Neurocomputing 269 (2017) 40–46

Contents lists available at ScienceDirect

Neurocomputing

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

Bootstrapping integrative hypothesis test for identifying biomarkers that differentiates lung cancer and chronic obstructive pulmonary disease

Kai-Ming Jiang[a,c], Ya-Jing Chen[a,d], Jin-Xiong Lv[a,d], Bao-Liang Lu[a,c], Lei Xu[b,d,*]

[a] Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China
[b] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China
[c] The Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, 800 Dong Chuan Road, Shanghai 200240, China
[d] Centre for Cognitive Machines and Computational Health (CMaCH), The School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China

ARTICLE INFO

Article history:
Received 3 December 2015
Revised 7 October 2016
Accepted 31 October 2016
Available online 3 June 2017

Keywords:
Integrative hypothesis test
Bootstrapping
Differential gene expression
Rank reliability

ABSTRACT

Different from the common approaches that use either hypothesis test or classifier for biomarker discovery, we applied the integrative hypothesis test (IHT) that combined both to identifying miRNAs for differentiation between lung cancer and Chronic Obstructive Pulmonary Disease (shortly L-C differentiation) on GEO data set GSE24709, and further extended IHT implementation by bootstrapping aided ranking and mean-variance based reliability check, which outputs a list of the top-15 differentially expressed miRNAs that confirmed the previously reported 14 miRNAs for L-C differentiation from a very different perspective plus an additional one. Moreover, we conducted a literature survey for a further explanation via dividing the 15 miRNAs into subclasses based on known relevances to the two diseases. Also, every pair of 15 miRNAs is exhaustively examined on their joint effect via *p*-value, misclassification, and correlation, which identifies core pairs and linked cliques as joint miRNAs biomarkers.

1. Introduction

Lung cancer is closely related to Chronic Obstructive Pulmonary Disease (COPD), a common pulmonary affliction encompassing chronic obstructive bronchitis and lung emphysema [1]. COPD is a global burden affecting 10.15% of adults older than 40 years [2] and precedes lung cancer in 50.90% of cases [3]. Based on the differential expression of miRNAs in tumors, efforts have been made on finding miRNA expression signatures of lung cancer and subtypes via not only tumor cells [4,5] but also sera and peripheral blood cells from cancer patients [6,7,8,9].

Differentiation analyses on miRNA expression and gene expression are made in one of two typical methods that are generally used in various tasks of case-control studies or binary classification. Under the name of two sample test or model comparison in general, the first method evaluates the overall difference between two populations of samples with each population described by a parametric model, usually a normal distribution. One widely used example on gene expression differentiation is t-test and Welch test. Under the name of classification or model prediction, the second method evaluates the performance of discriminative boundary that classifies each sample into its corresponding population. Each of the two methods has been extensively studied individually.

Previous studies tend to merely use one of the two methods to identify biomarkers, though some study may also compute the measure of the other for a reference or a double check. According to our knowledge, there lack efforts on systematically integrating both of them to jointly identify biomarkers. Integrative Hypothesis Test (IHT) has been recently proposed to suit this purpose [10], [11]. The data were downloaded from Gene Expression Omnibus data set GSE24709 (GEO, http://www.ncbi.nlm.nih.gov/geo/, GSE24709) [3,12], which contained the expression data in blood cells of 863 miRNAs for lung cancer patients and patients suffering from COPD. We further extended IHT implementation by bootstrapping aided ranking (shortly called bootstrapping-IHT) and mean-variance based reliability check, resulting in a so-called IHT rank of miRNA biomarkers for distinguishing lung cancer and COPD. We obtained a list of top-15 miRNAs (See Table 2) that covered all the 14 differentially expressed ones identified

* Corresponding author at: Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China.
E-mail address: lxu@cse.cuhk.edu.hk (L. Xu).

http://dx.doi.org/10.1016/j.neucom.2016.10.092
0925-2312/

Table 1
A list of top-15 obtained by IHT.

Rank	Gene id	*p*-value	Accuracy (%)
1	hsa-miR-369-5p	1.24E−05	81.32
2	hsa-miR-675	2E-05	77.78
3	hsa-miR-662	9.25E−06	76.88
4	hsa-miR-641	4.86E−05	76.77
5	hsa-miR-767-3p	4.55E−06	76.46
6	hsa-miR-888*	0.00076	78.47
7	hsa-miR-26a	1.55E−06	75.59
8	hsa-miR-1299	0.000567	75.45
9	hsa-miR-95	7.46E−05	74.20
10	hsa-miR-636	0.000192	73.68
11	hsa-miR-1308	0.000196	72.95
12	hsa-miR-513b	0.000366	72.15
13	hsa-miR-668	0.000934	72.92
14	hsa-miR-130b	0.000349	71.18
15	hsa-miR-875-3p	0.001364	74.41

Table 2
A list of top-15 obtained with rank bootstrapping.

