The P–R curves of two real datasets. Figure(a-b) was the P–R curves of the BRCA and the KIRC dataset, respectively. Each comparing method was lined by a different color. *X*-axis represented the number of top N ranked genes, which were selected to calculate the accuracy. *Y*-axis represented the accuracy of each top N genes between ground truth genes and prediction selected genes. Please note that the P–R curves was also fragmentary, as we only considered these tDEGs, where the BRCA and the KIRC datasets have 1257 and 1810 tDEGs, respectively.

Table 3
Compared the ability of finding tDEGs (real datasets).

Top N genes	methods	BRCA		KIRC	
		TP	Precision	TP	Precision
Top 100 genes **(tDEGs = 100)**	Paired *t* test	25	0.250	34	0.340
	Tan's method	70	0.700	60	0.600
	Cao's method	16	0.160	30	0.300
	Tan + Cao's method	58	0.580	53	0.530
	Our method	**79**	**0.790**	**69**	**0.690**
Top 500 genes **(tDEGs = 500)**	Paired *t* test	242	0.484	195	0.390
	Tan's method	382	0.764	321	0.642
	Cao's method	189	0.378	148	0.296
	Tan + Cao's method	383	0.766	355	0.710
	Our method	**460**	**0.920**	**432**	**0.864**
Top 1000 genes **(tDEGs = 1000)**	Paired *t* test	521	0.521	510	0.510
	Tan's method	763	0.763	719	0.719
	Cao's method	453	0.453	397	0.397
	Tan + Cao's method	800	0.800	791	0.791
	Our method	**926**	**0.926**	**904**	**0.904**

Note: tDEGs, the number of true differentially expressed genes. TP, the number of true positive genes, which was the number of genes which were found in the tDEGs. Precision was the rate of TP and tDEGs. The bold font represented the result of our method.

we built simulated dataset to compare our method with the other four test statistic methods. And the simulated purities fitted the distributions of the empirical purities of BRCA and KIRC datasets from the TCGA website. Results showed that our proposed method had larger area under P–R curve than the other methods, which illustrated that our method performed well on simulated data. Moreover, we evaluated the performance of those methods on two real datasets from TCGA. Our method demonstrated the best sensitivity (BRCA 0.992, KIRC 0.992) and specificity (BRCA 0.967, KIRC 0.943) and also had a larger area under P–R curve than the other methods. At last, we conducted gene set enrichment analysis against GO along with three aspects of BP, CC and MF. In the results from enrichment analysis of the BRCA dataset (Table 4), 90% GO terms appeared in both methods, but our method had a higher significance with the lower adjust p-value than that of Tan + Cao's method. Similar findings were obtained in the enrichment analysis of the KIRC dataset (Table 5), where 60% GO terms were identified by both methods. Among these experiments, our method achieved a better performance in identifying real tDEGs with high sensitivity and specificity.

In the survey of Liang et al. [15], they compared the running time between test statistic family, conditional logistical regression family and boosting strategy family, and found that test statistic methods suited for high genes counts analysis as they required the least running time. Our method was based on test statistic with tumor purity information considered, so we just compared it with four test statistic methods, which were paired *t*-test [21], Tan et al.'s method [17], Cao et al.'s method [16], The joint approach of Tan et al.'s and Cao et al.'s method [16,21]. We didn't include conditional logistic regression family, boost strategy family and others traditional feature selection method like minimum-Redundancy-Maximum-Relevancy (mRMR) [30,31], Bayesian method [32] and Entropy method [33], as we believed that the first aim of proposing a new method was to prove its effectiveness when taking tumor purity into consideration. Although there is not enough evidence to conclude that our proposed method has the best performance among all kinds of matched-pairs feature selection methods, our experiments showed it outperformed these test statistic methods without considering tumor purity method.

