Screening of Serum miRNAs as Diagnostic Biomarkers for Lung Cancer Using the Minimal-Redundancy-Maximal-Relevance Algorithm and Random Forest Classifier Based on a Public Database

Xiaoyan Huang  Xiong Chen  Xi Chen  Wenling Wang
Medical Oncology, 900 Hospital of the Joint Logistics Team, Fuzhou, China

Keywords
Lung cancer · Minimal redundancy-maximal relevance · Receiver operator characteristic · Principal component analysis · Serum biomarker

Abstract
Background: Lung cancer is one of the deadliest cancers, early diagnosis of which can efficiently enhance patient’s survival. We aimed to screening out the serum miRNAs as diagnostic biomarkers for patients with lung cancer. Methods: A total of 416 remarkably differentially expressed miRNAs were acquired using the limma package, and next feature ranking was derived by the minimal-redundancy-maximal-relevance method. An incremental feature selection algorithm of a random forest (RF) classifier was utilized to choose the top 5 miRNA combination with the optimum predictive performance. The performance of the RF classifier of top 5 miRNAs was analyzed using the receiver operator characteristic (ROC) curve. Afterward, the classification effect of the 5-miRNA combination was validated through principal component analysis and hierarchical clustering analysis. Analysis of top 5 miRNA expressions between lung cancer patients and normal people was performed based on GSE137140 dataset, and their expression was validated by qPCR. The hierarchical clustering analysis was used to analyze the similarity of 5 miRNAs expression profiles. ROC analysis was undertaken on each miRNA. Results: We acquired top 5 miRNAs finally, with the Matthews correlation coefficient value as 0.988 and the area under the curve (AUC) value as 0.996. The 5 feature miRNAs were capable of distinguishing most cancer patients and normal people. Furthermore, except for the lowly expressed miR-6875-5p in lung cancer tissue, the other 4 miRNAs all expressed highly in cancer patients. Performance analysis revealed that their AUC values were 0.92, 0.96, 0.94, 0.95, and 0.93, respectively. Conclusion: By and large, the 5 feature miRNAs screened here were anticipated to be effective biomarkers for lung cancer.

Introduction
Lung cancer has triggered abundant associated deaths as an aggressive tumor [1, 2]. Despite immense advancement and outcomes thus far, more investigations into early diagnosis and novel targets are still warranted [3]. The recent years have witnessed the development and research of molecular biomarkers becoming a crucial part in optimizing early diagnosis of lung cancer [4].
Blood-derived biomarkers are expected to be mainstream ones due to their stability and efficiency. Blood-derived miRNAs display excellent diagnostic performance as a new biomarker type [5]. For stance, Tobias et al. [6] published a multicenter cohort study on the early diagnosis of lung cancer and built a 15-miRNA signature with machine learning. Though they fully analyzed differences in miRNA expression mode between cancer and normal population, the gradient algorithm used alone is a limitation [6].

Hanchuan Peng first raised the minimal-redundancy-maximal-relevance (mRMR) algorithm in 2005 and screened excellent features out of many [7]. mRMR can recognize genes with independent effect and remove the redundant features of the selected genes [8], by which feature set with the maximum relevance and minimum redundancy can be obtained [9]. In 2008, Zhang et al. [10] established different classifiers based on feature genes screened with mRMR and RliefF to distinguish diseases and normal status. Lu et al. [11] employed mRMR and a simple decision tree model to predict benign and malignant ovarian tumors, manifesting a rather favorable performance. Besides, mRMR has been applied in prediction of lysine ubiquitination [12], protein-protein interaction [13], and human immunodeficiency virus progression-associated genes [14, 15]. In general, mRMR is applicable to screen feature genes of tumors from large gene sets.

Random forest (RF) is a classifier consisting of several decision trees [16] and a classifying method of integrated learning, which is active and can prevent overfitting [17]. This method has been employed in resolving many bioinformatics problems, such as identification of proteins and peptide [18] and prediction of transcription factor binding [17]. This paper utilized this method for the development of a novel diagnosis for lung cancer.