Rank	Gene ID	Avg. rank	Std. rank
1	hsa-miR-662	2.8	1.30384
2	hsa-miR-636	3	1.224745
3	hsa-miR-675	3.4	1.516575
4	hsa-miR-369-5p	4.2	4.494441
5	hsa-miR-940	7.6	3.361547
6	hsa-miR-92a	8.4	5.029911
7	hsa-miR-1224-3p	8.6	4.505552
8	hsa-miR-26a	10.6	4.615192
9	hsa-miR-328	11.2	5.80517
10	hsa-miR-641	14.2	3.701351
11	hsa-miR-383	17	5.43139
12	hsa-let-7d*	21.2	4.086563
13	hsa-miR-93*	24	10.90871
14	hsa-miR-323-3p	24.8	6.220932
15	hsa-miR-513b	26.6	4.27785
Excluded genes due to large rank std.			
	hsa-miR-875-3p	20.2	17.32628
	hsa-miR-30e*	22	13.32291
	hsa-miR-139-5p	22.6	13.01153
	hsa-miR-1911	23.8	12.43785
	hsa-miR-130b	24	15.23155

in [12] (See Table 3) plus an additional one. Further literature survey divided the list into a subclass of 5 miRNAs that were validated by qRT-PCR for the separation between lung cancer and COPD, a subclass of 3 miRNAs that were reported to be related to lung diseases, while the roles of the rest 7 ones in lung diseases remained unclear. The subclass containing the 7 miRNAs were further divided into two subgroups by whether they were relevant to other cancers. (See Table 4)

Additionally, we performed pathway analysis by TarBase [13] on the target mRNAs of the top-15 miRNAs listed in Table 2. The target genes of hsa-miR-26a-5p (previously, hsa-miR-26a) and hsa-miR-92a-3p (previously, hsa-miR-92a) were identified to be associated with pathways closely related to cancer.

Moreover, we enumerated every pair of miRNAs to compute the corresponding p-values, misclassification rates and correlation coefficients, and obtained a list of top-20 pairs of miRNAs (See Table 7). Interestingly, 19 out of them contained one miRNA that ranked top-10 in Table 2. But the top-ranking pairs did not necessarily include both top-ranking miRNAs.

Within the top-20 pairs shown in Table 7, 16 pairs were uncorrelated, with correlation coefficients $|r| < 0.3$; 3 pairs were weakly correlated, with correlation coefficients $0.3 \leq |r| < 0.5$; 1 pair was moderately correlated, with correlation coefficient $|r| \geq 0.5$. We can see that most of the top-20 pairs were weakly correlated or even uncorrelated. Moreover, these pairs displayed high differentiation performance with high classification accuracies (89–95%) and low *p*-values (around $10^{-7} - 10^{-10}$). Three distinct joint miRNA signatures were generated, the core miRNAs of which were hsa-miR-675, hsa-miR-369-5p and hsa-miR-92a, respectively. These three signatures were further linked by the pair of hsa-miR-369-5p and hsa-miR-92a, and the pair of hsa-miR-369-5p and hsa-miR-675.

2. Methods

2.1. Integrative hypothesis tests

Integrative hypothesis tests (IHT) was previously advocated in [10,11] for an integrative study of case-control problems. It is featured with two perspectives (namely, model-based versus boundary-based), and four different tasks (namely, modelling, comparison, classification and assurance [11,14]). To facilitate the explanation, we define the case and control samples as $X_\omega = \{x_{t,\omega}, t = 1, \ldots, N_\omega\}$, $\omega = 0, 1$ with $\omega = 0$ for control and $\omega = 1$ for case, where N_ω is the number of samples. We assume that the case samples were generated by a parametric model $q(x|\theta_1)$, while the control samples were generated by $q(x|\theta_0)$.

From the model-based perspective, the main focus is to test the null hypothesis H_0 (i.e., no difference between $q(x|\theta_0)$ and $q(x|\theta_1)$). First, we need to find two parameters θ_0 and θ_1 that fit the two populations well, i.e., performing the task of modelling. Then, statistics should be proposed to measure the difference between the two models, i.e., performing the task of comparison. From the boundary-based perspective, we consider whether samples are well separated by a separating boundary of two populations. Considering a linear separation on two populations, the performance is featured with either the distances of samples to the linear separating boundary or the differences between the projection values of two populations along the normal direction w of the linear boundary. First, we need to find the best boundary to separate samples of two populations, namely, performing the task of classification. Then, statistics should be proposed to test whether samples are well separated or whether the resulted boundary breaks the null hypothesis H_0: $w = 0$ significantly, namely, performing the task of assurance.

Each of four tasks has been studied separately in the existing efforts, having its strengths and limitations. However, the performances of these tasks are coupled, and also the best set of features for one task are not necessarily be the best for the others. It naturally motivates that better results might be generated if the performance of all the four tasks can be jointly optimized. The necessity and feasibility of the joint consideration have been addressed in details in [11].

In this paper, we focus on combining the two commonly encountered ones for biomarker discoveries, i.e., model comparison testing and sample classification. For making comparison, we aimed at testing whether the difference between lung cancer and COPD populations were significant. We evaluated each miRNA by Welch's t-test and each miRNA pair by the Hotelling T-squared test, with the expressions of every miRNA pair modelled by bivariate normal distributions, where θ consists of its mean μ and covariance matrix Σ, i.e., $\theta = \{\mu, \Sigma\}$. The testing result is featured by the corresponding *p*-value.

For classification, we first used Fisher's discriminant analysis (FDA) to find the best direction w and projected the expressions of every miRNA pair onto w to get a scalar expression. This best direction w was obtained by maximising the between-class variance $\sigma^2_{between}$ and minimising the within-class variance $\sigma^2_{between}$, i.e.,

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(w^T\mu_1 - w^T\mu_0)^2}{w^T\Sigma_1 w + w^T\Sigma_0 w}, \quad (1)$$

where μ_1 and Σ_1 are the mean and covariance matrix for lung cancer samples, μ_0 and Σ_0 are the mean and covariance matrix

Table 3
Significant markers identified in [12] for differentiation of lung cancer versus COPD (p-value < 0.01).

