Translational Oncology
www.transonc.com

Volume 11 Number 1 February 2018 pp. 157–167 157

Machine Learning Based Prediction of Brain Metastasis of Patients with IIIA-N2 Lung Adenocarcinoma by a Three-miRNA Signature[1,2,3]

Shuangtao Zhao, Jiangyong Yu and Luhua Wang

Department of Radiation Oncology, National Cancer Center/ Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China

Abstract

OBJECTIVES: MicroRNAs (miRNAs) play a key role in governing posttranscriptional regulation through binding to the mRNAs of target genes. This study is to assess miRNAs expression profiles for identifying brain metastasis-related miRNAs to develop the predictive model by microarray in tumor tissues. *METHODS:* For this study, we screened the significant brain metastasis-related miRNAs from 77 lung adenocarcinoma (LUAD) patients with brain metastasis (BM+) or non-brain metastasis (BM−). A predictive model was developed from the training set (n = 42) using a random Forest supervised classification algorithm and a Class Centered Method, and then validated in a test set (n = 35) and further analysis in GSE62182 (n = 73). The independence of this signature in BM prediction was measured by multivariate logistic regression analysis. *RESULTS:* From the training set, the predictive model (including hsa-miR-210, hsa-miR-214 and hsa-miR-15a) stratified the patients into two groups with significantly different BM subtypes (90.4% of accuracy). The similar predictive power (91.4% of accuracy) was obtained in the test cohort. As an independent predictive factor, it was closely associated with BM and had high sensitivity and specificity in predicting BM in clinical practice. Moreover, functional enrichment analysis demonstrated that this signature involved in the signaling pathways positively correlated with cancer metastasis. *CONCLUSION:* These results suggested that the three-miRNA signature could develop a new random Forest model to predict the BM of LUAD patients. These findings emphasized the importance of miRNAs in diagnosing BM, and provided evidence for selecting treatment decisions and designing clinical trials.

Translational Oncology (2018) 11, 157–167

Introduction

As the most common cancer, the lung cancer is still the main cause of cancer death worldwide including China for several decades [1]. The lung cancer was the top 1 of common cancers for men and top 2 for women in 2015, and there were about 733,000 newly cases of lung cancer and more than 610,000 deaths in China [2,3]. Lung cancer consisted of two leading types: non-small cell lung cancer (NSCLC), accounting for approximately 85% and small cell lung cancer. The incidence of LUAD has increased dramatically and about half of NSCLC are lung adenocarcinoma (LUAD) [4]. Although surgery plus adjuvant therapy provides a potential

Address all correspondence to: Luhua Wang, 17 Panjiayuan Nanli, Beijing, 100021, China. Tel.: +86 10 87788221; fax: +86 10 87712804.
E-mail: wlhwq@yahoo.com

[1] Competing interests: The authors declare that they have no competing interests.

[2] Funding: This work was supported by National High-tech Research and Development Projects (863) (No. 2015AA020106, L. Wang)

[3] Importance of the Study: MiRNAs are usually discovered to have important regulatory roles in LUAD. Here we reported a predictive model constructed by random Forest supervised classification algorithm with a three-miRNA signature to predict brain metastasis (BM) of LUAD. When the new model was applied into clinical practice by measuring the expression profiles of these three miRNAs with microarray in tumor tissues, the probability of BM could be obtained with high sensitivity and specificity. And multivariate logistic regression analysis was performed between this signature and the clinical factors to identify this signature as an independent biomarker. This new model could detect the early changes accurately and predict BM status to reduce its risk such as neurologic, cognitive and emotional difficulties and finally poor prognosis in LUAD management. Generally, this new model was more clinically practice since it only uses three miRNAs.

Received 12 October 2017; Accepted 4 December 2017

1936-5233/18
https://doi.org/10.1016/j.tranon.2017.12.002

curative strategy, lung cancers often develop recurrence with a low survival rate especially in patients with metastatic disease, for example, brain metastasis (BM). Up to 25% of these patients could be influenced by BM during their lifetime [5]. BM can bring about significant neurologic, cognitive and emotional difficulties [6] and finally poor prognosis [7]. Currently, no common therapeutic measures appeared to reduce the risk of BM in LUAD. Therefore, it is urgently needed to develop novel biomarkers for detecting the early tumor changes accurately and predicting BM status as the principle warrant in LUAD management.

microRNAs (miRNAs) are an endogenous conserved class of non-coding 20–24 nucleotide small RNAs, which can regulate gene expression at post-transcriptional level by binding to 3′-UTR of targeted mRNAs and give rise to mRNA degradation or translation inhibition [8,9]. Recently, miRNAs have been studied to characterize tumors [10]. Meanwhile, a miRNA appears to regulate numerous protein-coding genes and as a result, miRNA profiling could work as a preferable classifier compared to gene expression profiling [11]. Previous studies have shown that the expression values of hsa-miR-210 [12] and hsa-miR-214 [13] are higher in LUAD tissues or A549 cells than normal control, and identified that they are metastasis-related miRNAs. Oppositely, hsa-miR-15a acts as a tumor suppressor induces cell apoptosis and inhibits cancer metastasis in NSCLC [14]. However, there is no data regarding a signature model including these three miRNAs developed by machine learning algorithm in predicting BM of patients with LUAD up to now.

This study reports the examination of miRNA expression profiles in tumor and normal tissues in a large cohort of 150 samples. A random Forest model including the three-miRNA signature was identified in the training set with the ability to predict the BM of LUAD patients and validated its diagnostic power in an independent test cohort.

Materials and Methods

Ethics Statement

And this study was approved by the medical ethics committee of the Cancer Hospital of the Chinese Academy of Medical Sciences (CAMS). All patients provided written informed consent.

Patients and Samples

The tumor tissues were retrospectively collected from 77 patients with LUAD with following-up information and examined the miRNA expression profile of the tissues by microarray analysis (Supplementary text). The formalin-fixed paraffin-embedded (FFPE) specimens of the 77 cases were pathologically confirmed with IIIA-N2 LUAD. All patients had surgically proven primary LUAD and received lobectomy or pneumonectomy in the CAMS between 2003 and 2005.

To validate the expression level of these three miRNAs in the signature, the miRNA profiles and RT-PCR value of the LUAD patients and normal controls were downloaded from the GEO databases (http://www.ncbi.nlm.nih.gov/geo/). A total of 46 normal lung samples and 27 LUAD samples from GSE62182 dataset were analyzed in this study.

Statistical Analysis

Pearson χ^2 test and the Kruskal-Wallis H test analysis of variance were used to identify the statistical differences in clinical characteristics and pathological factors. The paired Student *t* test was utilized to compare the distributive differences of the researching miRNAs between BM+ and BM− group. Considering the interrelated relationship among the candidate factors, multivariate logistic regression analysis was adopted to identify the independent signature in predicting the BM subtypes. The correlation between the three-miRNA signature and the overall survival (OS), disease-free survival (DFS), first brain metastasis (FBM) or single brain metastasis (SBM) of patients was evaluated by univariate Cox regression analysis. Survival differences between BM+ and BM− group in each condition were assessed by the Kaplan–Meier estimation, and compared by the log-rank test. The receiver-operating characteristic (ROC) curves analysis was used to evaluate discriminatory accuracy of each gene, and allowed us to visualize the sensitivity and specificity of the correlated genes in assigning LUAD patients to PBM (predictive brain metastasis) + or PBM− group before further categorization just like the application reported previously. The performance of each gene could be quantified by the area under the ROC curve (AUC). Statistical analysis was performed with SPSS 13.0, and presented with GraphPad Prism 5.0 and R3.2.5 software. Results were considered statistically significant at $P < .05$.