In this study, mRMR was used for feature ranking of differentially expressed miRNAs, and top 5 miRNAs were identified. Subsequently, the optimal features were selected using incremental feature selection (IFS) first, and then receiver operator characteristic (ROC) plot was used to examine the selected classifier. In addition, principal component analysis (PCA) and hierarchical clustering analysis revealed that top 5 miRNAs could effectively distinguish lung cancer samples from normal samples. Finally, our speculation was further verified by expression detection of cancer and normal samples. In general, these 5 feature miRNAs were screened and validated by combining bioinformatics methods and qPCR detection. Their diagnostic effects enable them to work as efficient serum biomarkers for lung cancer.

Methods

Acquirement of miRNA Expression Data and Differential Expression Analysis

Expression data of serum miRNAs in lung cancer and non-tumor groups were accessed from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). The data (GSE137140), including serum miRNA expression of 1,566 preoperative lung cancer patients and 2,178 nontumor population, were acquired by GPL21263 platform. Data were standardized by the limma package [19] and then proceeded to differential expression analysis (logFC > 1.5, FDR < 0.05).

Feature Ranking Using mRMR Algorithm

mRMR is capable of selecting excellent features for clinical feature classification [7]. The algorithm works by finding a feature with the max-relevance with the output outcome and maintaining the minimum redundancy between features. Given two random variables $x$ and $y$, their probability density function was $p(x, y)$, $p(x, y)$. The mutual information was therefore as follows (the larger $I(x; y)$, the larger the relevance between the two variables):

$$I(x; y) = \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dx \, dy. \quad (1)$$

Based on the function above, max-relevance satisfies the following formula:

$$\max D(S, c), D = \frac{1}{|S|} \sum_{i \in S} I(x_i; c). \quad (2)$$

$x_i$, features in the feature set $S$; $c$: classification results. However, when features in S have high relevance, removing one of them affects little on the classification effect. Thus, the minimum redundancy formula was introduced to reduce relevance between features:

$$\min R(S), R = \frac{1}{|S|^2} \sum_{i, j \neq i} I(x_i; x_j). \quad (3)$$

$x_i/x_j$, features in S. The mRMR algorithm could be constructed by combining max $D$ and min $R$. The algorithm was simplified by setting the operator ($\phi$):

$$\max \phi (D, R), \phi = D - R \quad (4)$$

Here, remarkably differentially expressed miRNAs were taken as the feature set and ranked using mRMR.

Construction of the RF Classifier and IFS

Important feature miRNAs were selected with the mRMR algorithm, but which combination of miRNAs had the optimum classification performance was not determined. Thus, a series feature miRNA sets $F = \{F_1, F_2, \ldots, F_n\}$ (fn: feature miRNA, n: integer from 1 to 5) were acquired based on top 5 miRNAs ranked by mRMR. Afterward, the python package (sklearn) was applied to establish the RF classifier [20] for each set ($F_1, F_2, \ldots, F_5$). RF is an ensemble algorithm based on decision trees. Each decision tree classifies samples, whose results are summarized by voting and then outputted according to the voting results. In consideration of unequal sample numbers, the python package (imblearn) was used for simulation training by the upsampling method [21]. To detect the classification performance, we calculated the 10-fold cross-val-
Validation Matthews correlation coefficient (MCC) in the training set of each RF classifier and drew the IFS curve [22]. According to the MCC value in the IFS curve, the optimum feature miRNA combination was chosen. In addition, the ROC curve was drawn based on the 5-miRNA RF classifier with the python package (sklearn) to evaluate the classification performance [23].

