Bootstrapped Integrative Hypothesis Test, COPD-Lung Cancer Differentiation, and Joint miRNAs Biomarkers

Kai-Ming Jiang[1,2], Bao-Liang Lu[1,2], and Lei Xu[1,2,3(✉)]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China
lxu@cse.cuhk.edu.hk
[2] The Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China
[3] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China

Abstract. Integrative Hypothesis Test (IHT) has been recently proposed for an integrated study of hypothesis test, classification analysis and feature selection. This paper not only applies IHT to identifying miRNAs biomarkers for the differentiation of lung cancer and Chronic Obstructive Pulmonary Disease (COPD), but also proposes a bootstrapping method to enhance the reliability of IHT ranking on samples with a small size and missing values. On the GEO data set GSE24709, the previously reported fourteen differentially expressed miRNAs have been re-confirmed via one by one enumeration of their IHT ranking, with two doubtful miRNAs identified. Moreover, every pair of miRNAs is also exhaustively enumerated to examine the pairwise effect via the p-value, misclassification, and correlation, further identifying those that take core roles in coordinated effects. Furthermore, linked cliques are found featured with joint differentiation performances, which motivates us to identify such clique patterns as joint miRNAs biomarkers.

Keywords: IHT · Bootstrapping · Differential gene expression

1 Introduction

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 [1, 2] but also sera and peripheral blood cells from cancer patients [3–6]. Lung cancer closely relates to Chronic Obstructive Pulmonary Disease (COPD), a common pulmonary affliction encompassing chronic obstructive bronchitis and lung emphysema [7]. COPD is a global burden affecting 10–15 % of adults older than 40 years [8] and precedes lung cancer in 50–90 % of cases [9]. This paper also works such a topic via performing differentiation analyses on the expression of 863 miRNAs in blood cells of lung cancer patients and patients suffering from COPD, with data available in the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/, GSE24709) [9, 10].

X. He et al. (Eds.): IScIDE 2015, Part II, LNCS 9243, pp. 538–547, 2015.
DOI: 10.1007/978-3-319-23862-3_53

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.

Though there are also some studies that separately use two methods on one experiment and report the performances obtained by both the methods, there lacks effort on systematically integrating the performances of two methods. Integrative Hypothesis Test (IHT) has been recently proposed towards this purpose [11, 12]. This paper applies IHT to identifying miRNAs biomarkers for the differentiation of lung cancer and COPD. Moreover, a bootstrapping method is proposed to enhance the reliability of IHT ranking on a small size of samples and many missing values among the samples.

First, we adopt typical practice of evaluating each miRNA one by one by using this bootstrapped based IHT ranking, resulting in a list of miRNAs with the top 15 IHT ranks (See Table 2) that covers all the 14 differentially expressed ones identified in [9, 10] but in a different order of reliability. Also, we found that among the 14 ones, hsa-miR-513b and hsa-miR-93* are really doubtful in their reliability. We checked the Human microRNA Disease Database (HMDD) version two [13], and find no report of these two miRNAs related to any kind of cancer.

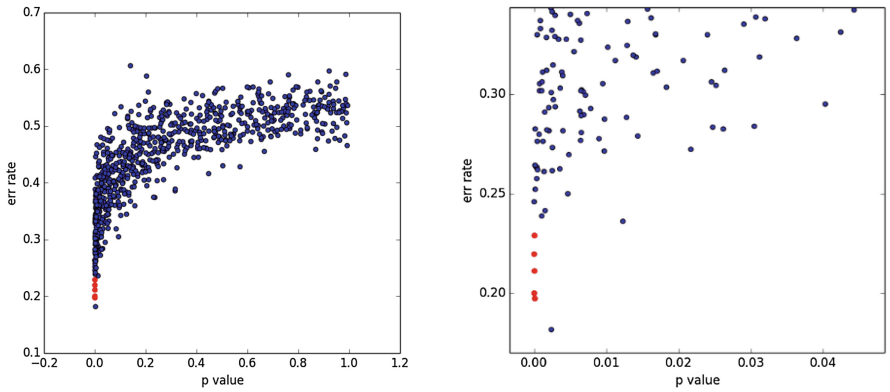
Moreover, we examine paired miRNAs not just following typical practice via Pearson correlation, but also exhaustively evaluating every pair of miRNAs by the integrated performances of the p-value and classification accuracy with help of this bootstrapped based IHT ranking, resulting in a list of top-20 pairs of miRNAs (See Table 4). Interestingly, each of 19 pairs contains one miRNA that locates within top-10 in Table 2. Especially hsa-miR-675 and hsa-miR-92a each in 6 pairs while hsa-miR-369-5p in 5 pairs, hsa-miR-641 and hsa-miR-662 each in 2 pairs seemly take important roles in the differentiation of lung cancer and COPD, featured with classification accuracy (90 % – 94.9 %) and the p-value (in around 10^{-7}– 10^{-10}).