Tumor purity was an important factor in gene expression analysis. The purity of case samples will bias the difference level estimation between case data and its matched control data, and enlarge the within-group variances for case samples, which lead to biased results and decreased power in the statistical results [14]. It is becoming essential about how to take purity information into gene expression analysis. Petralia et al. proposed a new method, Tumor Specific Net (TSNet), to construct tumor-cell specific gene (or protein) co-expression network based on gene (or protein) expression profiles [34]. TSNet explicitly modeled tumor purity percentage in each tumor samples and treated each observed expression profile as a mixture of expressions from different cell types. Zhang et al. developed a novel differential gene calling method called DECtp by integrating tumor purity information into a generalized least square procedure, followed by the Wald test, which achieved more sensitive, consistent, and biologically meaningful results in the validation experiments from 14 TCGA tumor types [24].

In order to extend our method to other fields, there are two essential conditions: matched-pairs data and purity information. One disadvantage of our method is that it didn't consider the interaction among features, while conditional logical regression model [34] has this

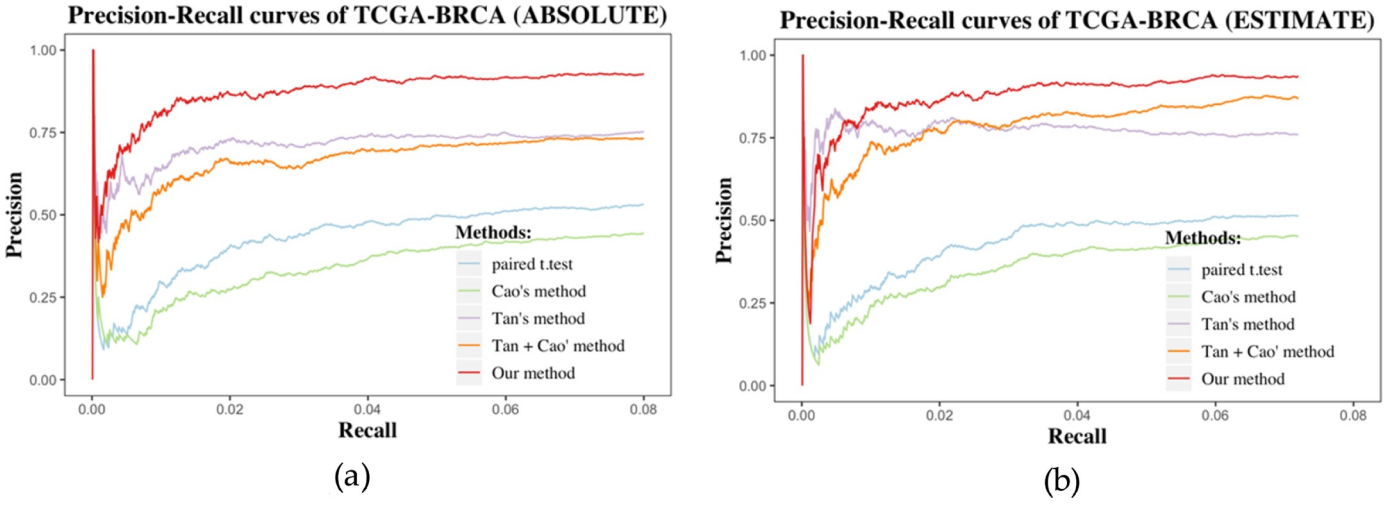

Fig. 6. The P–R curves with tumor purity estimated by others methods for TCGA-BRCA dataset. Figure(a-b) were the P–R curves which purity information estimated by ABSOLUTE and ESTIMATE, respectively. Each method being compared was lined by a different color. *X*-axis represented the number of top N ranked genes, which were selected to calculate the accuracy. *Y*-axis represented the accuracy of each top N genes between ground truth genes and prediction selected genes. Please note that the P–R curves was also fragmentary, as we only considered these tDEGs, where the BRCA and the KIRC datasets have 1257 and 1810 tDEGs, respectively.