miRNA	Control	COPD	Lung cancer	Control vs. COPD	Control vs. lung cancer	Lung cancer vs. COPD
hsa-miR-641	76.68	143.15	59.58	0.00013	0.90088	0.00075
hsa-miR-662	90.65	23.1	95.46	0.0003	0.5175	0.0001
hsa-miR-369-5p	33.46	97.1	33.25	0.00041	0.60298	0.0001
hsa-miR-383	74.96	142.06	73.83	0.00122	0.87052	0.00316
hsa-miR-636	246.59	106.39	222.87	0.00186	0.72712	0.00016
hsa-miR-940	225.92	152.89	247.83	0.00583	0.94678	0.00683
hsa-miR-26a	7269.84	7975.44	5568.45	0.00931	0.21746	0.00047
hsa-miR-92a	13651.44	9554.17	13651.44	0.00957	0.80809	0.00156
hsa-miR-328	59.92	76.93	208.31	0.96379	0.00428	0.00126
hsa-let-7d*	70.76	102.75	250.42	0.05763	0.00006	0.00278
hsa-miR-1224-3p	137.63	109.61	233.37	0.08731	0.86406	0.00316
hsa-miR-513b	66.76	80.41	39.04	0.03264	0.12765	0.00411
hsa-miR-93*	893.5	1303.7	2321.35	0.99299	0.01562	0.0068
hsa-miR-675	254.2	149.11	287.83	0.04421	0.04842	0.00156

Table 4
Literature survey results for top-15 miRNAs.

Subclass	miRNA	Description
I	hsa-miR-26a	Validated by qRT-PCR for the separation between COPD and NSCLC [15]
	hsa-miR-328	Validated by qRT-PCR for the separation between COPD and NSCLC [15]
	hsa-miR-93*	Validated by qRT-PCR for the separation between COPD and NSCLC [15]
	hsa-miR-1224-3p	Validated by qRT-PCR for the separation between COPD and NSCLC [15]
	hsa-miR-383	Validated by qRT-PCR for the separation between COPD and NSCLC [15]
II	hsa-miR-675	Down-regulation leads to tumor progression in non-small cell lung cancer [16]
		Significantly dysregulation in the small airway epithelium despite smoking cessation [17]
	hsa-miR-636	Variant miRNAs in control vs. COPD [18]
	hsa-miR-940	Associated with regulation of ASM actin cytoskeleton [19]
IIIa	hsa-miR-92a	Increased expression in ALL patients [20]
	hsa-let-7d*	Contribute to cross-targeting of cancer-related factors [21]
	hsa-miR-323-3p	Play a role in the transformation from oral leukoplakia to cancer [22]
	hsa-miR-662	Part of signature for predicting prognosis for neural GBM [23]
IIIb	hsa-miR-369-5p	
	hsa-miR-641	
	hsa-miR-513b	

Table 5
Significant pathways of the target genes of miR-26a-5p.

Pathway	p-values	Bonferroni
hsa04110:cell cycle	3.90E−07	6.94E−05
hsa05200:pathways in cancer	2.93E−06	5.22E−04
hsa05210:colorectal cancer	3.97E−05	0.007042
hsa04115:p53 signaling pathway	7.82E−05	0.013817
hsa04114:oocyte meiosis	9.56E−05	0.01688
hsa04310:Wnt signaling pathway	1.60E−04	0.028087
hsa04350:TGF-beta signaling pathway	2.12E−04	0.037095
hsa05220:chronic myeloid leukemia	2.84E−04	0.049303

Table 6
Significant pathways of the target genes of miR-92a-3p.

Pathway	p-values	Bonferroni
hsa04110:cell cycle	1.17E−09	2.12E−07
hsa05200:pathways in cancer	7.38E−07	1.34E−04
hsa03010:ribosome	9.71E−06	0.001756
hsa05215:prostate cancer	1.44E−05	0.002601
hsa05220:chronic myeloid leukemia	1.31E−04	0.023382
hsa04510:focal adhesion	1.81E−04	0.032187
hsa04520:adherens junction	1.87E−04	0.033332
hsa05222:small cell lung cancer	5.87E−04	0.100802

for COPD samples. In the case of testing one miRNA, no projection was needed because we already dealt with a scalar expression. Next, the misclassification rates were computed on scalar expressions according to the separating boundary point that is simply the middle of the mean values of the lung cancer and COPD patients.

We may further observe the joint performances of p-values and misclassification rates with help of the scatterplot called the IHT plot (See Fig. 1), where each point represents one miRNA. The x-axis denotes the p-value while the y-axis the misclassification rate. A small p-value indicated significant difference between two populations and a small misclassification rate indicated a good classification of samples by its separating boundary. Thus, we prefer the candidate points that were closest to the origin of the coordinate space. However, as addressed in [11], this IHT plot is limited to merely integrating two measures and also encounters a challenge of how to appropriately scaling each measure.

2.2. Bootstrapping-IHT

We further turned the IHT plot shown in Fig. 1 into a list of miRNAs sorted according to their distances to the origin of the coordinate space. Table 1 showed the resulted top-15 miRNAs, with the p-values and accuracies obtained by the Welch's t-test and Fisher's discriminant analysis respectively. Because of the small sample size ($N_1 = 28$ for lung cancer patients and $N_0 = 24$ for COPD), the resulted p-values and accuracies are actually random variables, namely, each point in Fig. 1 is actually random dot, which makes the IHT ranks in Table 1 unreliable.

To tackle the small sample size, we applied the bootstrapping strategy. Bootstrapping is a common practice to estimate properties when the sample size is insufficient. It measures the properties via sampling with replacement, which makes the estimation of almost any statistics possible. Also, bootstrapping provides a way to account for the distortions caused by specific samples that may not be fully representative of the population.

In each bootstrapping implementation, we obtained the corresponding initial IHT ranks for all miRNAs by integrating p-values and classification accuracies, after which the specific values of

Table 7
A list of top-20 pairs obtained by IHT (Pearson Correlation did not join the sorting).

