

Which statistical significance test best detects oncomiRNAs in cancer tissues? An exploratory analysis

Wei Tang^{1,*}, Zhijun Liao^{2,3,*} and Quan Zou^{3,4}

¹ Department of Biological Engineering, School of Chemical Engineering, Tianjin University, Tianjin, China

² Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China

³ School of Computer Science and Technology, Tianjin University, Tianjin, China

⁴ State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin, China

* These authors have contributed equally to this work

Correspondence to: Quan Zou, email: zouquan@nclab.net

Keywords: microRNA; differential expression; statistical significance test; MARS; oncomiRNA

Received: September 24, 2016

Accepted: October 14, 2016

Published: October 23, 2016

ABSTRACT

MicroRNAs(miRNAs) often exert their oncogenic and tumor suppressor functions by suppressing protein-coding genes expressions in cancers and thus have a strong association with cancers' generation, development and metastasis. Through comprehensively understanding differentially expressed miRNAs (oncomiRNA) in tumor tissues, we can elucidate the underlying molecular mechanisms in tumorigenesis and develop novel strategies for cancer diagnosis and treatment. The differential expression of miRNAs can now be analyzed through numerous statistical significance tests based on different principles, which are also available in various R packages. However, the results can be notably different. In this study, we compared miRNAs obtained from 6 common significance tests/R packages (t-test, Limma, DESeq, edgeR, LRT and MARS) with the miRNAs archived in two databases; HMDD 2.0 database, which collects experimentally validated differentially expressed miRNAs, and Infer microRNA-disease association database, which contains the potential disease-associated miRNAs by network forecasting. Finally, we sought the MARS method in DEGseq package more effectively searched out differentially expressed miRNAs than other common methods.

INTRODUCTION

MicroRNAs (miRNAs) are short (18-25-nucleotide) non-coding RNAs that function as posttranscriptional gene regulators by binding to the 3'UTR of mRNAs, consequently, either repress translation or initiate mRNA degradation [1, 2]. Since their discovery [3, 4], miRNAs have been implicated in the control of various cellular processes [5, 6], including cell proliferation [1], cell death [7-12] and differentiation [3, 13]. Therefore, many miRNAs could function as oncogenic miRNAs (oncomiRNAs), which cause cancer by down-regulating genes through both translational repression and mRNA destabilization mechanisms [14, 15], such as breast tumors [16, 17], esophageal carcinoma [18, 19] and lung cancer [1, 3]. miRNAs are also potential prognostic markers of chronic lymphocytic leukemia [20], colon tumors [15,

21], pancreatic cancer [22], and neuroblastoma [23]. Associations between differentially expressed (DE) miRNAs and the cancer occurrence have been the focus of intense cancer biology investigation [24-28].

Next Generation Sequencing (NGS) technology can rapidly and accurately perform large-scale DNA/RNA sequencing through a series of high-throughput technologies. These technologies facilitate genomic research and are increasingly replacing microarrays with gene expressing profiling of epigenetics and transcriptomics (RNA-seq) [29, 30]. Transcriptomic sequencing includes mRNA, small RNA and non-coding RNA (ncRNA), of which miRNAs are among the most important components [31-33]. Aided by the advantages of NGS, molecular biology has acquired a vast number of large-scale sequence data, which has also posed many challenges for high-throughput analysis. These challenges

include finding suitable statistical tests for large data and affirming their statistical assumptions by biological experiments, such as quantitative RT-PCR, northern blot, and overcoming the shortcomings of genetic sequencing technologies through statistical methods, which should fully uncover the essence of biology. Selective miRNAs expression profiling based on high-throughput test can strongly support the prognosis prediction of various cancers [34, 35]. Therefore, a significance test or R package which can efficiently screen out DE miRNAs in tumor tissues will guide the subsequent validation by low-throughput experiments.

Various normalizations and statistical hypotheses have been incorporated into statistical significance tests and R packages which can detect DE miRNAs in cancer tissues. For example, the t-test (Student's t-test) is widely used for comparing independent samples by statistical hypothesis test. This test examines whether the expressions of certain miRNAs significantly differ among different parent population samples. A 2011 study compared the miRNAs expressions in 20 patients with glioblastoma and other 20 age- and sex-matched healthy controls [11]. The researchers identified 52 significant DE miRNAs among 1158 tested miRNAs in glioblastoma tissues, however, only 2 miRNAs (miR-128 and miR-342-3p, which are up- and down-regulated respectively) of these 52 miRNAs were validated by low-throughput real-time PCR experiments, which means only two miRNAs were suitable biomarkers for blood-derived glioblastoma-associated characteristic miRNA fingerprints [36]. The Limma package analyses gene expression data obtained from microarrays or RNA-seq technologies. The core capability of this package is the evaluation of differential expression in multifactor-designed experiments by linear modeling [37]. Sun [38] used the Limma package to screen out numerous DE miRNAs in ductal carcinoma in situ compared with normal controls. The DESeq [39] and edgeR [40] package solve the overdispersion problem in RNA sequencing data by applying the negative binomial distribution. Hamfjord [41] used both tools to statistically test the miRNA expression differences in read counts per miRNA between two samples. In DESeq, they treated the tumor and normal samples as independent groups; in edgeR, they considered paired information. According to their results, 37 miRNAs were identified to be DE miRNAs (19 up-regulated and 18 down-regulated) in colorectal cancer both in DESeq and edgeR [41], however, among these miRNAs, 16 miRNAs had not been validated in previously documented experiments. Thus, DESeq and edgeR both have limited screening ability for DE miRNAs. In 2010, Wang [42] proposed MA-plot-based method with random sampling (MARS) in DEGseq package. This method incorporates the random sampling method and is based on MA plot, which is widely used to detect and visualize the intensity-dependent ratios in microarray data [43]. DEGseq package also includes

another commonly used method called Likelihood Ratio Test (LRT) [44].