Results

MiRNA Expression Profiles Displayed Significant Differences Between BM− and BM+ LUAD Patients in the Microarrays Data

To screen the most befitting variables for constructing a predictive model in IIIA-N2 LUAD, 77 patients with BM− or BM+ were selected to measure the miRNAs' expression in FFPE specimens by microarray assay (Figure 1*A*). Then a total of 330 miRNAs (accounting for 42%; Figure 1*A*) were evaluated with the expression value in the first cohort (n = 42; Figure 1*B*) and second cohort (n = 35; Figure 1*C*). As a result, 8 significant miRNAs (including 6 down-regulated genes and 2 up-regulated genes) were discovered with the different expression analysis between patients with BM− and those with BM+ in the two cohorts (*q-value*< 0.05 and Fold Change >2; Figure 1*D*). These results indicated that these common significantly alterative miRNAs might be associated with the presence of brain metastasis in LUAD.

Construction of a Novel Model with a Three-miRNA Signature in the Training Set

To investigate a novel model for predicting the risk of brain metastasis, the association between miRNA expression and brain metastasis probability of patients with LUAD was explored (Supplementary text). A three-miRNA signature (including hsa-miR-15a, hsa-miR-210 and hsa-miR-214, Figure 2*E*) and a machine learning algorithm (random Forest, Figure 2*F*) were screened to construct a predictive model from the training dataset considering a balance between the error rates and the number of miRNAs. The expression value of the three miRNAs measured by high throughput sequencing was verified between normal lung (n = 46) and LUAD (n = 27) tissues in GSE62182 (Supplementary Fig. 1A&B), and validated between primary and metastatic LUAD cell lines with expression profiling by RT-PCR in GSE63819 (Supplementary Fig. 1C), which showed that three miRNAs expression tendency was in line with that in the training dataset. In this signature, the "BM+" and "BM−" centroids were (68.71, 862.54, 1425.19) and (141.16, 346.32, 527.86), which represented the average expression value of the three miRNAs for BM+ and BM− patients, respectively. This signature was defined as follows:

$$D_{ip} = \sqrt{\left(x_1^i\text{-}68.71\right)^2 + \left(x_2^i\text{-}862.54\right)^2 + \left(x_3^i\text{-}1425.19\right)^2}$$

$$D_{in} = \sqrt{\left(x_1^i\text{-}141.16\right)^2 + \left(x_2^i\text{-}346.32\right)^2 + \left(x_3^i\text{-}527.86\right)^2}$$

where $x_1^i x_2^i x_3^i$ denoted the expression value of hsa-miR-15a, hsa-miR-210 and hsa-miR-214 for sample i, respectively. A patient

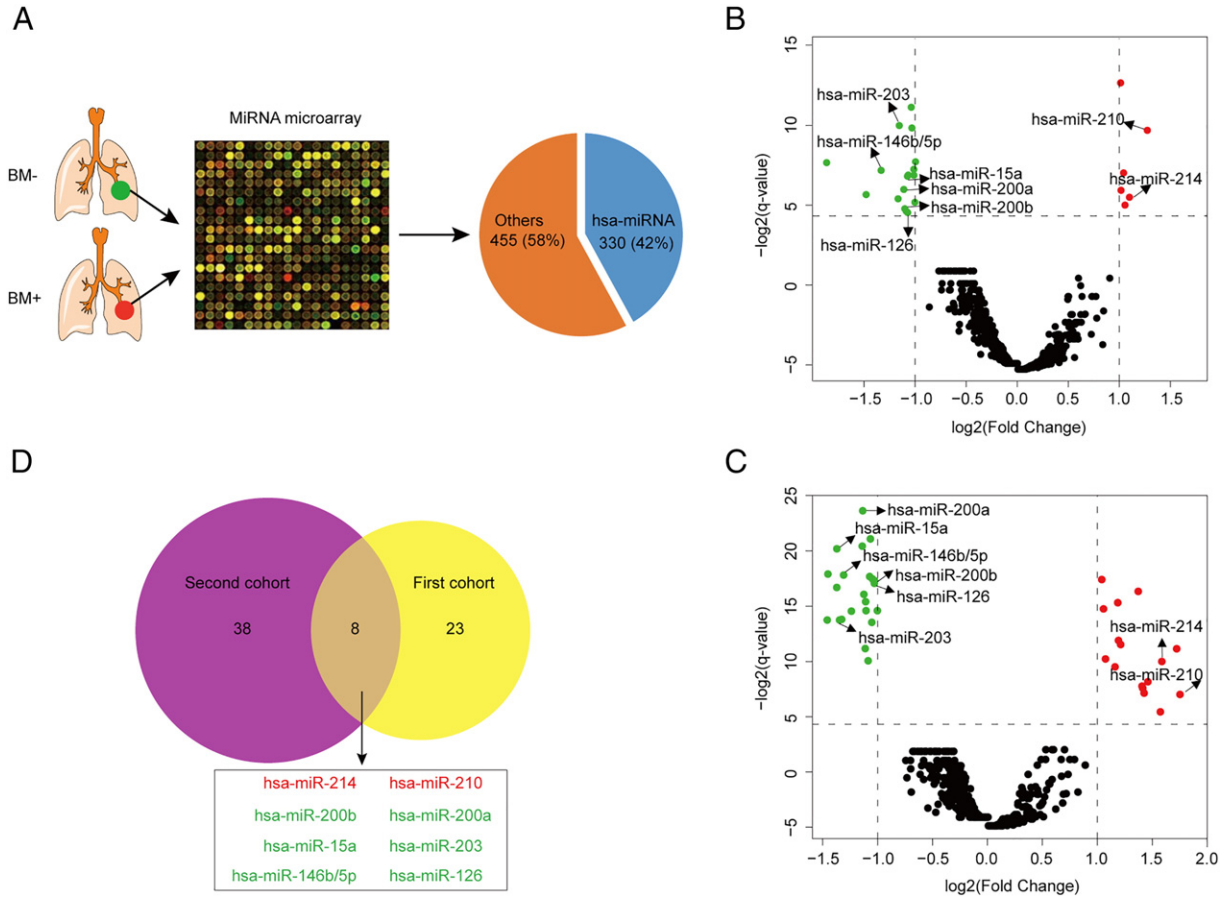


Figure 1. Eight significant miRNAs were selected between BM+ and BM− patients with LUAD in the miRNA microarray data. miRNA microarray test (**A**) was performed in LUAD tissues from BM+ or BM− patients and 330 (42%) hsa-miRNAs and 455 (58%) other genes were presented with pie-chart. Volcano plots show the significant miRNAs between BM+ and BM− patients with LUAD from the first cohort (**B**) and the second cohort (**C**); Red dots mean the up-regulated genes, green dots mean the down-regulated genes and black dots mean the unaltered genes. Venn diagram (**D**) shows the overlapped genes which are significant between BM+ and BM− patients with LUAD in both the first cohort and the second cohort; Red color represents the up-regulated genes and green color represents the down-regulated genes for BM+ vs. BM− patients.

was classified as "BM+" if $D_{ip} < D_{in}$ according to the three-miRNA expression profiles and as "BM−" if not.