In most of pairs, two miRNAs in a pair are not correlated or weakly correlated. There are also cliques that demonstrate joint miRNAs activities. One is featured by hsa-miR-369-5p pairing each of five miRNAs, with considerable negative or positive correlations. The other is featured by hsa-miR-675 pairing each of six miRNAs, all with negative correlations, and another is featured by hsa-miR-92a pairing each of six miRNAs, with a half in weak correlations and the other half in considerable correlations. Each of these pairs also reaches high differentiation performances with classification accuracy (89 % – 95 %) and the p-value (around 10^{-7}– 10^{-10}). Therefore, we may identify such clique patterns as joint miRNAs biomarkers.

2 Methods

2.1 Integrative Hypothesis Tests

The name of integrative hypothesis tests (IHT) was previously advocated in Reference [11] for an integrative study of case-control problems from not only a model based perspective such as two-sample test or model comparison to evaluate an overall difference but also a boundary based perspective such as classification or model prediction about boundary distinguishability and prediction performance. There are two basic tasks for each of the two perspectives, as outlined in Reference [12] by its Table 1.



(a) (b)

Fig. 1. 2D scatter plot (a) overall view, (b) zoom in view.

Given a set $X\omega = \{\mathbf{x}_{t,\omega}, t = 1, \cdots, N\omega\}$ of samples from two populations $\omega = 0,1$ (i.e., $\omega = 0$ for COPD and $\omega = 1$ for lung cancer), where each $\mathbf{x}_{t,\omega}$ is a vector that consists of all the features (i.e., miRNAs) and N_ω is the number of patients for the ωth population. A model based perspective study involves Task A and Task B, namely getting each population of samples to be described by a parametric model $q(x|\theta_\omega)$ (e.g., usually a normal distribution) and then compares the resulted models to examine the overall difference between two populations of samples. On the other hand, a boundary based study involves Task C and Task D, namely, classifying each sample to either $\omega = 1$ or 0 and examining whether a reliable separating boundary exists between the two populations of samples. Moreover, all the four tasks are associated with another problem called feature selection (i.e., identification of miRNAs).

Each of four tasks has been studied individually in the existing efforts, with each having its strength and limited coverage. However, performances of these tasks are coupled, and thus the best set of features for one task may not be necessarily the best for the others. Naturally, it was motivated to consider whether the performances of all the four tasks or at least more than one tasks can be jointly optimized. The necessarily

and feasibility have been addressed in Reference [12] and also empirically justified via a so called 2D scattering plots (see Fig. 4 in [12]).

In this paper, we adopt a simplified IHT implementation that only considers Task B and Task C. For Task B, we consider $q(x|\theta_\omega)$ to be either a univariate normal distribution when we evaluate each miRNA one by one exhaustively by the Welch's t-test or a two-variate normal distribution when we examine every pair of miRNAs exhaustively by the Hotelling T-squared test. The performance of Task B is the resulted p-value. For Task C, in the current preliminary study we simply use the linear discriminating analysis, with its performance measured by the misclassification rate. How comparative studies by using Bayes classifier and support vector machine (SVM) will be further conducted.

Also, we adopt the above mentioned 2D scattering plots to help us interactively to observe the joint performances of p-value and misclassification rate. As illustrated in Fig. 1, a small p-value indicates big difference of the two distributions and a small misclassification rate indicates a well classification of samples by a separating boundary. We are interested in those candidate points that are nearest to the origin of the coordinate space.