Table 4
GO enrichment analysis with pDEGs of BRCA (BP).

Tan + Cao's method GO terms	*p*-value	*p*-value (adjust)	Our method GO terms	*p*-value	*p*-value (adjust)
Extracellular structure organization	2.76e−16	1.45e−12	**Sister chromatid segregation**	**5.78e−18**	**3.04e−14**
Sister chromatid segregation	6.64e−15	1.74e−11	**Mitotic nuclear division**	**1.02e−16**	**2.33e−13**
Extracellular matrix organization	3.26e−14	5.71e−11	**Extracellular structure organization**	**1.33e−16**	**2.33e−13**
Mitotic sister chromatid segregation	3.09e−13	3.41e−10	**Nuclear chromosome segregation**	**2.84e−16**	**3.40e−13**
Chromosome segregation	3.25e−13	3.41e−10	**Chromosome segregation**	**3.23e−16**	**3.40e−13**
Nuclear chromosome segregation	4.48e−13	3.92e−10	**Mitotic sister chromatid segregation**	**5.22e−16**	**4.57e−13**
Ossification	5.36e−13	4.02e−10	**Extracellular matrix organization**	**4.18e−15**	**3.14e−12**
Mitotic nuclear division	9.55e−13	6.26e−10	Positive regulation of cell migration	5.03e−15	3.30e−12
Microtubule cytoskeleton organization involved in mitosis	1.12e−11	6.54e−09	Nuclear division	4.14e−14	2.42e−11
Biomineral tissue development	6.80e−11	3.57e−08	**Microtubule cytoskeleton organization involved in mitosis**	**5.45e−13**	**2.86e−10**

Note: Left column was the GO enrichment analysis result of Tan + Cao's method, and the right column was our method. GO Terms was the name of GO items which enriched by our method, *p*-value was the significant level, adjust p-value was the result with Benjamini Hochberg method. The bold font represented the same GO terms in both methods.

Table 5
GO enrichment analysis with pDEGs of KIRC (BP).

Tan + Cao's method GO terms	*p*-value	*p*-value (adjust)	Our method GO terms	*p*-value	*p*-value (adjust)
Leukocyte migration	2.11e−19	1.17e−15	**T cell activation**	**2.20e−24**	**1.23e−20**
Leukocyte cell-cell adhesion	6.73e−18	1.86e−14	**Leukocyte cell-cell adhesion**	**2.06e−23**	**5.79e−20**
T cell activation	1.62e−17	3.01e−14	**Regulation of cell-cell adhesion**	**1.17e−20**	**2.20e−17**
Cell chemotaxis	9.25e−16	1.28e−12	Regulation of leukocyte cell-cell adhesion	3.04e−20	3.90e−17
Nephron development	4.23e−15	4.69e−12	**Leukocyte migration**	**3.47e−20**	**3.90e−17**
Positive regulation of cell adhesion	5.50e−15	5.09e−12	Regulation of T cell activation	1.97e−18	1.84e−15
Regulation of cell-cell adhesion	8.48e−15	6.72e−12	**Positive regulation of cell adhesion**	**3.15e−17**	**2.52e−14**
Leukocyte chemotaxis	1.44e−14	1.00e−11	Regulation of lymphocyte activation	2.32e−16	1.62e−13
Response to lipopolysaccharide	1.97e−14	1.14e−11	Leukocyte differentiation	5.40e−16	3.08e−13
Kidney development	2.18e−14	1.14e−11	**Positive regulation of cell-cell adhesion**	**5.50e−16**	**3.08e−13**

Note: Left column was the GO enrichment analysis result of Tan + Cao's method, and the right column was our method. GO Terms was the name of GO items which enriched by our method, *p*-value was the significant level, adjust p-value was the result with Benjamini Hochberg method. The bold font represented the same GO terms in both methods.

capability. Balasubramanian et al. proposed a random penalized conditional logistic regression method to conduct feature selection by accounting the two-way interactions among features in high dimensional data setting with matched case-control design [19]. We expect that integrating this characteristic into our method in the future will improve the robustness and effectiveness.