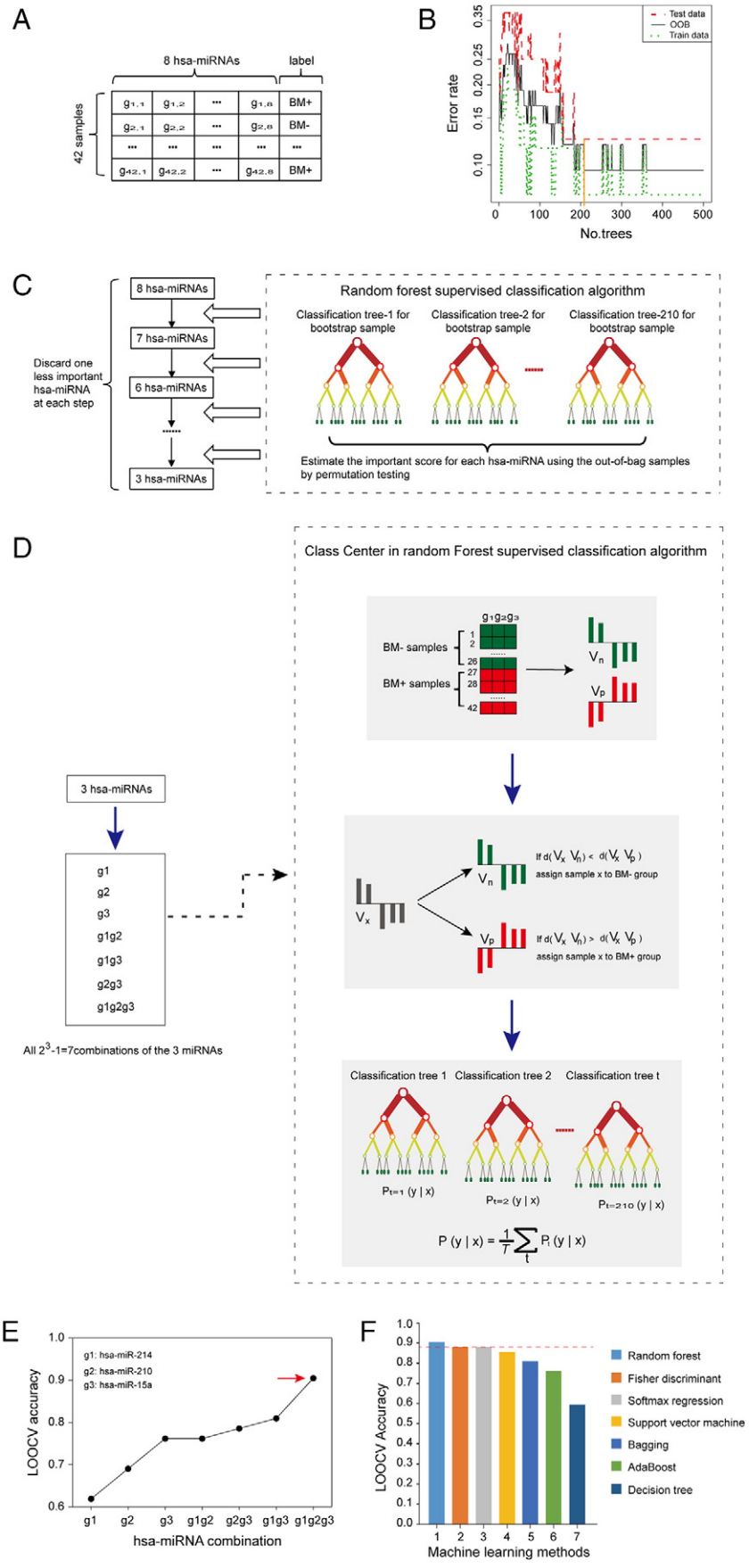
A Three-miRNA Signature Predicted Effectively the Brain Metastasis of Patients with LUAD

The classified analysis was conducted with the three-miRNA signature in the novel model. As a result, the predictive model correctly characterized 90.5% (38/42) of patients as either PBM− (2 mismatches out of 26 BM− patients) or PBM+ (2 mismatches out of 16 BM+ patients) subtype in the training dataset (Figure 3*A*–*B*). The three-miRNA signature model was then tested for its predictive power in the second cohort of 35 patients. The same criteria as those derived from the training set divided these 35 patients into PBM− and PBM+ groups, respectively (Figure 3*C*). Similarly, the whole error rate (8.6%, Figure 3*D*) of the test cohort was as low as the training set. Next, the two cohorts were combined into one group to verify the predictive power of this new model (Figure 3*E*). As a result, the new model with this three-miRNA signature had a concordance of 90.9% with the primary result classified by clinical imaging for these 77 patients as having either PBM+ (2 mismatches out of 32 BM+ patients) or PBM− (5 mismatches out of 45 BM− patients) subtype (Figure 3*F*). All the results indicated that the new model based on a three-miRNA signature could be more feasible for clinical use as its higher accuracy in predicting the brain metastasis of LUAD patients.

Brain Metastasis Prediction by the Three-miRNA Signature Model was Independent of Clinical and Pathological Factors

To gain further insights into the predictive role of the three-miRNA signature in LUAD brain metastasis, the association between this signature expression and the basic clinical characteristics from the perioperative period was conducted with univariate analysis in these 77 patients (Figure 4*A*). Obviously, the results showed that the signature was closely associated with the performance status (P = .013), the positive lymph nodes (P = .040), the number of positive lymph nodes stand (P = .029), the number of N2 positive lymph nodes stand (P = .036), N2 positive lymph nodes stand classification (P = .003), Post-radiotherapy (P = .018) and Treatment model classification (P = .038).

Next, to assess whether the brain metastasis prediction ability of the three-miRNA signature was independent of other clinical or pathological factors of patients with LUAD, multivariate analysis was performed in the 77 patients by the stepwise variable selection method. The result showed that the three-miRNA signature was the most crucial factor among these eight significant variables according


A
8 hsa-miRNAs
label
42 samples
BM+
BM-
B
Test data
OOB
Train data
Error rate
No.trees
C
8 hsa-miRNAs
7 hsa-miRNAs
6 hsa-miRNAs
3 hsa-miRNAs
Discard one less important hsa-miRNA at each step
Random forest supervised classification algorithm
Classification tree-1 for bootstrap sample
Classification tree-2 for bootstrap sample
Classification tree-210 for bootstrap sample
Estimate the important score for each hsa-miRNA using the out-of-bag samples by permutation testing
D
Class Center in random Forest supervised classification algorithm
BM- samples
BM+ samples
3 hsa-miRNAs
g1
g2
g3
g1g2
g1g3
g2g3
g1g2g3
All 2^3-1=7combinations of the 3 miRNAs
assign sample x to BM- group
assign sample x to BM+ group
Classification tree 1
Classification tree 2
Classification tree t
E
g1: hsa-miR-214
g2: hsa-miR-210
g3: hsa-miR-15a
LOOCV accuracy
hsa-miRNA combination
F
LOOCV Accuracy
Machine learning methods
Random forest
Fisher discriminant
Softmax regression
Support vector machine
Bagging
AdaBoost
Decision tree

to mean decrease accuracy or mean decrease Gini in the random Forest analysis (Figure 4*B*). Also, the three-miRNA signature was significantly correlated with BM+ patients' discrimination from the multivariate logistic regression analysis (OR = 121.156, 95%CI: 15.247–962.751, P = .000; Figure 4*B*), which showed that it might be a high-risk factor of brain metastasis for patients with LUAD. The multivariate analysis thus demonstrated that the predictive ability of this three-miRNA signature was independent of other clinical factors for the brain metastasis of patients with LUAD.