Though such 2D or 3D plots provide a possible joint evaluation, how to appropriately scaling each measure is a challenging issue. As addressed in Reference [12], we need to integrate multiple measures into a scalar index based on which the joint performance can be evaluated.

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 %

2.2 Rank Bootstrapping

We may turn the scattering plots in Fig. 1 into a rank list from increasingly sorting the distance to the origin of the coordinate space. Given in Table 1 is such a list obtained from evaluating each miRNA one by one exhaustively, with the p value obtained by the

Welch's t-test and the accuracy is equal to one minus the misclassification rate resulted from a linear discriminating analysis. Only the top 15 miRNAs are listed. However, the result maybe not reliable enough because there are only $N_1 = 28$ lung cancer patients and $N_0 = 24$ COPD samples. That is, we encounter a typical small sample size problem, which is actually widely encountered in the case-control studies.

Such a small sample size problem makes the resulted p value and the accuracy become random variables and thus each point in Fig. 1 may randomly move, resulting in its distance to the origin varies too, namely, the resulted rank in Table 1 is unreliable.

Bootstrapping is a widely used practice of estimating properties of an estimator when the sample size is insufficient, by measuring those properties via sampling from an approximating distribution. Also, this technique allows estimation of the sampling distribution of almost any statistic. Moreover, bootstrapping provides a way to account for the distortions caused by the specific sample that may not be fully representative of the population.

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

Therefore, the bootstrapping provides a useful tool for improving the unreliable rank in Table 1, not just for a small size of samples but also for the problem of missing data, which also happens seriously in miRNAs expression profile. The bootstrapping

will help to estimate the p-value and the misclassification rate, not for estimating their means and variances, but for estimating the mean and variance of the ranks of the corresponding miRNA, in the list obtained from increasingly sorting the distance to the origin of the coordinate space, in a way similar to Table 1 but not just for the top 15 but all the miRNAs in consideration.

Instead of sampling from an approximating distribution, resampling is implemented by constructing a number of resamples with replacement from the miRNAs expression data set. In each bootstrapping implementation, we will get a ranking list. The rank of each miRNA is a statistic in our consideration, the smaller the statistic is, the more we are interested in considering it as a candidate. After a large enough number of bootstrapping implementations, we may get the mean ranking and standard deviation ranking of each statistics, by which we may further generate a new ranking such that each miRNA is sorted increasingly according its mean ranking, subject to that its standard deviation is less than a threshold.

Though all the ranking lists addressed in this subsection consider each miRNA one by one, i.e., each row in Table 1 and Table 2 represents one miRNA, extension can be easily made to examine every pair of miRNAs simply with each row replaced by a pair of two miRNAs.

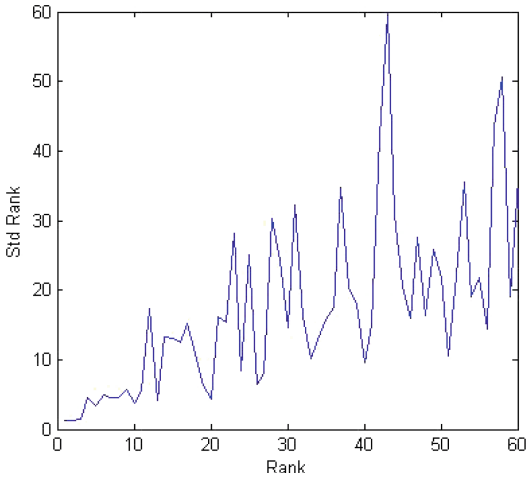


Fig. 2. Selection of a threshold for standard deviation.

3 Empirical Results

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

Evaluating each miRNA one by one exhaustively, the counterpart of Table 1 is shown in Table 2, obtained by rank bootstrapping. In each bootstrapping implementation, a size 45 of resampling samples are obtained. After 100 bootstrapping implementations, we get the mean ranking and standard deviation ranking for each statistics, by which we list the top-15 in Table 2 with miRNAs sorted increasingly according its mean ranking, subject to that its standard deviation is less than a threshold 12. As illustrated in Fig. 2, this threshold is obtained by sorting the mean ranks along the horizontal axis with the vertical axis for standard deviation, with the threshold found at the point that the standard deviation suddenly changes. The excluded genes with relatively small mean ranking but large standard deviation is also shown in Table 2.