Gene ID1	Gene ID2	P value	Accuracy	Pearson
hsa-miR-1271	hsa-miR-675	5.74E-10	0.948958	-0.19585
hsa-miR-369-5p	hsa-miR-92a	3.19E-09	0.930903	-0.42127
hsa-miR-641	hsa-miR-675	8.1E-09	0.929514	-0.21791
hsa-miR-1204	hsa-miR-92a	8.74E-08	0.91875	0.013134
hsa-miR-1302	hsa-miR-662	4E-08	0.917014	-0.00473
hsa-miR-369-5p	hsa-miR-675	2.05E-10	0.915972	-0.30058
hsa-miR-1299	hsa-miR-675	2.17E-06	0.9125	-0.03089
hsa-miR-627	hsa-miR-675	4.21E-08	0.909375	-0.02967
hsa-miR-610	hsa-miR-662	5.5E-11	0.907639	0.114439
hsa-miR-26a	hsa-miR-369-5p	4.11E-09	0.905208	0.524418
hsa-miR-376a*	hsa-miR-92a	3.5E-08	0.904167	-0.11793
hsa-miR-767-3p	hsa-miR-92a	1.23E-07	0.900694	-0.16173
hsa-miR-183	hsa-miR-369-5p	7.83E-10	0.899306	0.211596
hsa-let-7d*	hsa-miR-1226*	9.54E-09	0.898958	-0.27561
hsa-miR-636	hsa-miR-888*	2.33E-07	0.897222	-0.28287
hsa-miR-92a	hsa-miR-93*	2.51E-07	0.896181	0.212414
hsa-miR-489	hsa-miR-675	1.99E-08	0.895139	-0.29558
hsa-miR-1248	hsa-miR-641	9.83E-08	0.893403	-0.07723
hsa-miR-875-5p	hsa-miR-92a	7.4E-07	0.893403	0.003946
hsa-miR-369-5p	hsa-miR-940	4.88E-09	0.892014	-0.35694

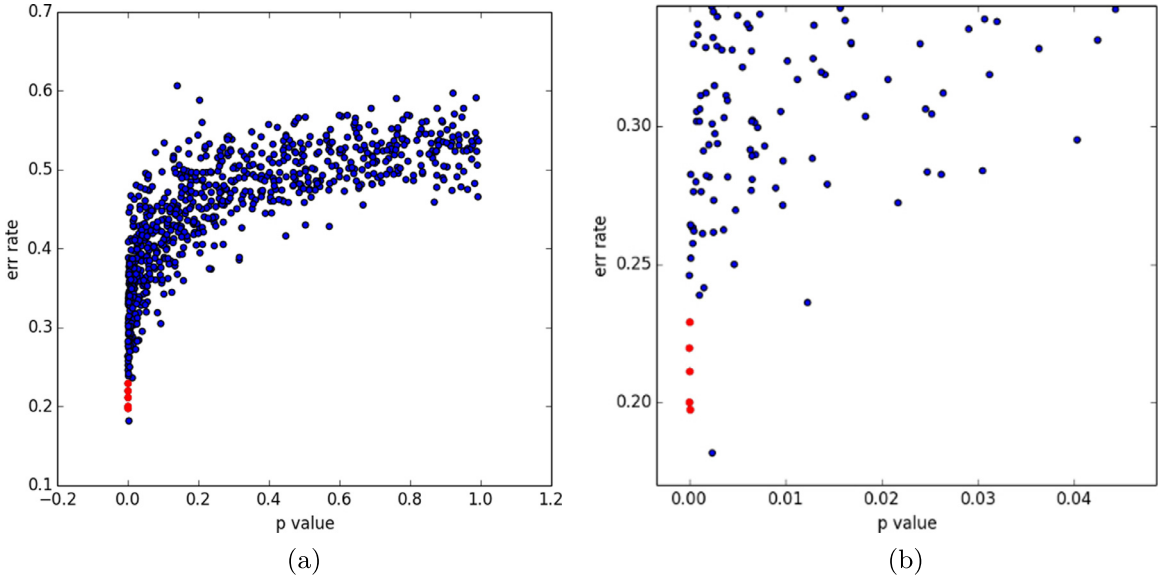

Fig. 1. IHT plot(a) overall view, (b) zoom in view.

p-values and classification accuracies were no longer important. What we focus on are these initial IHT ranks that randomly vary every time we perform bootstrapping. Then, we reorder all miRNAs according to the means of random initial IHT ranks and check the corresponding variances for reliability. One miRNA with a small mean and variance in its IHT ranks is regarded as stably significant for differentiation. As a result, we obtain a list of final IHT ranks after discarding those miRNAs with the corresponding variances bigger than a pre-specified threshold.

Empirically, illustrated in Fig. 2 is the relationship between the mean and standard deviation of IHT ranks. As the mean ranks increase, we observed that the standard deviation displays an overall growing tendency though this tendency is not monotonic. We are thus motivated to choose the threshold at the point where the standard deviation suddenly changed to a great amount so that a spike appeared in Fig. 2.

Similarly, we may further extend such a bootstrapping-IHT to examine every pair of miRNAs.