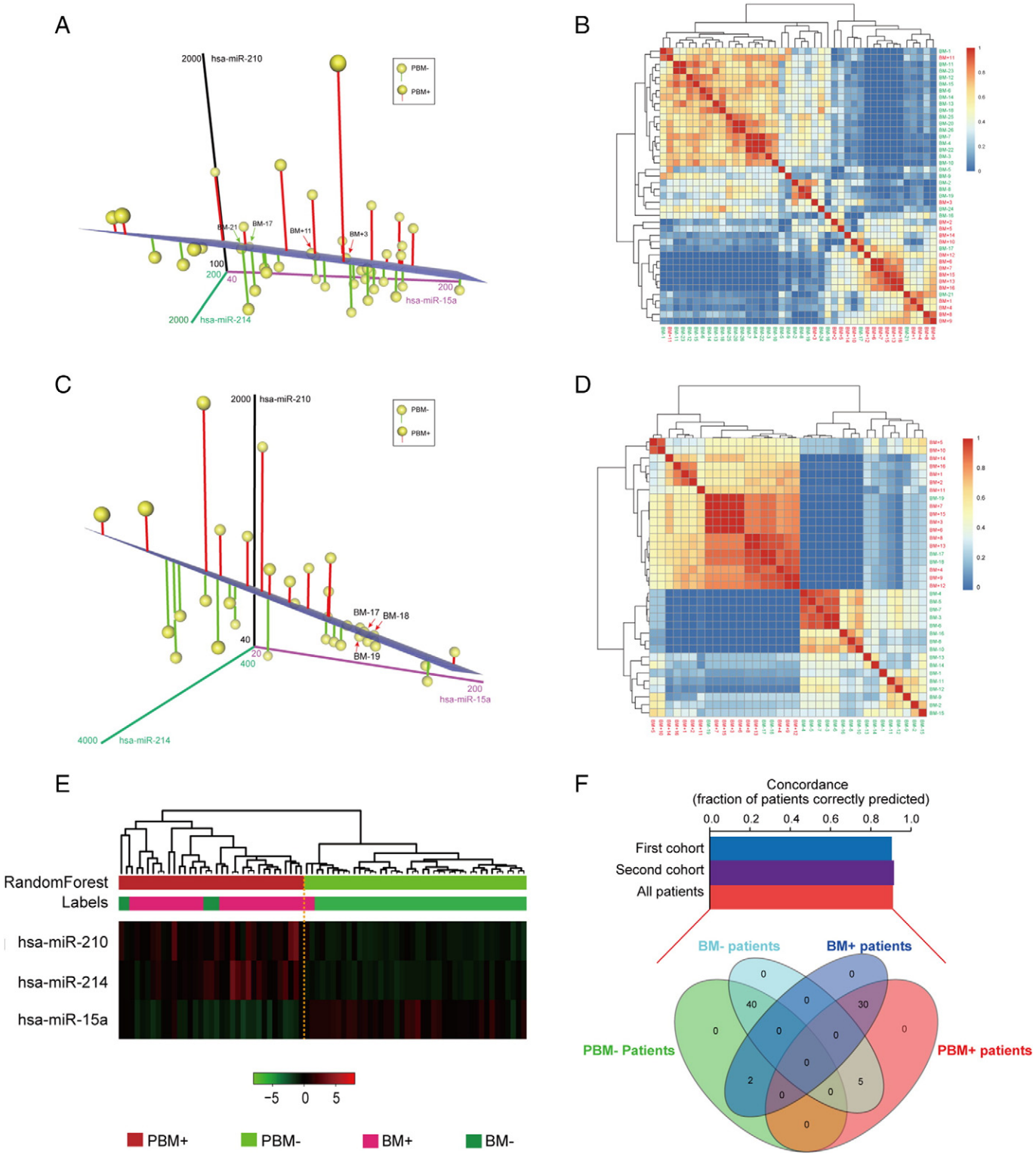
The Three-miRNA Signature was Closely Associated with Clinical Profile of LUAD Brain Metastasis

To determine whether the three-miRNA signature was singly clinical relevant with the brain metastasis, a different expression analysis of these three miRNAs was performed between metastatic and non-metastatic group in 77 patients with LUAD. In line with the previous result from the PBM or BM analysis, the expression value of these three miRNAs were all significantly different between FBM– or SBM– and PBM+ or SBM+ group patients ($P < .05$, Figure 5A). However, the similar result was not discovered in the other disseminated metastasis analysis. The data suggested that the three-miRNA signature could be exclusive in the brain metastasis of patients with LUAD.

Some studies reported that the survival of LUAD patients with brain metastasis was poor than non-brain metastasis [15]. To validate the close correlation between the three-miRNA signature and brain metastasis, a further survival analysis was carried out to evaluate whether the three-miRNA signature could predict survival of patients with LUAD. The results showed that the expression value of the three miRNAs was significantly different between the two groups of LUAD patients divided by distinct survival status ($P < .05$, Figure 5*A*). Then the log-rank test of the combined patients showed that the signature could significantly stratify these 77 patients with LUAD into high (PBM+) or low-risk (PBM–) group in OS (P = .009), DFS (P = .017), FBM (P = 3.19e-06) or SBM analysis (P = .0006) by the three-miRNA signature model with random Forest algorithm (Figure 5*F–I*), which was similar with the raw BM classification method in the predictive power (Figure 5*B–E*). In detail, the OS analysis showed that these LUAD patients were divided into PBM+ (n = 35) and PBM– (n = 42) with significantly different survival rates (P = .009, Figure 5*F*) in the random Forest classification. Moreover, the three-miRNA signature was significantly associated with survival of PBM+ LUAD patients (HR = 2.103, 95CI: 1.202–3.680, $P < .01$; Figure 5*F*), and the median survival of PBM+ patients was 34.0 months, which was significantly lower than those of PBM– patients with median survival 62.1 months ($P < .01$; Figure 5*F*). Analogously, the OS was significantly different when LUAD patients were stratified by the raw classified method (median survival months: 34.3 for PBM+ vs. 54.1 for PBM–; HR = 2.020, 95%CI: 1.151–3.544; P = .014, Figure 5*B*). Next, another three analyses including DFS, FBM and SBM analysis were adopted to confirm the reliability of this model in predicting survival and brain metastasis, and there were no different comes of survival analysis in these patients divided by raw BM classified method (Figure 5*C–E*) or random Forest algorithm (Figure 5*G–I*). In brief, the LUAD patients with BM+ or BM– subtypes predicted by the three-miRNA signature model, did not differ significantly with raw classification in terms of clinical survival status at presentation. Together with prior studies, these data established that brain metastasis influenced the clinical outcomes and the three-miRNA signature played a powerful role in predicting the brain metastasis classification.