**PCA and Hierarchical Clustering Analysis**

PCA is able to reduce the feature dimensionality of data while retaining data variables as much as possible [24]. In such an algorithm, new variables are generated via linear combinations of the original features, namely principal components. The classification effect could be validated by observing data distribution in a space that consists of these new principal components. Using the FactoMineR package, 5 feature miRNAs in this investigation were transformed as PC1 and PC2 that made up a 2-dimensional plane, the distribution of each sample on which was then observed. Additionally, hierarchical clustering analysis was performed on samples using the pheatmao package according to the expression of 5 feature miRNAs [25]. In this way, whether cancer and normal samples could be classified based on the 5-gene expression was verified.

**Research Subjects and Sample Collection**

Hundred lung cancer patients and 80 normal people from June 2020 to June 2021 in 900 Hospital of the Joint Logistics Team were...
enrolled as subjects. Their clinical information and serum samples were gathered. All patients did not receive any treatment while they were diagnosed according to the American Joint Committee on Cancer: The 7th Cancer Staging Manual [26]. All subjects’ peripheral blood (5 mL) was collected in a 5-mL clotting tube (Greiner Bio-One, Austria). Serums were separated by centrifugation and stored at −80°C for miRNA extraction. The above steps were approved by the Ethics Committee of 900 Hospital of the Joint Logistics Team.

**qPCR**

According to the instructions of manufacturers, TRIzol (Invitrogen, USA) was applied to isolate total RNA from serums. The miScript II RT Kit (QIAGEN, Germany) was used to reversely transcribe total RNA into cDNA. The cDNA was quantified and amplified on Bio-Rad CFX96 qPCR apparatus using miScript SYBR green (Bio-Rad, USA). Primers (See Table 1) of miR-1290, miR-663a, miR-3192-5p, miR-1343-3p, miR-6875-5p, U6 (internal reference) were all procured from GENEWIZ, China. The relative expression of miRNAs was calculated by the 2−ΔΔCT method.

**Statistics and Analysis**

GraphPad Prism (GraphPad Software, USA) was applied for statistical processing. The $T$ test was employed to test expression difference significance which was identified with $p$ value <0.05. ROC analysis of GraphPad Prism was employed to analyze the predictive performance of each miRNA. * refers to $p$ less than 0.05.

### Table 3. Clinical information of subjects

|                        | Cancer | Normal |
|------------------------|--------|--------|
| Number                 | 100    | 80     |
| Gender, n (%)          |        |        |
| Male                   | 70 (70)| 45 (56.2)|
| Female                 | 30 (30)| 35 (43.8)|
| Age, n (%)             |        |        |
| ≤60                    | 25 (25)| 37 (46.2)|
| >60                    | 75 (75)| 43 (53.8)|
| Smoking status, n (%)  |        |        |
| Current                | 20 (20)| 32 (40) |
| Formal                 | 74 (74)| 12 (15) |
| Never                  | 6 (6)  | 36 (45) |
| Stage, n (%)           |        |        |
| 1                      | 6 (6)  |        |
| 2                      | 23 (23)|        |
| 3                      | 48 (48)|        |
| 4                      | 23 (23)|        |
| EGFR mutation, n (%)   |        |        |
| Y                      | 16 (16)|        |
| N                      | 20 (20)|        |
| NA                     | 64 (64)|        |

Y, yes; N, no; NA, not available.
Results

Feature Ranking and Expression Analysis of the Differentially Expressed miRNAs

Analysis of the expression profiles of serum miRNAs revealed 416 remarkably differentially expressed miRNAs (online suppl. Table 1; for all online suppl. material, see www.karger.com/doi/10.1159/000525316). These miRNAs were ranked with the mRMR method according to feature importance. Top 5 miRNAs were used herein after (Table 2).

Evaluation of the Classification Performance of the Top 5 miRNAs

This part focused on the probability of top 5 feature miRNAs as serum biomarkers for lung cancer. To this end, IFS and ROC curves of the RF classifier based on different feature sets (F₁, F₂, ..., F₅) were utilized to analyze its accuracy. In the IFS curve, the top 5 miRNA-based RF classifier displayed the maximum MCC value (0.988) (Fig. 1a). According to the ROC curve, the AUC value was 0.996 (Fig. 1b). Integrated analyses demonstrated that the top 5 miRNA-based RF classifier could effectively distinguish cancer samples from normal ones.