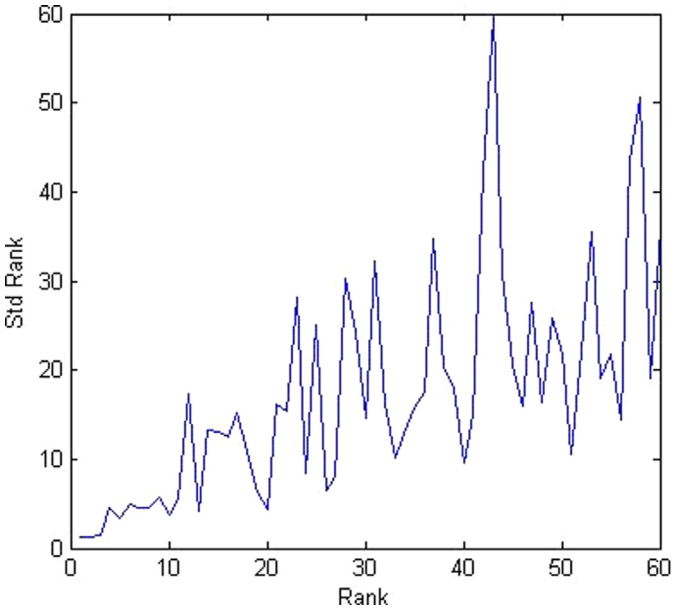

Fig. 2. Selection of a threshold for standard deviation.

3. Empirical studies

3.1. Top-15 miRNAs identified

The Microarray data used in this paper were downloaded from the Gene Expression Omnibus profile GSE24709 (GEO, http://www.ncbi.nlm.nih.gov/geo/, GSE24709). It includes 863 human miRNAs annotated in miRBase version 12.0. The data were the miRNA expression values in blood cells for 71 individuals, including 28 lung cancer patients, 24 COPD patients, and 19 healthy controls. In this study, we only considered the COPD patients and lung cancer patients.

Table 1 shows the results obtained by evaluating each miRNA merely based on the original samples, while Table 2 shows the results obtained by bootstrapping. In each bootstrapping implementation, we resampled and obtained a set of 45 samples. We set the number of bootstrapping implementation to be 100, as it reached a balance of efficient computation and stable results. After 100 bootstrapping implementations, we computed the mean and standard deviation of the IHT ranks for each miRNA. We listed the top-15 miRNAs in Table 2, sorting increasingly according to their mean IHT ranks but discarding ones bigger than a threshold of 12 for standard deviation that is chosen from Fig. 2. Some discarded miRNAs with relatively smaller mean IHT ranks but big standard deviation were also shown in Table 2.

Fourteen of the top-15 miRNAs listed in Table 2 were also included in Table 3 that consists of the significant miRNAs for the differentiation between lung cancer and COPD found in [12]. In other words, the findings of [12] are confirmed here independently by a very different method, which may enhance attentions to these findings.

To get a further understanding on the top-15 miRNAs shown in Table 2, we also conducted literature survey and divided the 15 miRNAs into 3 subclasses (Table 4). Subclass I contains 5 miRNAs that were validated by qRT-PCR to be differentially expressed between COPD and NSCLC. Subclass II contains 3 miRNAs that were known to be related to lung diseases, including lung cancer, COPD, etc. The remaining 7 miRNAs were not reported to have relevance to lung diseases, forming subclass III. But we could find some cancer-related reports for 4 out of the 7 miRNAs, which were further classified into subclass III a, while the rest 3 with few reports were classified into subclass III b.

3.2. Functional analysis

We also performed the pathway analysis on the target genes of the top-15 miRNAs listed in Table 2. All the target genes were obtained by TarBase [13]. A biological pathway is a series of molecular actions that produce substances or cause changes in a cell. Investigating the relationship between target mRNAs and pathways provides knowledge about the potential biological functions of these miRNAs.

Tables 5 and 6 respectively demonstrated the pathway analysis results of hsa-miR-26a-5p (another name of hsa-miR-26a) and hsa-miR-92a-3p (another name of hsa-miR-92a) obtained from DAVID [24,25], where the relevant pathways were sorted increasingly according to the p-values corrected by Bonferroni method. Smaller p-values means more significant relevant with the pathways.

Both the two strongly correlate to two cancer-related factors, namely cell cycle and pathways in cancer. Moreover, Hsa-miR-26a-5p were related to p53 signaling, Wnt signaling and TGF-beta signaling. p53 signaling is important for the pathogenesis of cancer [26]. Wnt signalling influences the growth of tumor, and thus can serve as a therapeutic target [27]. TGF-beta signalling can determine the behaviour of cancer cell and have an effect on cancer progression [28]. Hsa-miR-92a-3p were relevant to focal adhesion and adherens junction. Focal adhesion plays a role in tumor formation [29]. Adherens junctions have significant changes in cancer cells compared to normal cells [30].

These results indicated that the miRNAs found by bootstrapping-IHT were actively involved in the tumor development process by regulating the expression of their target genes.

3.3. Joint miRNA analysis

We further proceeded to study two or more miRNAs jointly. First, we applied hierarchical clustering on the expression data of the miRNAs listed in Table 2. The clustering result was visualised in a heat map shown in Fig. 3. We can see that the miRNAs are clustered into two groups, including the one that contains hsa-miR-641, hsa-miR-26a, hsa-miR-369-5p, hsa-miR-875-3p, hsa-miR-383, and the other that contains hsa-miR-139-5p, hsa-miR-662, hsa-miR-30e*, hsa-let-7d*, hsa-miR-328, hsa-miR-636, hsa-miR-675, hsa-miR-940, hsa-miR-1224-3p and hsa-miR-92a. The two groups show different expression patterns: miRNAs in the first group up-regulate in COPD (red) and down-regulate in lung cancer (green), while miRNAs in the second group show an opposite pattern.

Also, we examined every pair of miRNAs by Pearson correlation, and evaluated them by boostrapping-IHT which generated the IHT ranks that reflected the integrated performance of the p-value and classification accuracy. Table 7 gives a list of top-20 pairs of miRNAs sorted by IHT ranks. The Pearson correlation coefficients between the two miRNAs are listed in the last column, but they did not join the sorting process. We see that the top-ranking miRNA pairs did not necessarily contain both of the top-ranking individual miRNAs.