Despite the wide range of statistical significance tests and R packages for detecting DE miRNAs in RNA expression profiles, few studies have considered which method gives the most accurate result. Along with the flourishing development of bioinformatics and applications of machine learning [24, 26, 45, 46], literature mining [47, 48] has greatly assisted biomedicine and genomics research. The HMDD 2.0 database (<http://www.cuilab.cn/hmdd>) collects experimental evidences of DE miRNAs and disease associations through literature mining [47]. The Infer microRNA-disease association database (<http://lab.malab.cn/soft/ifmda>) [49] predicts the underlying interactions between miRNAs and disease for further confirmation of biological experiments. The Infer method constructs a heterogeneous network that connects the disease similarity subnetwork to the miRNA similarity subnetwork by validated experimental miRNA-disease associations. The HMDD 2.0 database contains the experimentally validated DE miRNAs, and the Infer microRNA-disease association database offers miRNAs that are potentially associated with diseases.

In this study, we selected 5 miRNA expression profiles of cancers (BRCA, ESCA, LUAD, PAAD and THCA), and their corresponding controls miRNA profiles. We then screened the DE miRNAs in the cancer tissues by the six abovementioned methods (t-test, Limma, DESeq, edgeR, LRT and MARS), then compared and classified the six sets of results with the miRNAs in HMDD 2.0 and Infer microRNA-disease association. By calculating the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC), we identified the method with the best screening results. Among the six methods, MARS delivered the highest performance.

RESULTS

Expression analysis of miRNAs in cancer vs normal groups

After applying the six methods to the five datasets (Figure 1 and Table 1), we acquired the significant DE miRNAs in cancer tissues (P or $P_{adj} < 0.05$) among the miRNAs. The numbers of DE miRNAs returned by the six methods have huge differences (Figure 2).

The DE miRNAs in BRCA obtained by those 6 methods were compared in a Venny distribution using the Venny web server 2.0 (<http://bioinfogp.cnb.csic.es/tools/venny/index.html>) [50] (Figure 3). The miRNAs obtained from the six methods were also very different. For example, in BRCA dataset, hsa-mir-4482 was classified as a DE miRNA by Limma and edgeR, but as a normally expressed miRNA in DESeq.

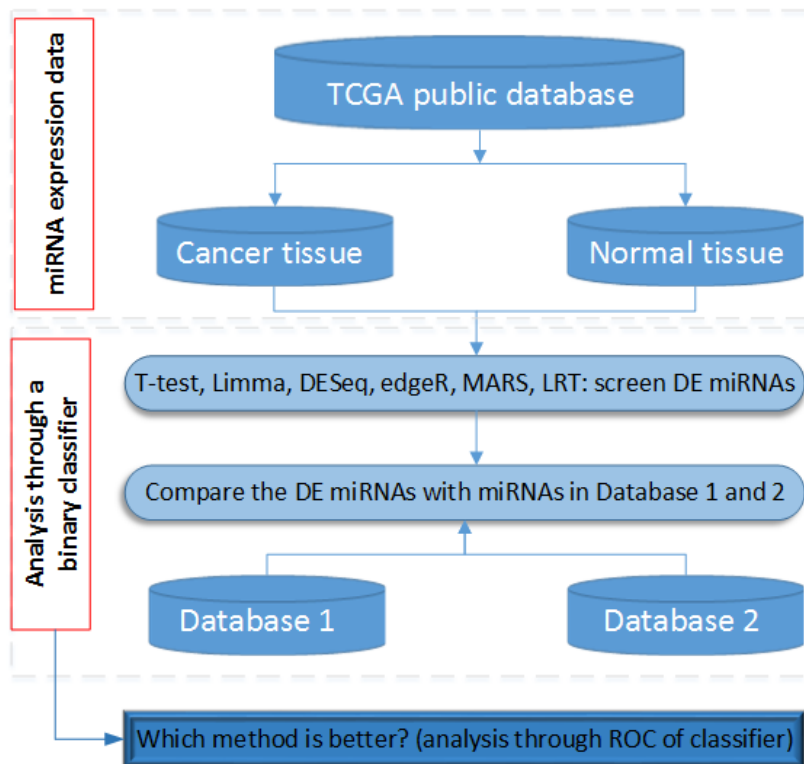


Figure 1: The flowchart to analyze miRNA and to compare with the other two databases in this study. Note: DE: differentially expressed. Database 1: HMDD 2.0 database. Database 2: Infer microRNA-disease association database