PCA and Hierarchical Clustering Analysis

To further verify whether these 5 feature miRNAs could efficiently classify samples, PCA and hierarchical clustering analysis were, respectively, undertaken. As shown by PCA results, two kinds of samples could be distinguished in the 2-dimensional plane consisting of PC1

Fig. 3. Analysis of top 5 miRNA expression based on GSE137140 dataset. a–e Expression of miR-6875-5p, miR-3192-5p, miR-1290, miR-663a, and miR-1343-3p, respectively. Green: normal samples. Pink: tumor samples.
and PC2 (Fig. 2a). As also exhibited by hierarchical cluster analysis, the expression of the 5 miRNAs could distinguish most of the sample types (Fig. 2b). Next, expression of these miRNAs in samples was analyzed. MiR-6875-5p was downregulated in tumor samples, while the others were upregulated (Fig. 3a–e). Altogether, these 5 feature miRNAs could efficiently distinguish cancer samples from normal ones.

Validation of the Expression and Predictive Performance of Top 5 miRNAs

To verify whether these miRNAs could be used as serum biomarkers for lung cancer, we collected clinical information (Table 3) and serum samples of subjects (100 tumor patients and 80 normal people). Expression of these miRNAs was measured with qPCR. Except for the remarkably low expression of miR-6875-5p in the tumor group (Fig. 4a), others (miR-3192-5p, miR-1290, miR-663a, miR-1343-3p) displayed markedly high expression when compared to normal group (Fig. 4b–e). Such results were consistent with the expression mode of top 5 miRNAs in section 2.3. Next, hierarchical clustering analysis revealed that most cancer patients could be well distinguished from normal people (Fig. 4f). The ROC curve of each miRNA was drawn to analyze corresponding predictive performance. The predicted AUC of miR-6875-5p, miR-3192-5p, miR-1290, miR-663a, and miR-1343-3p were, respectively, 0.92, 0.96, 0.94, 0.95, and 0.93 (Fig. 5a–e). In conclusion, each feature miRNA exhibited a favorable predictive performance for lung cancer.
Discussion

Evidence has described strong relevance between certain miRNA expression and cancer development. Zhang et al. [27] discovered a high miR-31 level in colorectal carcinoma tissue and its association with radiotherapy sensitivity. This miRNA is a promising prognostic biomarker for colorectal carcinoma, which was also revealed in the subsequent cellular and molecular experiments. Zhou et al. [28] applied expression detection, cell viability experiment, and protein detection to reveal the promoting effect of miR-665 in ovarian cancer, suggesting its probability as a predictive biomarker. Generally, different miRNAs might serve as diagnostic or predictive biomarkers of cancers. Five serum miRNAs, miR-1290, miR-663a, miR-3192-5p, miR-1343-3p, and miR-6875-5p, were obtained through differential expression analysis and the mRMR method. MiR-1343-3p has been reported to be highly expressed in lung cancer patients [29], which coincided with our analysis. In Yining Wu’s study, the authors found that miR-1290 is a potential diagnostic and prognostic marker of lung adenocarcinoma [30]. Huang et al. [31] discovered that miR-663a is a potential biomarker for the diagnosis of osteosarcoma. Neda Gilani’s work demonstrated that miR-1343-3p can be used as a reliable biomarker for gastric cancer diagnosis [32]. These results suggest that these the 3 miRNAs can be used as potential markers for cancer diagnosis. Our results also observed 5 miRNAs (including the above 3 miRNAs) as indicators of lung cancer diagnosis.