To further verify that the association between the three-miRNA signature and the brain metastasis, the three miRNAs from this signature and another six miRNAs (hsa-miR-146a [16], hsa-miR-21 [17], hsa-miR-145 [18], hsa-miR-378 [19], hsa-miR-330/3p [20], hsa-miR-197[21]) reported to be correlated with brain metastasis were compared with unsupervised hierarchical clustering method in 77 patients with LUAD. Hierarchical clustering based on Spearman correlation clearly divided these nine miRNAs into three groups (Figure 5*J*), which showed that hsa-miR-146a was significantly associated with hsa-miR-15a (r = 0.63, $P < .001$; Figure 5*L*) in Cluster1 and another two miRNAs (hsa-miR-21, hsa-miR-378) were closely adjacent to hsa-miR-210 and hsa-miR-214 ($r \geq 0.32$, $P < .01$; Figure 5*L*) in Cluster3. Then the expression values of these genes were evaluated to show that hsa-miR-21 and hsa-miR-378 had a significant higher expression in BM+ cases as compared with BM– patients ($P < .05$), which agreed with the results from hsa-miR-210 and hsa-miR-214 outlined above (Figure 5*K*). Inversely, the gene expression of hsa-miR-146a was significantly down-regulated in the BM+ compared to the BM– group ($P < .01$), which was also in line with the hsa-miR-15a expression distribution (Figure 5*K*). Interestingly, the expression of hsa-miR-146a was significantly negative correlation with hsa-miR-21 and hsa-miR-378 ($r \leq -0.40$, $P < .001$), which was similar with the relationship between hsa-miR-15a and the other two miRNAs (hsa-miR-210 and hsa-miR-214) in this signature (Figure 5*L*). These results indicated that this signature was closely associated with brain metastasis in patients with LUAD.

Figure 2. Identification of the miRNA signature and selection of the machine learning algorithm in the training set. (**A**) Through microarray processing, the data was described by an 8 × 42 matrix with a 'BM+' or 'BM−' label column. (**B**) The optimal number of trees was selected in the random Forest algorithm, the line with orange color represents a cutoff value. (**C**) Selection process for the three miRNAs with highest classification power for BM prediction. A random Forest supervised classification algorithm was used to narrow down the number of miRNAs by several iterative steps, in which the least important miRNA was discarded at each step according to their importance score. (**D**) Development of BM classifier with the three-miRNA signature from all combinations ($n = 2^3–1 = 7$) using the Class Center Method. V_n and V_p are the mean expression profiles of the miRNA combination ($g_1g_2g_3$) for BM− patients and BM+ patients, respectively. V_x is the expression profiles of patient x. The Euclid distance $d(V_x V_n)$ and $d(V_x V_p)$ are used to classify patient x into a BM− or BM+ group. For each case x in the training set, the nearest neighbors are found. And then, for each class in one classification tree, the case x which has the most neighbors of that class y is identified as the maximum probability P, t represents the classification tree number. The final classification for each case x is defined by the average value of the maximum probability P of that class y from all the classification trees (n = 210), T represents the number of BM labels. (**E**) The procedure for identifying the final signature. The LOOCV accuracies of all seven combinations were calculated and shown in the plot. The signature containing three miRNAs (hsa-miR-210, hsa-miR-214 and hsa-miR-15a) was selected as the final signature. (**F**) Histogram of the classified accuracies of seven machine learning methods with LOOCV in the training set. The classified accuracy of random Forest algorithm was the highest within the three-miRNA signature.



The Three-miRNA Signature had Strongly Diagnostic Power for Predicting Brain Metastasis of LUAD Patients

To identify the diagnostic value in BM subtypes, the ROC analysis was applied to compare the power between the six significant miRNAs and this signature in 77 LUAD patients. As a result, the area under the curve (AUC) values from the three miRNAs of this signature were much higher (0.859 for hsa-miR-15a, 0.886 for hsa-miR-210, 0.883 for hsa-miR-214, $P < .0001$) than the other three miRNAs (0.685 for hsa-miR-146a, 0.702 for hsa-miR-378, 0.662 for hsa-miR-21, $P < .05$; Figure 6*A*) correlated with brain metastasis reported previously. More important, the AUC of this signature was 0.913 ($P = .000$), which was much more than any single AUC value from the six miRNAs (Figure 6*A*). This result demonstrated that the three-miRNA signature model could predict

the BM subtypes with high sensitivity and specificity, and the new model could sufficient to work as a practical clinical tool in the current therapeutic era.

From the ROC curve analysis, the predictive power of this signature was identified to assign cases into BM+ or BM− classification. To further discover the distribution of the performance of each marker, radar map method was used based on the normalized expression value (Figure 6*B*). The result showed that hsa-miR-210 and hsa-miR-214 were distributed in BM+ subtype patients (Figure 6*B*) with higher AUC value (AUC> 0.88, $P = .000$; Figure 6*C*); meanwhile the AUC value of hsa-miR-15a was more in BM− group (AUC> 0.85, $P = .000$; Figure 6*B*and *C*). Generally, the three-miRNA signature comprised of hsa-miR-210, hsa-miR-214 and hsa-miR-15a had high diagnostic power in the brain metastasis classification.

Functional Enrichment Analysis of Genes Correlated with the Signature miRNAs

To explore the potential role of the miRNAs from this signature in regulating the LUAD brain metastasis, the correlation between their expression values and those of their targeted mRNAs was examined in the database of Diana Tools. The expression level of 2914 protein-coding genes (PCGs) was positively targeted by that of at least one of the three miRNAs in this signature. Then the functional enrichment analysis was performed on GO terms and KEGG pathways for these PCGs co-expressed with these three miRNAs (Supplementary text). And the result of GO term (Supplementary Fig. 2A) revealed that PCGs clustered most significantly in cell biosynthetic process, metabolic process, and mitotic cell cycle. The same analysis of KEGG pathway (Supplementary Fig. 2B) demonstrated the signaling pathways positively correlated with NSCLC carcinogenesis and cancer metastasis (TGF-β signaling pathway [22] and P53 signaling pathway [23]). These data suggested that the three-miRNA signature might positively or negatively regulate these significant PCGs to affect the development of brain metastasis in patients with LUAD. However, these discoveries should be verified by bio-experimentation.