There are some aspects in further developing and applying the matched-pairs feature selection method with purity information. As we know, our method was based on test statistic, which can be modified into one based on conditional logistic regression, and boosting strategy. The current study assumed that the control data was pure, however, control data might be mixed with some impurities in reality. And in gene co-expression networks analysis, we should consider purity information, as Aran et al. found it is problematic when identifying co-expression networks from genomics data without accounting for tumor purity [3]. The same consideration can be taken in biclustering analysis. Li et al. developed a qualitative biclustering algorithm for analyses of gene expression data [35–37], and make an improvement for large-scale bulk RNA-sequencing and single-cell RNA-sequencing data analysis. It may performance better when biclustering analysis with the consideration of tumor purity [38]. In the future we will work towards modeling these circumstances with purity information.

5. Conclusion

In this study, we developed a test statistics based matched-pairs feature selection method for gene expression data analysis. The unique feature of this method is that it integrated purity information with the observation that case data was often a mixture of case data and its matched control data. Our method was showed to be simple, effective and accurate by conducting validation experiments using both simulated and real gene expression datasets.

Funding

This research was funded by the National Natural Science Foundation of China (Nos. 61572227, 61772227), Projects of International Cooperation and Exchanges NSFC (No. 81320108025), and the Development Project of Jilin Province of China (Nos. 20170101006JC, 20190201293JC, 20180414012GH). This work was also supported by Jilin Provincial Key Laboratory of Big Data Intelligent Computing (No. 20180622002JC) .

Acknowledgments

We thank Rongguo Zhang and Xia Chen for providing writing support.

Declarations of interest

None.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mbs.2019.02.007.

References

[1] C. Zhang, W. Cheng, X. Ren, Z. Wang, X. Liu, G. Li, S. Han, T. Jiang, A. Wu, Tumor purity as an underlying key factor in glioma, Clin. Cancer Res. 23 (2017) 6279–6291, https://doi.org/10.1158/1078-0432.ccr-16-2598.

[2] M. Yihao, F. Qingyang, Z. Peng, Y. Liangliang, L. Tianyu, X. Yuqiu, Z. Dexiang, C. Wenju, J. Meiling, T. Yongjiu, R. Li, W. Ye, H. Guodong, X. Jianmin, Tumour purity as a prognostic factor in colon cancer, Biorxiv. (2018) 263723. doi:10.1101/263723.

[3] D. Aran, M. Sirota, A.J. Butte, Systematic pan-cancer analysis of tumour purity, Nat. Commun. 6 (2015) 8971, https://doi.org/10.1038/ncomms9971.

[4] K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P.W. Laird, D.A. Levine, S.L. Carter, G. Getz, K. Stemke-Hale, G.B. Mills, R. Verhaak, Inferring tumour purity and stromal and immune cell admixture from expression data, Nat. Commun. 4 (2013) 2612, https://doi.org/10.1038/ncomms3612.

[5] S.L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P.W. Laird, R.C. Onofrio, W. Winckler, B.A. Weir, Absolute quantification of somatic DNA alterations in human cancer, Nat. Biotechnol. 30 (2012) 413–421, https://doi.org/10.1038/nbt.2203.

[6] L. Oesper, A. Mahmoody, B.J. Raphael, THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data, Genome Biol. 14 (2013) R80, https://doi.org/10.1186/gb-2013-14-7-r80.

[7] H. Chen, J.M. Bell, N.A. Zavala, H.P. Ji, N.R. Zhang, Allele-specific copy number profiling by next-generation DNA sequencing, Nucleic Acids Res. 43 (2015) e23, https://doi.org/10.1093/nar/gku1252.