We observe the 15 miRNAs covers all the 14 differentially expressed ones given in Table 3 but in a different order. The last three ones in Table 2, including hsa-miR-513b and hsa-miR-93* in Table 3, are really doubtful because of their reliability, we checked the Human microRNA Disease Database (HMDD) version two [13], and find no report of these two miRNAs related to any kind of cancer, while most of the other 12 genes appear in a few reports.

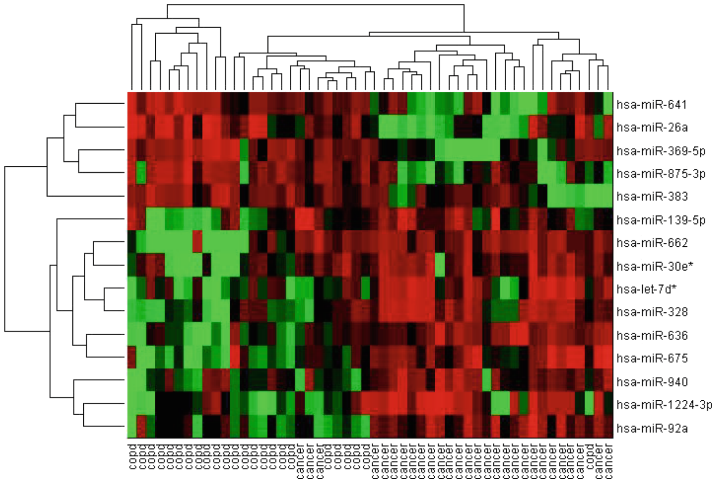


Fig. 3. Hierarchical clustering result of selected 15 miRNAs.

We apply hierarchical clustering to the miRNA expression data of those selected miRNAs and the clustering is visualized as heatmap in Fig. 3. As can be seen, obvious patterns of miRNAs can be found for COPD and lung cancer and the selected miRNAs can be divided into 2 clusters which one up-regulates with the disease and the other down-regulates.

Moreover, we examine paired miRNAs not just following the typical practice via Pearson correlation, but also exhaustively evaluate every pair of miRNAs by the integrated performances of the p-value and classification accuracy with help of this bootstrapped based IHT ranking. Table 4 gives a list of top-20 pairs of miRNAs. Being

different from Table 2, each row represents a pair of miRNAs. The Pearson correlation between the two miRNAs is listed in the last column but did not join the integrative sorting. As expected, two individually best miRNAs does not necessarily form the best pair miRNAs.

Table 3. Significant markers identified in [9, 10] 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

Interestingly, each of 19 pairs contains one miRNA that locates within top-10 in Table 3. Especially hsa-miR-675 and hsa-miR-92a each in 6 pairs while hsa-miR-369-5p in 5 pairs, hsa-miR-641 and hsa-miR-662 each in 2 pairs seemly take important roles in the differentiation of lung cancer and COPD, featured with the classification accuracy (90 % – 94.9 %) and the p-value (in around 10^{-7}– 10^{-10}). Moreover, hsa-miR-26a, hsa-let-7d*, hsa-miR-636, hsa-miR-93* are also appear in both Tables 2 and 4. Thus, a total of 9 of miRNAs in Table 2 have appeared in Table 4 and draw our particular attention for further investigation.

In most of pairs, two miRNAs in a pair are not correlated or weakly correlated. There are also cliques that demonstrate joint miRNAs activities. One consists of (has-miR-369-5p, hsa-miR-92a), (has-miR-369-5p, hsa-miR-675), (has-miR-369-5p, hsa-miR-26a), (has-miR-369-5p, hsa-miR-183), (has-miR-369-5p, hsa-miR-940), with each pair in a considerable negative or positive correlation. The other consists of (hsa-miR-675, *hsa-miR-1271*), (hsa-miR-675, hsa-miR-641), (hsa-miR-675, hsa-miR-1299), (hsa-miR-675, hsa-miR-627), (hsa-miR-675, hsa-miR-489), all with negative correlations. The another one consists of (hsa-miR-92a, has-miR-369-5p), (hsa-miR-92a, hsa-miR-1204), (hsa-miR-92a, hsa-miR-376a*), (hsa-miR-92a, hsa-miR-767-3p), (hsa-miR-92a, hsa-miR-93*), with a half in weak correlations and the other half in