Discussion

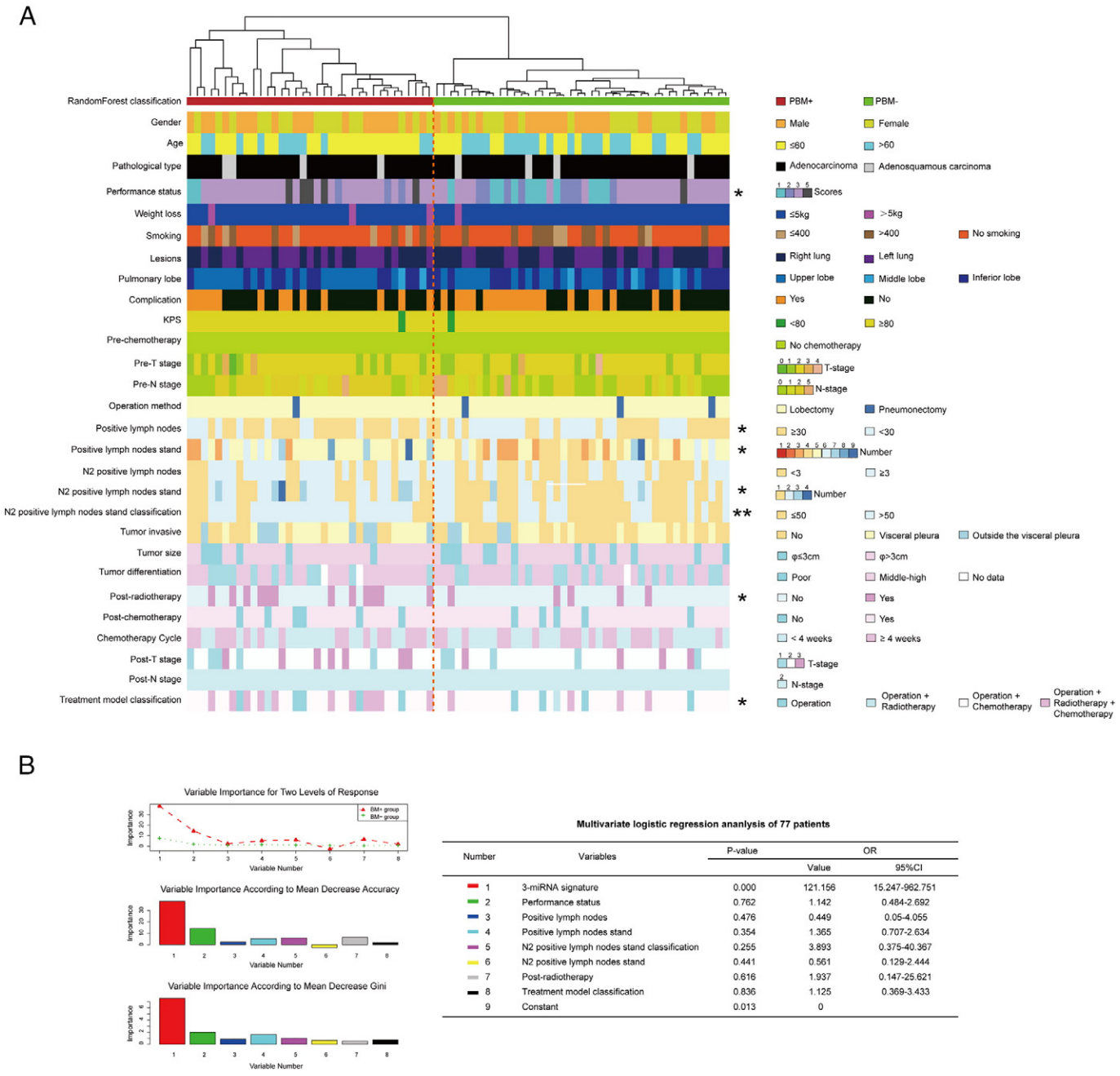
In recent years, the research on the roles of miRNAs in cancer progress were gradually increased [24–26]. However, the involvement of miRNAs in the LUAD brain metastasis prediction model developed by random Forest algorithm has not been reported. Here, the study on differential miRNA expression was presented in two cohorts and a three-miRNA signature was discovered to be powerful predictors for brain metastasis of patients with cancer. This signature identified in the training group showed similar predictive power in the test cohort. Although there are currently no other LUAD brain metastasis data sets with both mRNA and miRNA expression data publicly available that would allow further validation of miRNA-based BM classification, we still believe that the predictive value of this signature has a solid basis in LUAD patients. This is a pioneering research of the correlation between the machine learning model developed by miRNAs signature and brain metastasis of patients with LUAD.

In this study, to void the common 'curse-of-dimensionality' problem, a total of 330 miRNAs differentially expressed between BM+ and BM− samples were filtered out and then subjected to random Forest supervised classification to further narrow down the number of miRNAs correlated with brain metastasis. To reduce the predictive error rates, the random sampling and ensemble strategies were used in random Forest classification for the 'curse-of-dimensionality' datasets. And then the measures of gene importance were applied to filter the original gene set iteratively in the random Forest classification, resulting in superior performance in feature screening.

Next, a classifier was developed for each combination of the three selected miRNAs using the Class Centered Method. We agree that if more genes were selected, even with some redundancy, the predictive model might perform a better role in these patients with cancer. However, it would be greater to have fewer genes as possible to be analyzed to make the new model more competitive. For these reasons and the rule of Occam's razor, the three-miRNAs signature was selected as the final signature. And the result validated that this model, as the 'less-gene-possible' combination, could divide the brain metastasis subtypes of LUAD effectively.

In addition, to further demonstrate the independence of this new model in brain metastasis prediction, the association between this signature and the basic clinical characteristics was examined and identified 7 variables as significantly candidate predictive factors in the combined dataset. Then multivariate logistic regression analysis was performed and identified this signature as an independent biomarker. Moreover, this signature was closely associated with the brain metastasis through analyzing the clinical profiles of various metastasis. Because the brain metastasis negatively influenced the clinical outcome of patients with LUAD [15], the prognostic value of this signature was analyzed to further verify its significant association with the brain metastasis. The result proved again that this signature could stratify the patients with LUAD into BM+ and BM− group more effectively compared with the performance of the single miRNA.

Figure 3. Computational classification of BM based on the three-miRNA signature in patients with LUAD from the training set, the validated set and the combined set. (**A&C**) The 3D Scatter Plot shows the classification of BM+ versus BM− based on the profiles of three miRNAs (hsa-miR-210, hsa-miR-214 and hsa-miR-15a) in the training set n = 42 (**A**) and the validated set n = 35 (**C**). A case is classified into some group depending on whether its miRNA levels place it below or above the separation plane. The blue plane illustrates the automatically generated classifier. The dots with red or green vertical lines represent predictive BM+ (PBM+) or BM− (PBM−), respectively. The four mismatches of 42 patients highlight with red or green arrows. **(B&D)** Supervised random Forest clustering of the patients in the training set n = 42 (**B**) and the validated set n = 35 (**D**). The similarity coefficients of the three miRNAs in the two cohorts were used for clustering analysis. This clustering clearly divided the cases into BM+ (red color) and BM− (green color) group. Only four and three samples were misclassified in the training set and the validated set, respectively. (**E**) Heatmap shows the supervised random Forest classification of patients in the combined set (n = 77). The three significant miRNAs classified these patients into PBM+ and PBM − group, with seven mismatches out of 77 patients. Labels stand for the primary classification including BM+ and BM−. (**F**) The bar chart shows concordance estimates for BM prediction with the use of three miRNAs selected from this signature in training set (concordance, approximately 90.5%), validated set (concordance, approximately 91.4%) and the combined set (concordance, approximately 90.9%). The Venn diagram shows the overlaps among the BM+, BM−, PBM+ and PBM− subgroups in the combined set.