[8] X. Su, L. Zhang, J. Zhang, F. Meric-Bernstam, J.N. Weinstein, PurityEst: estimating purity of human tumor samples using next-generation sequencing data, Bioinformatics 28 (2012) 2265–2266, https://doi.org/10.1093/bioinformatics/bts365.

[9] Z. Luo, X. Fan, Y. Su, Y.S. Huang, Accurity: accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants, Bioinformatics 34 (2018) 2004–2011, https://doi.org/10.1093/bioinformatics/bty043.

[10] Y. Qin, H. Feng, M. Chen, H. Wu, X. Zheng, InfiniumPurify: an R package for estimating and accounting for tumor purity in cancer methylation research, Genes Dis. 5 (2018) 43–45, https://doi.org/10.1016/j.gendis.2018.02.003.

[11] X. Zheng, Q. Zhao, H.-J. Wu, W. Li, H. Wang, C.A. Meyer, Q. Qin, H. Xu, C. Zang, P. Jiang, F. Li, Y. Hou, J. He, J. Wang, J. Wang, P. Zhang, Y. Zhang, X. Liu, MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes, Genome Biol. 15 (2014) 419, https://doi.org/10.1186/s13059-014-0419-x.

[12] X. Zheng, N. Zhang, H.-J. Wu, H. Wu, Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies, Genome Biol. 18 (2017) 17, https://doi.org/10.1186/s13059-016-1143-5.

[13] N. Zhang, H.-J. Wu, W. Zhang, J. Wang, H. Wu, X. Zheng, Predicting tumor purity from methylation microarray data, Bioinformatics 31 (2015) 3401–3405, https://doi.org/10.1093/bioinformatics/btv370.

[14] F. Wang, N. Zhang, J. Wang, H. Wu, X. Zheng, Tumor purity and differential methylation in cancer epigenomics, Brief. Funct. Genom. (2016) elw016, https://doi.org/10.1093/bfgp/elw016.

[15] S. Liang, A. Ma, S. Yang, Y. Wang, Q. Ma, A Review of matched-pairs feature selection methods for gene expression data analysis, Comput. Struct. Biotechnol. J. 16 (2018) 88–97, https://doi.org/10.1016/j.csbj.2018.02.005.

[16] Z. Cao, Y. Wang, Y. Sun, W. Du, Y. Liang, A novel filter feature selection method for paired microarray expression data analysis, Int. J. Data Min. Bioinform. 12 (2015) 363–386, https://doi.org/10.1504/ijdmb.2015.070071.

[17] Q. Tan, M. Thomassen, T.A. Kruse, Feature selection for predicting tumor metastases in microarray experiments using paired design, Cancer Inform. 3 (2007) 213–218.

[18] J. Asafu-Adjei, M.G. Tadesse, B. Coull, R. Balasubramanian, M. Lev, L. Schwamm, R. Betensky, Bayesian Variable selection methods for matched case-control studies, The International Journal of Biostatistics 13 (2017) 20160043, , https://doi.org/10.1515/ijb-2016-0043.

[19] R. Balasubramanian, A.E. Houseman, B.A. Coull, M.H. Lev, L.H. Schwamm, R.A. Betensky, Variable importance in matched case-control studies in settings of high dimensional data, J. R. Stat. Soc. 63 (2014) 639–655, https://doi.org/10.1111/rssc.12056.

[20] A.J. Adewale, I. Dinu, Y. Yasui, Boosting for correlated binary classification, J. Comput. Graph. Stat. 19 (2010) 140–153, https://doi.org/10.1198/jcgs.2009.07118.

[21] X. Cui, G.A. Churchill, Statistical tests for differential expression in cDNA microarray experiments, Genome Biol. 4 (2003) 210, https://doi.org/10.1186/gb-2003-4-4-210.