Interestingly, 19 out of the 20 pairs contained one miRNA that ranked top-10 in Table 2. Especially, hsa-miR-675 and hsa-miR-92a each appeared in 6 pairs; hsa-miR-369-5p appeared in 5 pairs; hsa-miR-641 and hsa-miR-662 appeared in 2 pairs. Besides, hsa-miR-26a, hsa-let-7d*, hsa-miR-636, hsa-miR-93* also appeared in both Tables 2 and 7. As these 9 miRNAs appeared in both Tables 2 and 7, they drew our particular attention for further investigation.

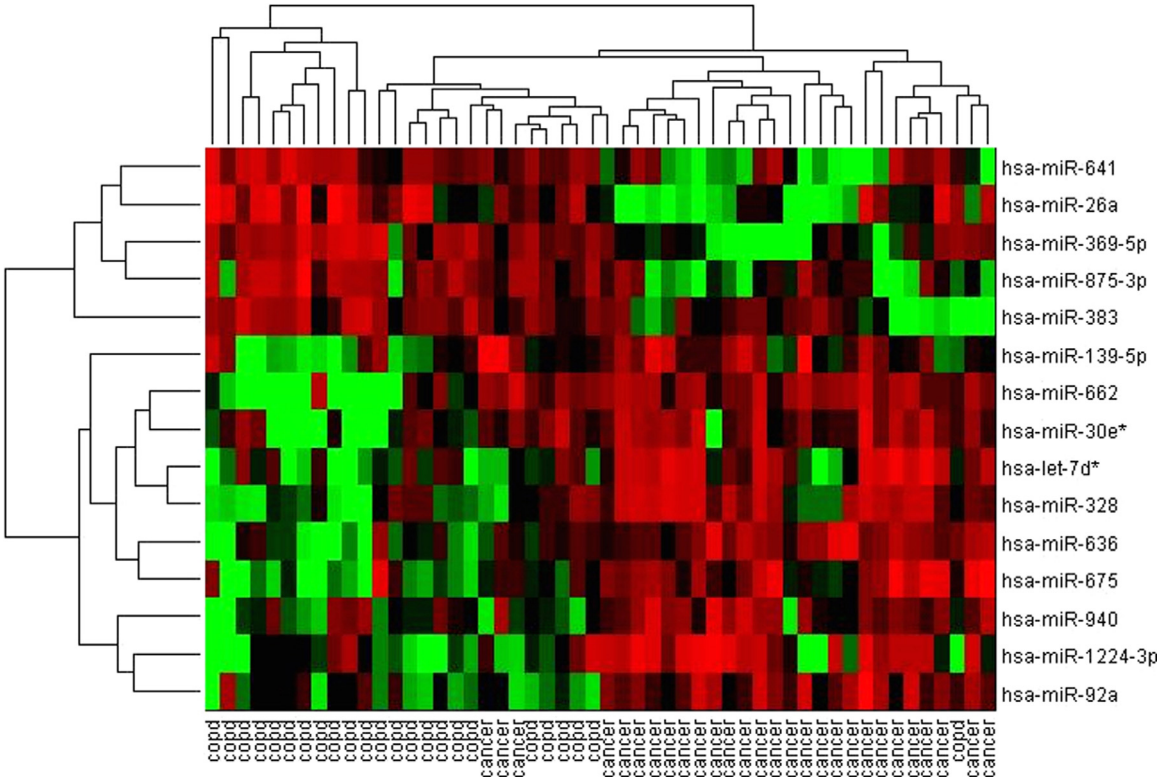

Fig. 3. Hierarchical clustering result of top-15 miRNAs. Red color represents up-regulation, while green color represents down-regulation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Within the top-20 pairs shown in Table 7, 16 pairs were uncorrelated, with correlation coefficients $|r| < 0.3$; 3 pairs were weakly correlated, with correlation coefficients $0.3 \leq |r| < 0.5$; 1 pair was moderately correlated, with correlation coefficient $|r| \geq 0.5$. We can see that most of the top-20 pairs were weakly correlated or even uncorrelated. Moreover, these pairs displayed high differentiation performance with high classification accuracies (89–95%) and low p-values (around $10^{-7} - 10^{-10}$). Three distinct joint miRNA signatures were generated (marked by blue, green and yellow), the core miRNAs of which were hsa-miR-675, hsa-miR-369-5p and hsa-miR-92a, respectively. These three signatures were further linked by the pair of hsa-miR-369-5p and hsa-miR-92a, and the pair of hsa-miR-369-5p and hsa-miR-675.

3.4. Some remarks

Machine learning and statistical methods have been widely used in medicine and biology related fields [3,4,5,12,31]. We used a bootstrapping integrative hypothesis test to study the data set GSE24709 of miRNA expressions in blood cells. We may further extend this bootstrapping-IHT study to analysing data set that contain both tumour and adjacent normal expression data, with the help of bilinear matrix-variate analysis [11]. Also, the S-space boundary-based test developed in [14] may be adopted for joint miRNA analysis. Besides, most of the existing studies, including this one, do not consider conditions or environment. The E-GPS theory may also be applied to carry out analysis conditioned on certain environments [32].

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61272248), the National Basic Research Program of China (Grant No. 2013CB329401), the Science and Technology Commission of Shanghai Municipality (Grant No.13511500200), and Shanghai Jiao Tong University start-up fund WF220103010 for Zhiyuan Chair Professorship.