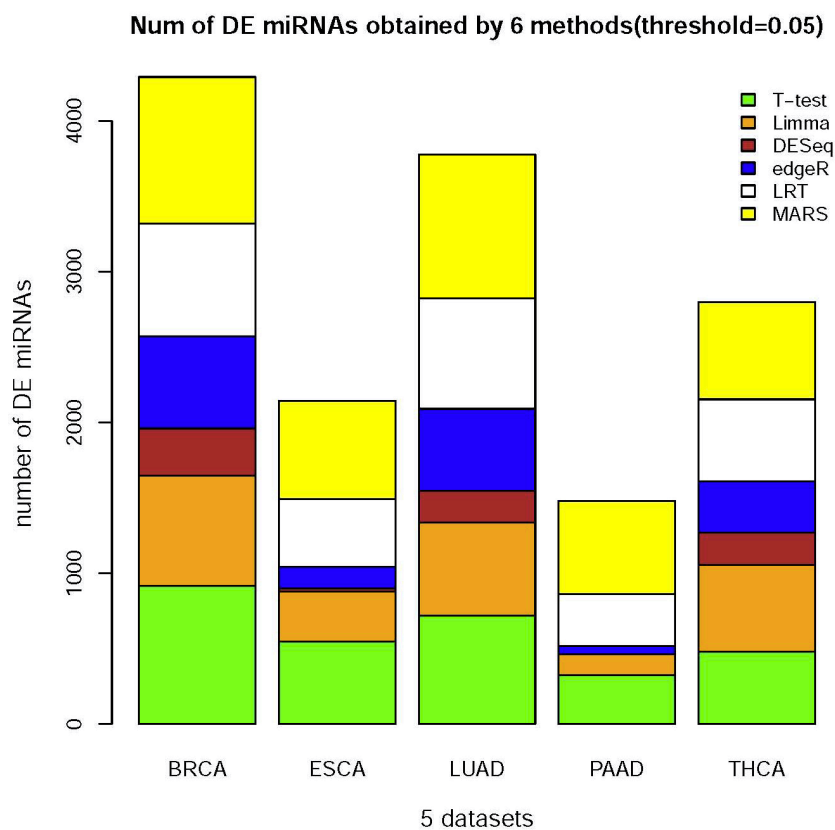


Figure 2: The histogram of the number of DE miRNAs obtained from 6 methods on 5 datasets. The DE miRNAs were obtained from 6 methods (*t*-test, Limma, DESeq, edgeR, LRT and MARS) and the threshold: P or $P_{adj} < 0.05$

Table 1.1: Number of samples of selected miRNA sequencing datasets from the TCGA database.

Cancers	NT samples	TN samples
BRCA	90	90
ESCA	13	13
LUAD	45	45
PAAD	4	4
THCA	57	57

Table 1.2: The number of selected miRNA from the database 1 and database 2

Cancers	TCGA	Database 1	Database 2
BRCA	1881	243	100
ESCA	1881	83	100
LUAD	1881	157	100
PAAD	1881	117	100
THCA	1881	56	100

Note:

NT: normal sample; TN: tumor sample;

Database 1: HMDD 2.0; Database 2: Infer microRNA-disease association;

BRCA: Breast invasive carcinoma

ESCA: Esophageal carcinoma

LUAD: Lung adenocarcinoma

PAAD: Pancreatic adenocarcinoma

THCA: Thyroid carcinoma.

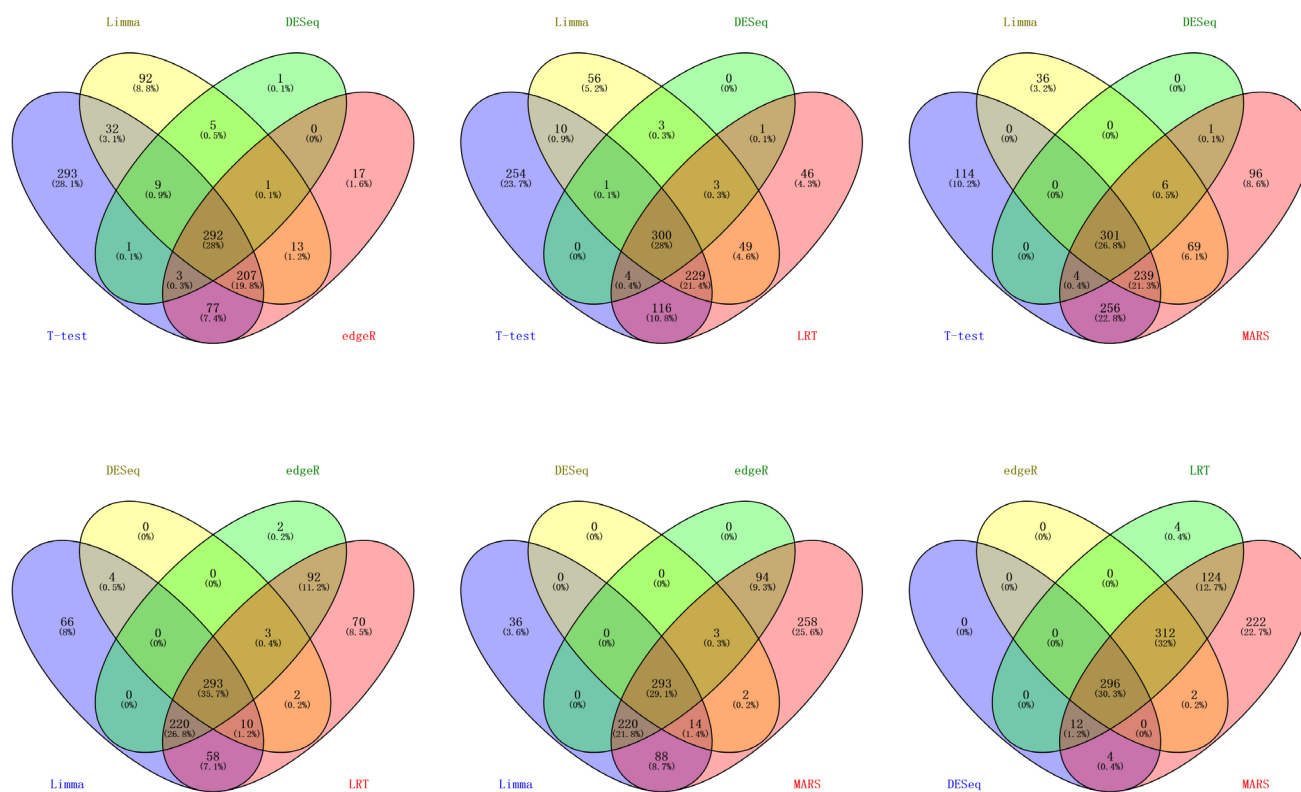


Figure 3: The venn diagrams of BRCA dataset. BRCA dataset: The DE miRNAs were obtained from 6 methods(*t*-test, Limma, DESeq, edgeR, LRT and MARS) and the threshold P or $P_{adj} < 0.05$, since venn diagrams based on 6 sets looks not intuitionistic, so we choose every 4 sets in 6 sets to draw venn diagrams.