Table 4. 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

considerable correlations. The three cliques are also linked via (hsa-miR-675 and hsa-miR-92a. Each pair in these cliques reaches high differentiation performances with classification accuracy (89 % – 95 %) and the p-value (around 10^{-7}– 10^{-10}). We are thus motivated to identify such clique patterns as joint miRNAs biomarkers.

4 Conclusion

This paper applies Integrative Hypothesis Test (IHT) to identifying miRNAs biomarkers for the differentiation of lung cancer and COPD via an integrative perspective of both hypothesis test and linear classification. A bootstrapping method is proposed to enhance the reliability of IHT ranking on samples with a small size and missing values. Empirical study has been made on the GEO data set GSE24709. First, we exhaustively evaluate miRNAs one by one by the bootstrapped based IHT ranking, re-discover the 14 differentially expressed ones identified in [9, 10] but two of them regarded as really doubtful in their reliability. We checked the Human microRNA Disease Database (HMDD) version two [13], and find no report about the two miRNAs are related to any kind of cancer. Second, we also exhaustively evaluate every pair of miRNAs by this IHT ranking and found that the majority of those found one by one take the roles in a list of top miRNA pairs. Furthermore, we found three mutually linked cliques that demonstrate joint miRNAs activities featured with differentiation performances in the

classification accuracy (89 % – 95 %) and the p-value (around 10^{-7}– 10^{-10}, which motivates us to identify such clique patterns as joint miRNAs biomarkers.

Acknowledgment. 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 fund for Zhiyuan Chair Professorship.

References

1. Lu, J., et al.: MicroRNA expression profiles classify human cancers. Nature **435**(7043), 834–838 (2005)
2. Barshack, I., et al.: MicroRNA expression differentiates between primary lung tumors and metastases to the lung. Pathol. Res. Pract. **206**(8), 578–584 (2010)
3. Cortez, M.A., Calin, G.A.: MicroRNA identification in plasma and serum: a new tool to diagnose and monitor diseases (2009)
4. Gilad, S., et al.: Serum microRNAs are promising novel biomarkers. PLoS ONE **3**(9), e3148 (2008)
5. Wang, J., et al.: MicroRNAs in plasma of pancreatic ductal adenocarcinoma patients as novel blood-based biomarkers of disease. Cancer Prev. Res. **2**(9), 807–813 (2009)
6. Tammemagi, C.M., et al.: Impact of comorbidity on lung cancer survival. Int. J. Cancer **103**(6), 792–802 (2003)
7. van Gestel, Y.R., et al.: COPD and cancer mortality: the influence of statins. Thorax **64**(11), 963–967 (2009)
8. Young, R.P., et al.: COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. Eur. Respir. J. **34**(2), 380–386 (2009)
9. Keller, A., Leidinger, P.: Peripheral profiles from patients with cancerous and non cancerous lung diseases, Gene Expression Omnibus (GEO, GSE24709) (2011). http://www.ncbi.nlm.nih.gov/geo/
10. Leidinger, P., et al.: Specific peripheral miRNA profiles for distinguishing lung cancer from COPD. Lung Cancer **74**(1), 41–47 (2011)
11. Xu, L.: Integrative hypothesis test and A5 formulation: sample pairing delta, case control study, and boundary based statistics. In: Sun, C., Fang, F., Zhou, Z.-H., Yang, W., Liu, Z.-Y. (eds.) IScIDE 2013. LNCS, vol. 8261, pp. 887–902. Springer, Heidelberg (2013)
12. Xu, L.: Bi-linear Matrix-variate Analyses, Integrative Hypothesis Tests, and Case-control Studies. To appear on Springer OA J. Appl. Inform., 1(1) (2015)
13. Human microRNA Disease Database. http://202.38.126.151/hmdd/tools/hmdd2.html