[22] N. Wang, T. Gong, R. Clarke, L. Chen, I.-M. Shih, Z. Zhang, D.A. Levine, J. Xuan, Y. Wang, UNDO: a bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples, Bioinformatics 31 (2015) 137–139, https://doi.org/10.1093/bioinformatics/btu607.

[23] Q. Shen, J. Hu, N. Jiang, X. Hu, Z. Luo, H. Zhang, contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples, Bioinformatics 32 (2016) 705–712, https://doi.org/10.1093/bioinformatics/btv657.

[24] W. Zhang, H. Long, B. He, J. Yang, DECtp: calling differential gene expression between cancer and normal samples by integrating tumor purity information, Front. Genet. 9 (2018) 321, https://doi.org/10.3389/fgene.2018.00321.

[25] T. Network, J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. uart, The cancer genome atlas pan-cancer analysis project, Nat Genet. 45 (2013) ng.2764, https://doi.org/10.1038/ng.2764.

[26] K. Tomczak, P. Czerwińska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, Contemp. Oncol. 19 (2015) A68–A77, https://doi.org/10.5114/wo.2014.47136.

[27] T. Gene Ontology Consortium, Expansion of the Gene Ontology knowledgebase and resources, Nucleic Acids Res. 45 (2017) D331–D338, https://doi.org/10.1093/nar/gkw1108.

[28] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, M.J. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29, https://doi.org/10.1038/75556.

[29] J. Ferreira, A. Zwinderman, On the Benjamini–Hochberg method, Ann. Stat. 34 (2006) 1827–1849, https://doi.org/10.1214/009053606000000425.

[30] M.A. Connolly, K.-Y. Liang, Conditional logistic regression models for correlated binary data, Biometrika 75 (1988) 501–506, https://doi.org/10.2307/2336600.

[31] F. Yuan, L. Lu, Y. Zhang, S. Wang, Y.-D. Cai, Data mining of the cancer-related lncRNAs GO terms and KEGG pathways by using mRMR method, Math. Biosci. 304 (2018) 1–8, https://doi.org/10.1016/j.mbs.2018.08.001.

[32] A. Bhattacharjee, G.K. Vishwakarma, A. Thomas, Bayesian state-space modeling in gene expression data analysis: an application with biomarker prediction, Math. Biosci. 305 (2018) 96–101, https://doi.org/10.1016/j.mbs.2018.08.011.

[33] Y.-W. Niu, H. Liu, G.-H. Wang, G.-Y. Yan, Maximal entropy random walk on heterogenous network for MIRNA-disease Association prediction, Math. Biosci. 306 (2018) 1–9, https://doi.org/10.1016/j.mbs.2018.10.004.

[34] F. Petralia, L. Wang, J. Peng, A. Yan, J. Zhu, P. Wang, A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity, Bioinformatics 34 (2018) i528–i536, https://doi.org/10.1093/bioinformatics/bty280.

[35] Y. Zhang, J. Xie, J. Yang, A. Fennell, C. Zhang, Q. Ma, QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data, Bioinformatics 33 (2016) btw635, https://doi.org/10.1093/bioinformatics/btw635.

[36] G. Li, Q. Ma, H. Tang, A.H. Paterson, Y. Xu, QUBIC: a qualitative biclustering algorithm for analyses of gene expression data, Nucleic Acids Res. 37 (2009) e101, https://doi.org/10.1093/nar/gkp491.

[37] F. Zhou, Q. Ma, G. Li, Y. Xu, QServer: a biclustering server for prediction and assessment of co-expressed gene clusters, PLoS One 7 (2012) e32660, https://doi.org/10.1371/journal.pone.0032660.

[38] J. Xie, A. Ma, Y. Zhang, B. Liu, C. Wang, S. Cao, C. Zhang, Q. Ma, QUBIC2: a novel biclustering algorithm for large-scale bulk RNA-sequencing and single-cell RNA-sequencing data analysis, Biorxiv. (2018) 409961. doi:10.1101/409961.