Table 2: The binary classifier

	True class	
Hypothesized class	<i>TP(True Positives)</i>	<i>FP(False Positives)</i>
	<i>FN(False Negatives)</i>	<i>TN(True Negatives)</i>

Note:

TP: true positive, the number of predicted miRNAs by statistical methods and also that appear in HMDD 2.0 or Infer microRNA-disease association

FP: false positive, the number of predicted miRNAs by statistical methods but not appear in HMDD 2.0 and Infer microRNA-disease association

TN: true negative, the number of not predicted miRNAs by statistical methods and also not appear in HMDD 2.0 and Infer microRNA-disease association

FN: false negative, the number of not predicted miRNAs by statistical methods but still appear in HMDD 2.0 or Infer microRNA-disease association

Hypothesized class: the predicted miRNA by those 6 methods

True class: the miRNA achieved in those 2 databases, HMDD 2.0 and Infer microRNA-disease association

These two figures indicated the large differences on the DE miRNAs obtained from those 6 methods. The differences not only lie in the number of those miRNAs but also in the predicted DE miRNAs.

Compared with the miRNAs in HMDD 2.0 and Infer microRNA-disease association

The total number of DE miRNAs can be varied by adjusting the threshold (P or P_{adj}) of the six methods. Here, we classified the miRNAs obtained by the six methods with miRNAs in the HMDD 2.0 and Infer microRNA-disease association database. As these classifications are binary classifications (Table 2), we can estimate the screening DE miRNA performance of these methods by plotting the ROCs and computing their AUCs. When integrating the miRNAs from HMDD 2.0 and Infer microRNA-disease association, we incremented the k -value from 0.5 to 1 in 0.05 steps as a weight factor (Figure 4, Figures S1-S4; Table S1). When individually considering these two databases, we constructed separate ROCs for the miRNA comparisons between each statistical method and HMDD 2.0, and between each method and Infer microRNA-disease association (Figure 5, Figures S5-S8; Table S2).

DISCUSSION

Abnormal mRNA expression are induced by DE miRNAs, which prevents the mRNA from executing its regular biological functions [51-53], which is a primary cause of cancer. Altered miRNA expression will likely contribute to the initiation and progression of human cancers [10, 11, 13, 14, 16, 47], and the relationship between miRNAs and cancers has become a major focus in cancer research. Vast numbers of miRNA expression profiles have been generated throughout the past decade, as rapid NGS development has continuously lowered the

gene sequencing cost. Although DE miRNAs in tumor tissues can be detected by various available methods, the accuracy of these methods remains a critical issue.

DE genes have been ubiquitously detected by the t-test, which is popular for its simple calculation and easily understandable characteristics. Even though the standard error in the t-test is based on a small sample size, some miRNAs with miniscule standard error will still inevitably exist among the great number of miRNAs. Consequently, the t-test will increase the false positives prediction for these miRNAs [54, 55]. Examining the ROC of PAAD dataset which has only 4 cancer samples and 4 control samples, we could observe clearly that the t-test cannot selectively screen the DE miRNAs in this dataset (Figure S3, Figure S7). In addition, the performance of t-test method on PAAD dataset was worse than any other methods when we compare the results with both integrated and independent considerations of HMDD 2.0 and Infer microRNA-disease association.

To improve the estimates stability in the traditional t-test, Limma introduces a prior distribution which can strengthen the sample variance estimation. The results of ROCs clearly showed that Limma delivered much better performance than the t-test in the small sample case, such as ESCA and PAAD datasets (Figure S1, Figure S3, Figure S5, Figure S7; Table S2). Limma also outperformed the t-test in the remaining datasets.

The DESeq and edgeR packages are based on the negative binomial (NB) distribution. The NB model corrects the overdispersion problem in RNA sequence data by an additional term in the variance of the Poisson model. The variance parameter is estimated differently in DESeq and edgeR; DESeq estimates the mean-dependent dispersion by a local regression method, whereas edgeR assumes that the mean and variance are related and thus share a single common estimate of the dispersion parameter across the read counts. edgeR also weakens each miRNA's the dispersion degree through an empirical Bayes method [40]. We note that many statistical methods