Early screening of high-risk lung cancer populations by detecting circulating miRNAs has been extensively applied. In 2010, Bianchi et al. [33] researched a 34-miRNA-based signature for early diagnosis of non-small-cell lung cancer. Researchers in this investigation constructed a predictive model using linear discriminant analysis and
hierarchical clustering analysis and undertook early screening of lung cancer in the large-scale high-risk lung cancer population. The accuracy of that model is 80%, and the AUC of ROC is 0.89, suggestive of a rather stable predictive performance. Thereafter, that team simplified 34 miRNAs to 13 miRNAs in 2015 and established a 13-miRNA model with a good early screening performance [34]. However, with the application of machine learning and pervasion of molecular diagnosis, more suitable algorithms might exist for feature gene screening and model construction for lung cancer. Different from Fabrizio Bianchi’s research, this investigation used the mRMR method to screen feature miRNAs, and these miRNAs-based RF classifiers showed well-predictive performance. Lastly, the performance was further validated by detecting their expression in clinical samples.

Overall, this investigation acquired serum miRNA expression profiles from the Gene Expression Omnibus database and used the mRMR method to explore feature miRNAs of lung cancer. Afterward, the 5-miRNA-based RF classifier showed optimum classification performance as determined by IFS curve analysis, and this was also validated by the ROC curve. Next, PCA and hierarchical clustering analysis further demonstrated that the classifier could effectively distinguish between lung cancer samples and normal samples. Lastly, the expression mode of these miRNAs was validated by qPCR in the blood samples. The predictive performance of each miRNA was verified by the curve. Although this investigation determined lung cancer-related feature genes and their expression modes in lung cancer patients using bioinformatics analysis and qPCR detection, some limitations should still be acknowledged. This investigation was not involved in corresponding cellular and animal experiments, and thus, the specific mechanism cannot be deeply researched. Associated experiments will be designed in our plans in pursuit of specific effects of these 5 feature genes in lung cancer.

Statement of Ethics

Our study did not require an ethical board approval and consent to participate statement because it did not contain human or animal trials.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

The authors declare there is no funding used in the article.

Author Contributions

Conceptualization: Xiaoyan Huang. Data curation: Xi Chen. Formal analysis: Xiaoyan Huang. Methodology: Xiong Chen. Validation: Xiaoyan Huang. Investigation: Xi Chen. Writing – original draft: Wenling Wang. Writing – review and editing: Wenling Wang and Xiong Chen.

Data Availability Statement

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66:115–32.
2. Reck M, Popat S, Reinmuth N, De Ruyscher D, Kerr KM, Peters S, et al. Metastatic non-small-cell lung cancer (NSCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2014;25(Suppl 3):iii27–39.
3. Zhang N, Nan A, Chen L, Li X, Jia Y, Qiu M, et al. Circular RNA circSATB2 promotes progression of non-small cell lung cancer cells. Mol Cancer. 2020;19:101.
4. Seijo LM, Peled N, Ajona D, Boeri M, Field JK, Sozzi G, et al. Biomarkers in lung cancer screening: achievements, promises, and challenges. J Thorac Oncol. 2019;14:343–57.
5. Sestini S, Boeri M, Marchiano A, Pelosi G, Galeone C, Verri C, et al. Circulating microRNA signature as liquid-biopsy to monitor lung cancer in low-dose computed tomography screening. Oncotarget. 2015;6:32868–77.
6. Fehlmann T, Kahraman M, Ludwig N, Backes C, Galata V, Keller V, et al. Evaluating the use of circulating microRNA profiles for lung cancer detection in symptomatic patients. JAMA Oncol. 2020;6:714–23.
7. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226–38.
8. Xu Y, Deng Y, Ji Z, Liu H, Liu Y, Peng H, et al. Identification of thyroid carcinoma related genes with mRMR and shortest path approaches. PLoS One. 2014;9:e94022.
9. Chen CW, Chang KP, Ho GW, Chang HP, Chu YW. KStable: a computational method for predicting protein thermal stability changes by K-star with regular-mRMR feature selection. Entropy. 2018;20:988.
10. Zhang Y, Ding C, Li T. Gene selection algorithm by combining reliefF and mRMR. BMC Genomics. 2008;9(Suppl 2):S27.
11. Lu M, Fan Z, Xu B, Chen L, Zheng X, Li J, et al. Using machine learning to predict ovarian cancer. Int J Med Inform. 2020;141:1041952020.
12. Cai Y, Huang T, Hu L, Shi X, Xie L, Li Y. Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids. 2012;42:1387–95.
Screening of Diagnostic Biomarkers for Lung Cancer by mRMR