Multivariate logistic regression ananlysis of 77 patients

Number	Variables	P-value	OR	
			Value	95%CI
1	3-miRNA signature	0.000	121.156	15.247-962.751
2	Performance status	0.762	1.142	0.484-2.692
3	Positive lymph nodes	0.476	0.449	0.05-4.055
4	Positive lymph nodes stand	0.354	1.365	0.707-2.634
5	N2 positive lymph nodes stand classification	0.255	3.893	0.375-40.367
6	N2 positive lymph nodes stand	0.441	0.561	0.129-2.444
7	Post-radiotherapy	0.616	1.937	0.147-25.621
8	Treatment model classification	0.836	1.125	0.369-3.433
9	Constant	0.013	0	

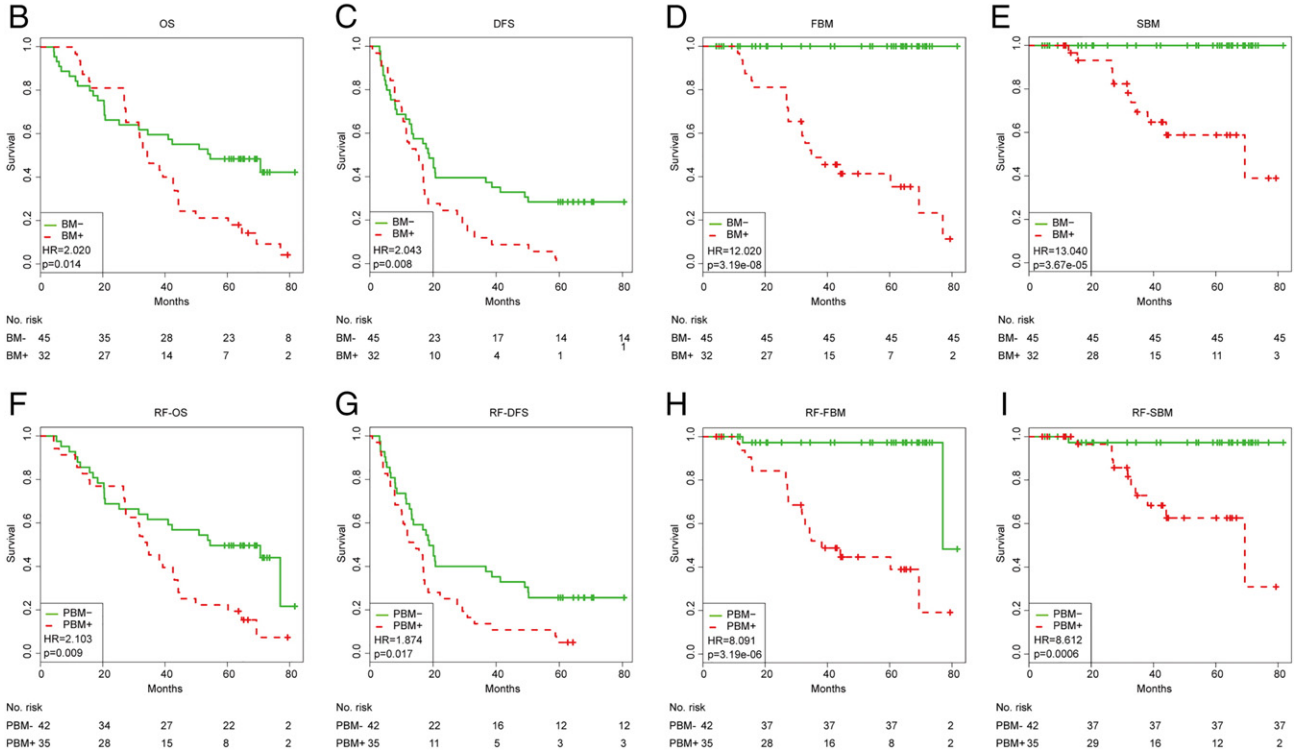
Figure 4. The correlation analysis between the three-miRNA signature and all clinical characteristics. (**A**) Univariate analysis of all clinical characteristics between PBM+ and PBM− subtypes. BM subtypes appear in columns, and the clinical features are displayed in rows. Categorical features analyzed using chi-square test or Fisher's exact tests; continuous features were analyzed using Kruskal-Wallis tests, * indicates $P < .05$, ** represents $P < .01$. Selected differently expressed feature labels are displayed on the right of each subtype. (**B**) Evaluate the importance of this three-miRNA signature and seven significant factors from univariate analysis with random Forest algorithm. At the top of this plot, the important score of each variable is shown between the BM+ and BM− group. In the middle portion of this plot, the importance of each variable is measured according to Mean Decrease Accuracy. At the bottom of this plot, the important scores of these eight variables were evaluated according to Mean Decrease Gini. (**C**) Multivariate logistic regression analysis was performed in the 77 LUAD patients. The *P*-value and Odds Ratio of these eight factors were displayed in the table. $P < .05$ is statistically significant.

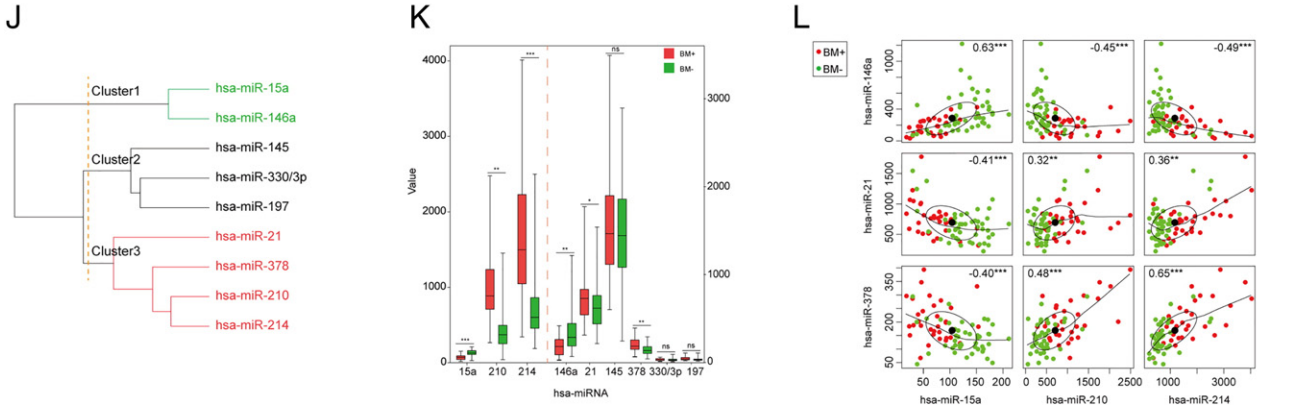
The functions of most miRNAs are not yet annotated. However, the possible function of the miRNAs in LUAD could be inferred by the mRNA co-expression data from the same cohort of patients. The targeted genes of each miRNA in this signature were selected to undergo gene set enrichment analysis (GSEA). And the bioinformatics analysis was performed to identify the correlated biological process and pathways by integrative analysis of miRNAs and PCGs. It is a plausible inference from the result that this signature may be involved in NSCLC carcinogenesis and metastasis-related biological process and pathways.