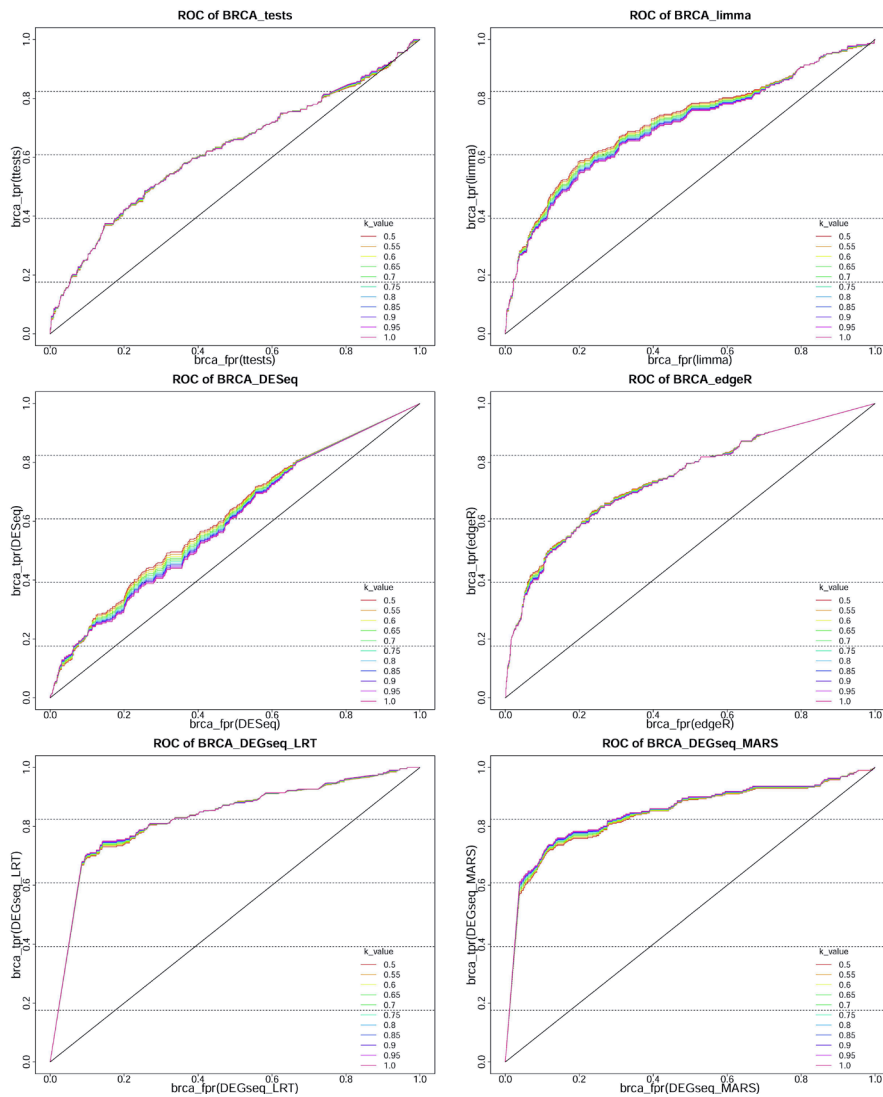


Figure 4: The ROC of 6 methods on BRCA dataset based on integrated HMDD 2.0 and Infer microRNA-disease association. These ROC are obtained from classification of miRNAs obtained from 6 methods (*t*-test, Limma, DESeq, edgeR, LRT and MARS) on 5 datasets based on the true class in integrated HMDD 2.0 and Infer microRNA-disease association and *k*-value is the weighting coefficient, which is arithmetic progression from 0.5-1 with the step size equaling to 0.05

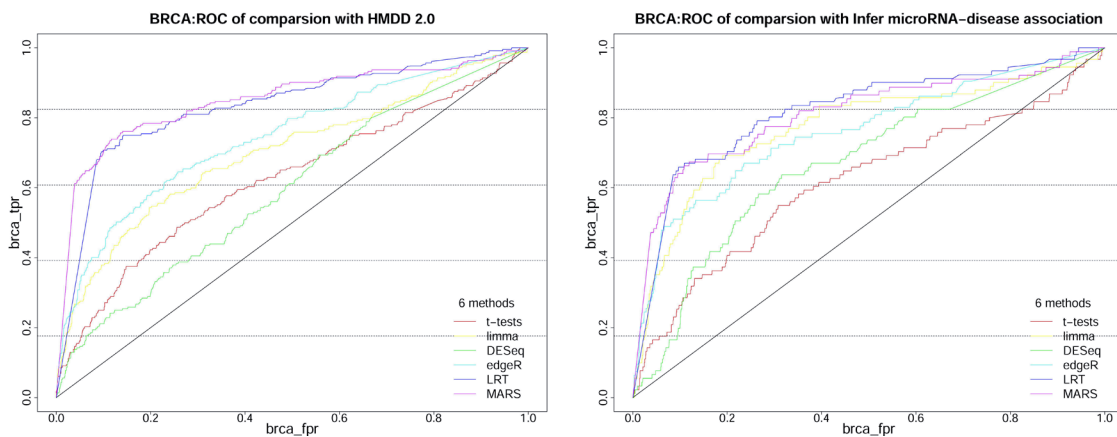


Figure 5: The ROC of 6 methods on BRCA dataset based on independently HMDD 2.0 and Infer microRNA-disease association. These ROC are obtained from classification of miRNAs obtained from 6 methods (*t*-test, Limma, DESeq, edgeR, LRT and MARS) on 5 datasets based on the true class in independently HMDD 2.0 and Infer microRNA-disease association.

cannot properly handle the small sample sizes which are very common in RNA sequencing experiments, for example, in DESeq, which are based on generalized linear models, small sample sizes consequently violated the assumptions of its statistical tests [39, 40]. As a result, DESeq was almost the worst performer in our experiments (Figures S1-S8). The edgeR method also inflated the type I error rates in the simulations. The results of both DESeq and edgeR deviated largely in relatively small samples, such as the PAAD (4) and ESCA (13) dataset (Table 1.1). The AUCs of both methods were also quite different (Tables S1-S2); in the ESCA dataset, the AUC in edgeR was 0.18-0.2 larger than in DESeq, and in the PAAD dataset, the DESeq was even ineffective due to the very small sample size. In other three datasets, which has relatively larger sample sizes than PAAD and ESCA, reduced the differences between performances in edgeR and DESeq (Tables S1-S2). Our simulations also confirmed a higher computational speed of edgeR than DESeq, moreover, the latter could even incur memory leakage at relatively large sample sizes.

According to the technical characteristic of RNA-Seq, Wang [42] proposed the MARS method in DEGseq, which can detect DE genes from MA plots and its test hypothesis is based on a random sampling model [42]. The DEGseq package includes LRT as well. Both methods were demonstrated higher DE miRNA screening performance in all five datasets than the other four methods (t-test, Limma, DESeq and edgeR). The results showed some AUCs of MARS and LRT were close to 0.90 and some were much higher than 0.9, suggesting that the both methods are very effective on detecting DE miRNAs (Figures S1-S8, Tables S1-S2). Moreover, the AUCs of the PAAD and ESCA dataset confirmed that MARS and LRT can also effectively identify DE miRNAs even in small samples. Although both MARS and LRT performed well, MARS achieved a higher True Positive Rate (TPR) at low False Positive Rate (FPR) than LRT in the BRCA, ESCA, LUAD and THCA datasets, in PAAD dataset, MARS and LRT achieved nearly identical TPR at the same low FPR (Figures S3 and S7), which indicated that MARS could correctly identify DE miRNAs with fewer misidentified miRNAs than LRT (Figures S1-S8). In summary, among the six tested methods, MARS could most accurately detect the DE miRNAs in cancer tissues.