13 Li BQ, Feng KY, Chen L, Huang T, Cai YD. Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. PLoS One. 2012;7:e43927.
14 Ma C, Dong X, Li R, Liu L. A computational study identifies HIV progression-related genes using mRMR and shortest path tracing. PLoS One. 2013;8:e78057.
15 Shen S, Gui T, Ma C. Identification of molecular biomarkers for pancreatic cancer with mRMR shortest path method. Oncotarget. 2017;8:41432–9.
16 Deist TM, Dankers F, Valdes G, Wijsman R, Hsu IC, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. Med Phys. 2018;45(7):3449–59.
17 Xu T, Li B, Zhao M, Szulwach KE, Street RC, Lin L, et al. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. Nucleic Acids Res. 2015;43:2757–66.
18 Ulintz PJ, Zhu J, Qin ZS, Andrews PC. Improved classification of mass spectrometry database search results using newer machine learning approaches. Mol Cell Proteomics. 2006;5:497–509.
19 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
20 Saraswat M, Arya KV. Feature selection and classification of leukocytes using random forest. Med Biol Eng Comput. 2014;52:1041–52.
21 Mishra S. Handling imbalanced data: SMOTE vs random undersampling. Int Res J Eng Technol. 2017;4:317–20.
22 Tan JX, Dao FY, Lv H, Feng PM, Ding H. Identifying phage virion proteins by using two-step feature selection methods. Molecules. 2018;23:2000.
23 Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. Med Phys. 2018;45:314–21.
24 Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26:303–4.
25 Tian R, Wang L, Zou H, Song M, Liu L, Zhang H. Role of the XIST-miR-181a-COL4A1 axis in the development and progression of keratoconus. Mol Vis. 2020;26:1–13.
26 Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010;17:1471–4.
27 Yang B, Xiong WY, Hou HJ, Xu Q, Cai XL, Zeng TX, et al. Exosomal miRNAs as biomarkers of cancer: a meta-analysis. Clin Lab. 2019;65.
28 Zhou P, Xiong T, Yao L, Yuan J. MicroRNA-665 promotes the proliferation of ovarian cancer cells by targeting SRCIN1. Exp Ther Med. 2020;19:1112–20.
29 Zhang X, Du L, Han J, Li X, Wang H, Zheng G, et al. Novel long non-coding RNA LINC02323 promotes epithelial-mesenchymal transition and metastasis via sponging miR-1343-3p in lung adenocarcinoma. Thorac Cancer. 2020;11:2506–16.
30 Wu Y, Wei J, Zhang W, Xie M, Wang X, Xu J. Serum exosomal miR-1290 is a potential biomarker for lung adenocarcinoma. Onco Targets Ther. 2020;13:7809–18.
31 Huang C, Sun Y, Ma S, Vadamooootoo AS, Wang L, Jin C. Identification of circulating miR-663a as a potential biomarker for diagnosing osteosarcoma. Pathol Res Pract. 2019;215:152411.
32 Gilani N, Arabi Belaghi R, Afabahi Y, Faramarzi E, Edgünlü T, Somi MH. Identifying potential miRNA biomarkers for gastric cancer diagnosis using machine learning variable selection approach. Front Genet. 2021;12:779455.
33 Bianchi F, Nicassio F, Marzi M, Belloni E, Dall’olio V, Bernard L, et al. A serum circulating miRNA diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer. EMBO Mol Med. 2011;3:495–503.
34 Montani F, Marzi MJ, Dezi F, Dama E, Carletti RM, Bonizzi G, et al. miR-Test: a blood test for lung cancer early detection. J Natl Cancer Inst. 2015;107:djw063.