References

[1] Y.R. van Gestel, S.E. Hoeks, D.D. Sin, V. Hzeir, H. Stam, F.W. Mertens, R.T. van Domburg, J.J. Bax, D. Poldermans, COPD and cancer mortality: the influence of statins, Thorax 64 (11) (2009) 963–967.
[2] R.P. Young, R.J. Hopkins, T. Christmas, P.N. Black, P. Metcalf, G. Gamble, COPD prevalence is increased in lung cancer, independent of age, sex and smoking history, Eur. Respir. J. 34 (2) (2009) 380–386.
[3] P. Leidinger, A. Keller, Peripheral profiles from patients with cancerous and non cancerous lung diseases, Lung Cancer 74 (1) (2011 Oct) 41–47 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSE24709.
[4] I. Barshack, G. Lithwick-Yanai, A. Afek, K. Rosenblatt, H. Tabibian-Keissar, M. Zepeniuk, L. Cohen, H. Dan, O. Zion, Y. Strenov, MicroRNA expression differentiates between primary lung tumors and metastases to the lung, Pathol. Res. Pract. 206 (8) (2010) 578–584.
[5] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet–Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, MicroRNA expression profiles classify human cancers, Nature 435 (7043) (2005) 834–838.
[6] M.A. Cortez, G.A. Calin, MicroRNA identification in plasma and serum: a new tool to diagnose and monitor diseases, Expert Opin. Biol. Ther. 9 (6) (2009 Jun) 703–711.
[7] S. Gilad, E. Meiri, Y. Yogev, S. Benjamin, D. Lebanony, N. Yerushalmi, H. Benjamin, M. Kushnir, H. Cholakh, N. Melamed, Serum microRNAs are promising novel biomarkers, PloS ONE 3 (9) (2008) e3148.
[8] C.M. Tammemagi, C. NeslundDudas, M. Simoff, P. Kvale, Impact of comorbidity on lung cancer survival, Int. J. Cancer 103 (6) (2003) 792–802.
[9] J. Wang, J. Chen, P. Chang, A. LeBlanc, D. Li, J.L. Abbruzzesse, M.L. Frazier, A.M. Killary, S. Sen, MicroRNAs in plasma of pancreatic ductal adenocarcinoma patients as novel blood-based biomarkers of disease, Cancer Prev. Res. 2 (9) (2009) 807–813.
[10] L. Xu, Integrative Hypothesis Test and A5 Formulation: Sample Pairing Delta, Case Control Study, and Boundary Based Statistics, in: Lecture Notes in Computer Science, vol. 8261Springer Berlin Heidelberg, pp. 887–902. 10.1007/978-3-642-42057-3-112.
[11] L. Xu, Bi-linear matrix-variate analyses, integrative hypothesis tests, and case-control studies, Appl. Inform. 2 (2015) 4, doi:10.1186/s40535-015-0007-5.
[12] P. Leidinger, A. Keller, A. Borries, H. Huwer, M. Rohling, J. Huebers, H.-P. Lenhof, E. Meese, Specific peripheral miRNA profiles for distinguishing lung cancer from COPD, Lung Cancer 74 (1) (2011) 41–47.

[13] P. Sethupathy, B. Corda, A.G. Hatzigeorgiou, TarBase: a comprehensive database of experimentally supported animal microRNA targets, RNA 12 (2) (2006) 192–197.
[14] L. Xu, A new multivariate test formulation: theory, implementation, and applications to genome-scale sequencing and expression, Appl. Inform. 3 (2016) 1, doi:10.1186/s40535-015-0016-4.
[15] P. Leidinger, T. Brefort, C. Backes, M. Krapp, V. Galata, M. Beier, J. Kohlhaas, H. Huwer, E. Meese, A. Keller, High-throughput QRT-PCR validation of blood microRNAs in non-small cell lung cancer, Oncotarget 7 (4) (2016) 4611.
[16] D. He, J. Wang, C. Zhang, B. Shan, X. Deng, B. Li, Y. Zhou, W. Chen, J. Hong, Y. Gao, et al., Down-regulation of mir-675-5p contributes to tumor progression and development by targeting pro-tumorigenic gpr55 in non-small cell lung cancer, Mol. Cancer 14 (1) (2015) 1.
[17] G. Wang, R. Wang, Y. Strulovici-Barel, J. Salit, M.R. Staudt, J. Ahmed, A.E. Tilley, J. Yee-Levin, C. Hollmann, B.-G. Harvey, et al., Persistence of smoking-induced dysregulation of miRNA expression in the small airway epithelium despite smoking cessation, PloS ONE 10 (4) (2015) e0120824.
[18] J. Ikari, L.M. Smith, A.J. Nelson, S. Iwasawa, Y. Gunji, M. Farid, X. Wang, H. Basma, C. Feghali-Bostwick, X. Liu, et al., Effect of culture conditions on microRNA expression in primary adult control and COPD lung fibroblasts in vitro, In Vitro Cell. Dev. Biol. Anim. 51 (4) (2015) 390–399.
[19] M.M. Perry, E. Tsitsiou, P.J. Austin, M.A. Lindsay, D.S. Gibeon, I.M. Adcock, K.F. Chung, Role of non-coding RNAs in maintaining primary airway smooth muscle cells, Respir. Res. 15 (1) (2014) 1.
[20] J.H. Ohyashiki, T. Umezu, C. Kobayashi, R.S. Hamamura, M. Tanaka, M. Kuroda, K. Ohyashiki, Impact on cell to plasma ratio of mir-92a in patients with acute leukemia: in vivo assessment of cell to plasma ratio of mir-92a, BMC Res. Notes 3 (1) (2010) 347.
[21] K.B. Choo, Y.L. Soon, P.N.N. Nguyen, M.S.Y. Hiew, C.-J. Huang, MicroRNA-5p and-3p co-expression and cross-targeting in colon cancer cells, J. Biomed. Sci. 21 (1) (2014) 1.
[22] G. Zhu, Y. He, S. Yang, B. Chen, M. Zhou, X.-J. Xu, Identification of gene and microRNA signatures for oral cancer developed from oral leukoplakia, BioMed Res. Int. 2015 (2015).
[23] R. Li, K. Gao, H. Luo, X. Wang, Y. Shi, Q. Dong, W. Luan, Y. You, Identification of intrinsic subtype-specific prognostic microRNAs in primary glioblastoma, J. Exp. Clin. Cancer Res. 33 (1) (2014) 1.
[24] D.W. Huang, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using david bioinformatics resources, Nat. Protoc. 4 (1) (2008) 44–57.
[25] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucl. Acids Res. 37 (1) (2009) 1–13.
[26] A.H. Stegh, Targeting the p53 signaling pathway in cancer therapy-the promises, challenges and perils, Expert Opin. Ther. Targ. 16 (1) (2012) 67–83.
[27] J.N. Anastas, R.T. Moon, Wnt signalling pathways as therapeutic targets in cancer, Nat. Rev. Cancer 13 (1) (2013) 11–26.
[28] R. Derynck, R.J. Akhurst, A. Balmain, TGF signaling in tumor suppression and cancer progression, Nat. Genet. 29 (2) (2001) 117–129.
[29] G.W. McLean, N.O. Carragher, E. Avizienyte, J. Evans, V.G. Brunton, M.C. Frame, The role of focal-adhesion kinase in cancer: a new therapeutic opportunity, Nat. Rev. Cancer 5 (7) (2005) 505–515.
[30] V. Vasioukhin, Adherens Junctions and Cancer, Springer, pp. 379–414.
[31] X. Lu, Z. Huang, Y. Yuan, MR image super-resolution via manifold regularized sparse learning, Neurocomputing 162 (2015) 96–104.
[32] L. Xu, Enviro-geno-pheno state approach and state based biomarkers for differentiation, prognosis, subtypes, and staging, in: Applied Informatics, 3, Springer, 2016, p. 4.