The detection of DE miRNAs in cancers (oncomiRNAs) has always been among the most important issue in cancer biology research. Prior accurate computational detection of DE miRNAs will effectively reduce the cost of clinical experiments. Current computational analyses focus on statistical significance tests [56], literature mining and networking prediction [57]. However, few works have considered all of these approaches. In the present study, we integrated these three approaches to maximize the uniformity of the results to see

which method could most accurately detect DE miRNAs. The DE miRNAs detected by the best performer (MARS) were highly consistent with the miRNAs extracted from literature mining and network prediction. This supports our inference that MARS outperforms other statistical significance tests (such as t-tests) in DE miRNAs detection. Complex genetic regulatory mechanisms in high-level organisms is considered to be achieved through controlled and coordinated miRNAs networks. The associations between miRNAs and disease are not only conducive to develop novel therapeutic applications for cancer patients by miRNA delivery and inhibition, but also help to construct the RNA network which is crucial to understand the underlying mechanisms of genetic network. In future work, we will continue to focus on improving the accuracy of oncomiRNA detection through integrating significance tests, literature mining and network prediction with including more data resources and also involving machine learning methods modify the MARS method.

MATERIALS AND METHODS

Flowchart

Figure 1 is a flowchart of the present study. After a comprehensive analysis of miRNA expression in cancerous and normal tissue samples, the results were compared with the miRNAs in HMDD 2.0 and the Infer microRNA-disease associations. Finally, we identified the best DE miRNA detection method among the six statistical significance tests/R packages.

Source data and sequence expression analysis

In this study, we selected five deep sequencing miRNA datasets from The Cancer Genome Atlas (TCGA) pilot project (<https://tcga-data.nci.nih.gov/tcga/>). All of these data were sequenced by the BCGSC (IlluminaHiSeq_miRNAseq) sequencing platform. HMDD 2.0 database collects DE miRNAs in various cancer types which were validated by biological experiments. However, many of the cancer types in HMDD 2.0 have not collected sufficiently experimentally validated miRNAs, only 5 cancer types not only collect more than 50 validated DE miRNAs in HMDD 2.0, included the miRNA sequences of breast invasive carcinoma (BRCA), esophageal carcinoma (ESCA), lung adenocarcinoma (LUAD), pancreatic adenocarcinoma (PAAD) and thyroid carcinoma (THCA), but also have corresponding miRNAs in Infer microRNA-disease association database. So in order to assure the statistical significance, we selected these 5 datasets (BRCA, ESCA, LUAD, PAAD and THCA) for further analysis. As the sample sizes differed between tumor and

normal samples, we randomly selected corresponding tumor samples with the same and similar characteristics (Table 1).

After obtaining the miRNA expression profiles from TCGA (the original sequencing data had been subjected to mapping analysis), we analyzed the miRNAs through abovementioned statistical methods (t-test, Limma, DESeq, edgeR, LRT and MARS), and computed the corresponding P or P_{adj} (the associated FDR (False Discovery Rate)) value of each miRNA in the five datasets. Results were deemed significant at the P or $P_{adj} = 0.05$ level.

Statistical comparison with HMDD 2.0 and Infer microRNA-disease association

If the P or P_{adj} value was below 0.05, the miRNA expression between the cancer and normal samples was considered as statistically significant, and the miRNA was assumed as a DE miRNA in the tumor tissue. By varying P or P_{adj} , we can vary the numbers of miRNAs that pass the hypothesis test.

All of the tested methods generated a P or P_{adj} value for each miRNA in the miRNAs differential expression analysis. Therefore, the comparison between the miRNAs obtained from these six methods and those in HMDD 2.0 database and Infer microRNA-disease association database can be regarded as a binary classification process. Here, the predicted outputs are the miRNAs obtained from the significance tests/R packages, and the true classes are the miRNAs of the corresponding cancers in HMDD 2.0 and Infer microRNA-disease association. The matrix so constructed is the basis of many common metrics (Table 2)

The performance of a binary classification could be characterized by two basic measurements; the recall rate and precision rate. However, these measurements are of limited usefulness due to their single-valued feature [58], we instead computed the True Positive Rate (TPR) and False Positive Rate (FPR), then plotted the ROC, which could be quantified by the AUC. In biomedical applications, the ROC is commonly used to judge the performance of a discriminant across varying decision thresholds [59, 60], and it has become increasingly important in the classification of unequally distributed categories, since its unique attributes can handle unequal costs incurred by classification errors [59, 60]. Thus, the ROC could provide a better metric than the accuracy measure in certain classifiers [61, 62]. In the binary classification of our DE miRNA analysis, we specified a threshold for the obtained outcome, such as 0.01, and assigned respectively all instances above and below this value as negative (no differential expression) and positive (differential expression). Increasing the threshold to 0.05 will increase the number of true positive instances, and

thereby the proportion of true positives among all positive instances (the true positive rate, or TPR) increases. However, a higher threshold also classifies more negative instances as positive, increasing the false positive rate (FPR). The AUC of the ROC is another indicator of the classifier performance [63, 64]. The TPR and FPR are respectively calculated as:

$$TPR(\text{TruePositiveRate}) = \frac{TP}{TP+FN}, (1)$$

$$FPR(\text{TruePositiveRate}) = \frac{FP}{FP+TN}, (2)$$

Where True Positive (TP) and False Positive (FP) denote the numbers of positive and negative samples that are classified as positive, respectively, and True Negative (TN) and False Negative (FN) represent the corresponding values of the negative samples.