Kaiming Jiang obtained his B.Eng. degree and the M.S. degree in Computer Science and Technology from Shanghai Jiao Tong University of China, in 2013, and 2016, respectively. His research interests include machine learning and Bioinformatics.

Yajing Chen obtained her B.Eng. degree in Computer Science and Technology from Shanghai Jiao Tong University of China, in 2015. She is currently a Ph.D. student of Department of Computer Science and Engineering in Shanghai Jiao Tong University. Her research interests include statistical learning and Bioinformatics.

Jinxiong Lv obtained his B.Eng. degree in Computer Science and Technology from Sichuan University of China, in 2014. He is currently a Ph.D. student of Department of Computer Science and Engineering in Shanghai Jiao Tong University. His research interests include statistical learning and Bioinformatics.

Bao-Liang Lu received the B.S. degree in instrument and control engineering from Qingdao University of Science and Technology, Qingdao, China, in 1982, the M.S. degree in computer science and technology from Northwestern Polytechnical University, Xian, China, in 1989, and the Dr. Eng. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1994. He was with Qingdao University of Science and Technology from 1982 to 1986. From 1994 to 1999, he was a Frontier Researcher with the Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research (RIKEN), Nagoya, Japan, and a Research Scientist with the RIKEN Brain Science Institute, Wako, Japan, from 1999 to 2002. Since 2002, he has been a Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He has also been an Adjunct Professor with the Laboratory for Computational Biology, Shanghai Center for Systems Biomedicine, since 2005. His current research interests include brain-like computing, neural network, machine learning, brain-computer interface, and affective computing. Dr. Lu was the President of the Asia Pacific Neural Network Assembly (APNNA) and the General Chair of the 18th International Conference on Neural Information Processing in 2011. He is currently Associate Editors of the IEEE Transactions on Cognitive and Developmental Systems, and the Neural Networks.

Lei Xu, Zhiyuan Chair Professor of Computer science and engineering department and Director of Brain Inspired Computing and Bio-Health Informatics Cente r, Shanghai Jiao Tong University; Professor of computer science and engineering, Chinese University of Hong Kong (CUHK); Guest Professor of Institute of Biophysics, CAS; Completed Ph.D. thesis at Tsinghua University by the end of 1986, joined Peking University as postdoc in 1987 and associate professor in 1988, became postdoc and visiting scientist in Finland, Canada and USA (Harvard and MIT) during 1989–1993. Then, joined CUHK as senior lecturer in 1993, professor in 1996, and chair professor during 2002–2016. He has published more than 350 academic papers and given over 50 keynote/invited/ tutorials at international conferences in the areas of statistical learning, artificial Intelligence, and bioinformatics. Served as the Editor-in-Chief of Springer OA J. Applied Informatics, and associate editors of several academic journals, including Neural Networks and IEEE Transactions on Neural Networks; Taken various roles in related academic societies, e.g., Governing Board of International Neural Networks Society (INNS) (2001–2003), the INNS award committee (2002–2003) and the Fellow committee of IEEE Computational Intelligence society (2006–2007), as well as general chairs, advisory committee chairs in a large number of intl. conferences. Prof. Xu has received several national and international academic awards, including 1993 National Nature Science Award, 1995 INNS Leadership Award and 2006 APNNA Outstanding Achievement Award. Elected to Fellow of IEEE in 2001; Fellow of intl. Association for Pattern Recognition in 2002 and of European Academy of Sciences (EAS) in 2002. Also, elected to EAS Scientific committee (2014–2017).