Because the AUC is a portion of a unit square area, its value always lies in the range 0-1.0. Another important statistical property of AUC is that the value equals the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. A higher AUC shifts the ROC toward the upper-left of the square, indicating higher performance of the classifier [60]. As the miRNAs in HMDD 2.0 were validated by biological experiments and those miRNAs in Infer microRNA-disease association were predicted by network, we must assign a weighting coefficient when considering the true classes in an integrated manner. Here, we applied a simple linear function. Specifically, we modified the TP and FN by introducing a parameter k :

$$TP(\text{True Positive}) = \frac{k+sum1}{(1-k)+sum2}, (3)$$

$$FN(\text{False Positive}) = \frac{k+(sum3-sum1)}{(1-k)+(sum3-sum2)}, (4)$$

In Eqs. (3) and (4), $sum1$ and $sum2$ are the sums of the miRNAs appearing in HMDD 2.0 (correctly classified positives in HMDD 2.0) and Infer microRNA-disease association (correctly classified positives in the Infer database), respectively, and $sum3$ is the total sum of the miRNAs obtained by all methods.

As the miRNAs obtained from Infer microRNA-disease association might not be associated with cancers, they should be weighted less heavily than those in HMDD 2.0, whose associations with cancer are confirmed. Hence, the k -value was varied as $0.5 \leq k \leq 1$

ACKNOWLEDGMENTS

The work was supported by the Natural Science Foundation of China (No. 61370010), the Natural Science Foundation of Fujian Province of China (No.2016J01152), and the State Key Laboratory of Medicinal Chemical Biology of China (No.201601013).

CONFLICTS OF INTERESTS

The authors declare no potential conflict of interests with respect to the authorship and/or publication of this paper.

REFERENCES

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*. 2004; 116: 281-97.
2. Liu B, Fang L, Liu F, Wang X, Chen J, Chou K-C. Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS ONE*. 2015; 10: e0121501.
3. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*. 1993; 75: 843-54.
4. Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*. 1993; 75: 855-62.
5. Chen X, Yan CC, Zhang X, You Z-H. Long non-coding RNAs and complex diseases: from experimental results to computational models. *Briefings in Bioinformatics*. 2016: bbw060.
6. Liu B, Fang L, Jie C, Liu F, Wang X. miRNA-dis: microRNA precursor identification based on distance structure status pairs. *Molecular BioSystems*. 2015; 11: 1194-204.
7. Hatfield S, Shcherbata H, Fischer K, Nakahara K, Carthew R, Ruohola-Baker H. Stem cell division is regulated by the microRNA pathway. *Nature*. 2005; 435: 974-8.
8. Cheng L-C, Tavazoie M, Doetsch F. Stem cells: from epigenetic to microRNAs. *Neuron*. 2005; 46: 363-7.
9. Houbaviy HB, Murray MF, Sharp PA. Embryonic stem cell-specific MicroRNAs. *Developmental cell*. 2003; 5: 351-8.
10. Suh M-R, Lee Y, Kim JY, Kim S-K, Moon S-H, Lee JY, Cha K-Y, Chung HM, Yoon HS, Moon SY. Human embryonic stem cells express a unique set of microRNAs. *Developmental biology*. 2004; 270: 488-98.
11. Zhang B, Pan X, Anderson TA. MicroRNA: a new player in stem cells. *Journal of cellular physiology*. 2006; 209: 266-9.
12. Wu D, Huang Y, Kang J, Li K, Bi X, Zhang T, Jin N, Hu Y, Tan P, Zhang L. ncRDeathDB: A comprehensive bioinformatics resource for deciphering network organization of the ncRNA-mediated cell death system. *Autophagy*. 2015; 11: 1917-26.
13. Chen C-Z, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. *science*. 2004; 303: 83-6.
14. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA. MicroRNA expression profiles classify human cancers. *nature*. 2005; 435: 834-8.
15. Volinia S, Calin GA, Liu C-G, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National academy of Sciences of the United States of America*. 2006; 103: 2257-61.
16. Yu F, Yao H, Zhu P, Zhang X, Pan Q, Gong C, Huang Y, Hu X, Su F, Lieberman J. *let-7* regulates self renewal and tumorigenicity of breast cancer cells. *Cell*. 2007; 131: 1109-23.
17. Iorio MV, Ferracin M, Liu C-G, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M. MicroRNA gene expression deregulation in human breast cancer. *Cancer research*. 2005; 65: 7065-70.
18. Feber A, Xi L, Luketich JD, Pennathur A, Landreneau RJ, Wu M, Swanson SJ, Godfrey TE, Litle VR. MicroRNA expression profiles of esophageal cancer. *The Journal of thoracic and cardiovascular surgery*. 2008; 135: 255-60.
19. Kano M, Seki N, Kikkawa N, Fujimura L, Hoshino I, Akutsu Y, Chiyomaru T, Enokida H, Nakagawa M, Matsubara H. miR-145, miR-133a and miR-133b: tumor-suppressive miRNAs target FSCN1 in esophageal squamous cell carcinoma. *International Journal of Cancer*. 2010; 127: 2804-14.
20. Calin GA, Liu C-G, Sevignani C, Ferracin M, Felli N, Dumitru CD, Shimizu M, Cimmino A, Zupo S, Dono M. MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101: 11755-60.
21. Michael MZ, O'Connor SM, van Holst Pellekaan NG, Young GP, James RJ. Reduced Accumulation of Specific MicroRNAs in Colorectal Neoplasia. Note: Susan M. O'Connor and Nicholas G. van Holst Pellekaan contributed equally to this work. *Molecular Cancer Research*. 2003; 1: 882-91.
22. Bloomston M, Frankel WL, Petrocca F, Volinia S, Alder H, Hagan JP, Liu C-G, Bhatt D, Taccioli C, Croce CM. MicroRNA expression patterns to differentiate pancreatic adenocarcinoma from normal pancreas and chronic pancreatitis. *Jama*. 2007; 297: 1901-8.
23. Schulte JH, Schowe B, Mestdagh P, Kaderali L, Kalaghatgi P, Schlierf S, Vermeulen J, Brockmeyer B, Pajtlar K, Thor T. Accurate prediction of neuroblastoma outcome based on miRNA expression profiles. *International Journal of Cancer*. 2010; 127: 2374-85.
24. Chen X, Clarence Yan C, Zhang X, You ZH, Huang YA, Yan GY. HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction. *Oncotarget*. 2016; 7: 65257-65269. doi: 10.18632/oncotarget.1125.
25. Chen X, You ZH, Yan GY, Gong DW. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. *Oncotarget*. 2016; 7: 57919-31. doi:

- 10.18632/oncotarget.11141.
26. Chen X, Yan G-Y. Semi-supervised learning for potential human microRNA-disease associations inference. *Scientific reports*. 2014; 4:doi:
 27. Chen X, Huang YA, Wang XS, You ZH, Chan K. FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model. *Oncotarget*. 2016; 7: 45948-58. doi: 10.18632/oncotarget.10008.
 28. Wang Y, Chen L, Chen B, Li X, Kang J, Fan K, Hu Y, Xu J, Yi L, Yang J. Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. *Cell Death & Disease*. 2013; 4: : e765.
 29. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. *Genomics*. 2008; 92: 255-64.
 30. Metzker ML. Sequencing technologies—the next generation. *Nature reviews genetics*. 2010; 11: 31-46.
 31. Keller C, Bühler M. Chromatin-associated ncRNA activities. *Chromosome Research*. 2013; 21: 627-41.
 32. Mattick JS, Makunin IV. Non-coding RNA. *Human molecular genetics*. 2006; 15: R17-R29.
 33. Li Y, Wang C, Miao Z, Bi X, Wu D, Jin N, Wang L, Wu H, Qian K, Li C. ViRBase: a resource for virus-host ncRNA-associated interactions. *Nucleic Acids Research*. 2014; 43: 578-82.
 34. Calin GA, Croce CM. MicroRNA-cancer connection: the beginning of a new tale. *Cancer research*. 2006; 66: 7390-4.
 35. Calin GA, Croce CM. MicroRNA signatures in human cancers. *Nature Reviews Cancer*. 2006; 6: 857-66.
 36. Roth P, Wischhusen J, Happend C, Chandran PA, Hofer S, Eisele G, Weller M, Keller A. A specific miRNA signature in the peripheral blood of glioblastoma patients. *Journal of neurochemistry*. 2011; 118: 449-57.
 37. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*. 2015: gkv007.
 38. Sun E, Zhou Q, Liu K, Wei W, Wang C, Liu X, Lu C, Ma D. Screening miRNAs related to different subtypes of breast cancer with miRNAs microarray. *Eur Rev Med Pharmacol Sci*. 2014; 18: 2783-8.
 39. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome biology*. 2010; 11: 1.
 40. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26: 139-40.
 41. Hamfjord J, Stangeland AM, Hughes T, Skrede ML, Tveit KM, Ikdahl T, Kure EH. Differential expression of miRNAs in colorectal cancer: comparison of paired tumor tissue and adjacent normal mucosa using high-throughput sequencing. *PLoS one*. 2012; 7: e34150.
 42. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*. 2010; 26: 136-8.
 43. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*. 2002; 30: e15-e.
 44. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*. 2008; 18: 1509-17.
 45. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, Dong Q, Chou K-C. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics*. 2014; 30: 472-9.
 46. Chen J, Wang X, Liu B. iMiRNA-SSF: Improving the Identification of MicroRNA Precursors by Combining Negative Sets with Different Distributions. *Scientific Reports*. 2016; 6: 19062.
 47. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic acids research*. 2013: gkt1023.
 48. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research*. 2009; 37: D98-104.
 49. Liu Y, Zeng X, He Z, Zou Q. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2016: Doi: 10.1109/TCBB.2016.2550432. doi: 10.1109/TCBB.2016.2550432.
 50. Oliveros JC. VENNY. An interactive tool for comparing lists with Venn Diagrams. 2007.
 51. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & development*. 2006; 20: 515-24.
 52. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*. 2010; 466: 835-40.
 53. Zhang B, Pan X, Cobb GP, Anderson TA. microRNAs as oncogenes and tumor suppressors. *Developmental biology*. 2007; 302: 1-12.
 54. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome biology*. 2003; 4: 1.
 55. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*. 2001; 17: 509-19.
 56. Guo L, Yu J, Liang T, Zou Q. miR-isomiRExp: a web-server for the analysis of expression of miRNA at the

- miRNA/isomiR levels. *Scientific Reports*. 2016; 6: 23700
57. Zou Q, Li J, Hong Q, Lin Z, Shi H, Wu Y, Ju Y. Prediction of microRNA-disease associations based on social network analysis methods. *BioMed Research International*. 2015; 2015: 810514.
 58. Hripscak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*. 2005; 12: 296-8. doi:
 59. Fawcett T. An introduction to ROC analysis. *Pattern recognition letters*. 2006; 27: 861-74.
 60. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*. 2008; 27: 157-72.
 61. Wang P, Tang K, Weise T, Tsang E, Yao X. Multiobjective genetic programming for maximizing ROC performance. *Neurocomputing*. 2014; 125: 102-18.
 62. Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. *Bioinformatics*. 2010; 26: 822-30.
 63. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*. 2008; 17: 145-51.
 64. Tang H, Chen W, Lin H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Mol Biosyst*. 2016; 12: 1269-75. doi: 10.1039/c5mb00883b.