Also, it remains unknown whether it has predictive value in patients with the other stage as the miRNA signature was derived from IIIA-N2 patients with LUAD. Another limitation of our study is that the test set is too small to allow for a sensible assessment of the

A

Clinical profile	Group 1		Group 2		hsa-miRNA expression (Group1 vs. Group2) P - value			
	Status	Number	Status	Number	15a	210	214	Legend
Supraclavicular lymph node metastasis	No	71	Yes	6				
Abdominal Metastasis	No	74	Yes	3				
Liver Metastasis	No	71	Yes	6				
Adrenal Metastasis	No	70	Yes	7				
Lung Metastasis	No	45	Yes	32				
Bone Metastasis	No	56	Yes	21				
Single Brain Metastasis	No	66	Yes	11				
First Brain Metastasis	No	56	Yes	21				
Brain Metastasis	Negative	45	Positive	32				
Predictive Brain Metastasis	Negative	42	Positive	35				
Cancer Associated with Dead	No	38	Yes	39				
Disease-Free Survival	No	13	Yes	64				
Overall Survival	Alive	24	Dead	53				





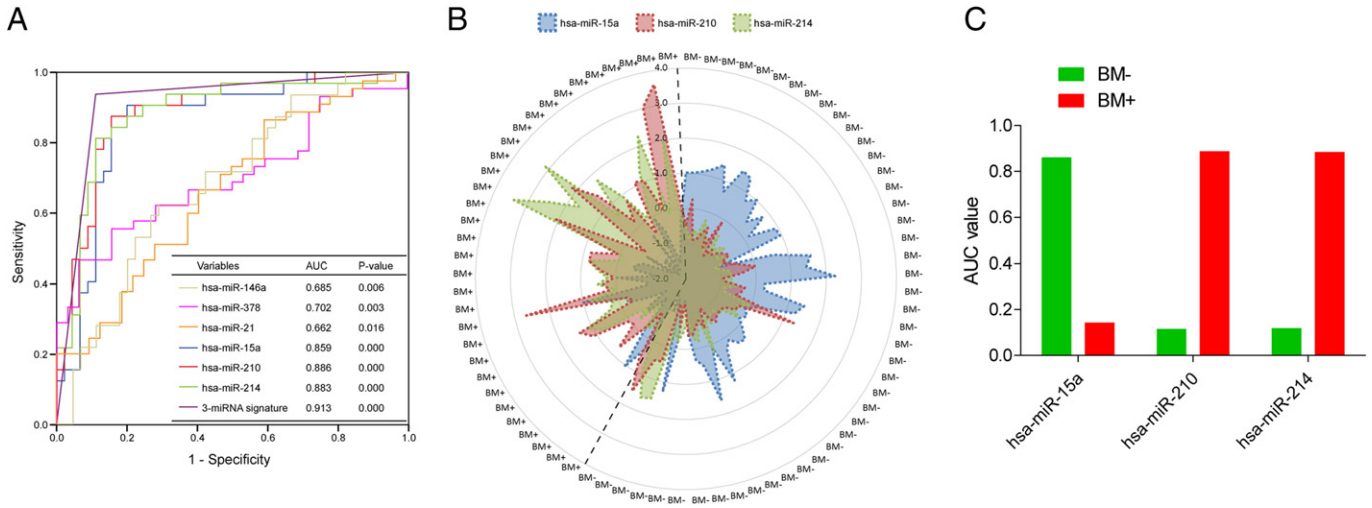


Figure 6. Access the discriminatory accuracy of the three-miRNA signature and each miRNA marker with ROC curve analysis. (**A**) The ROC curve analysis evaluated the discriminant power with AUC area of the signature and each miRNA. (**B**) Radar map was used for analyzing the distribution of the expression profiles of these three miRNAs from this signature in assigning cases to BM+ or BM− subgroup. (**C**) The histogram showed the performances of the three miRNAs quantified by the AUC value between BM+ and BM− subgroup.

generalizability of the three-miRNA signature. Data sets from other institutes and other countries are still necessary to validate its generalization ability. And its validity should be further confirmed in the prospective cohorts. In addition, the further efforts in the next study will be paid to verify the discoveries about the expression and function of these miRNAs with modern empirical method of molecular biology.

In conclusion, our study identified some miRNAs that were significantly changed between BM+ and BM− LUAD tissues. The three-miRNA signature we discovered independently and robustly predicted the brain metastasis of patients with LUAD. Furthermore, this signature could predict the brain metastasis of LUAD patients with high sensitivity and specificity. To our knowledge, it is the first random Forest model developed by the miRNA signature to predict the brain metastasis of patients with LUAD. And further validation researches in prospective cohorts and in patients with perfect adjuvant therapy information from different medical centers are required to verify the diagnostic value of this signature before its application into the clinical practice.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tranon.2017.12.002.

Acknowledgements

We thank Dr. Rong Xiang (Department of Immunology, School of Medicine, Nankai University, China) for assistance in writing the paper and provided important advices.

Contribution

S.Z. performed the biostatistics analysis and constituted the prediction model and prepared the manuscript, Y.J. collected the data samples and repaired the manuscript, L. W. designed the experiments and decided to publish the paper.

References

[1] Siegel RL, Miller KD, and Jemal A (2015). Cancer statistics, 2015. *CA Cancer J Clin* **65**(1), 5–29.

[2] Mao Y, Yang D, He J, and Krasna MJ (2016). Epidemiology of lung cancer. *Surg Oncol Clin N Am* **25**(3), 439–445.

Figure 5. The association analysis between the three-miRNA signature and the brain metastasis. (**A**) Evaluate the different expression of the three miRNAs among the different clinical profiles including metastasis and survival. The patients were divided into two groups according to their status of metastasis or survival. And the statistical results of these three miRNAs in each clinical profile were represented with distinct colors. $P < .05$ is statistically significant. (**B-I**) Overall survival (OS) analysis of 77 patients when stratified by the raw BM classified methods and the random Forest algorithm. Kaplan–Meier survival curves were displayed between BM+ and BM− subgroup in OS analysis (**B**), disease-free survival (DFS) analysis (**C**), first brain metastasis (FBM) analysis (**D**) and single brain metastasis (SBM) analysis (**E**). And the other Kaplan–Meier survival curves were shown between PBM+ and PBM− subgroup in random Forest (RF)-OS analysis (**F**), RF-DFS analysis (**G**), RF-FBM analysis (**H**) and RF-SBM analysis (**I**). (**J**) Unsupervised hierarchical clustering of nine miRNAs in the combined set. The correlation coefficients of these miRNAs in 77 patients were used for clustering analysis. Hierarchical clustering clearly separated these miRNAs into three subgroups with red, black and green colors. (**K**) Different expression analysis of these miRNAs was displayed between BM+ and BM− subgroup in 77 patients. Expression levels were presented as boxplots and compared using *t*-test method. (**L**) Correlation analysis between the signature miRNAs and another three miRNAs which were significant between BM+ and BM− subgroup. In each scatterplot, x and y axis represented the expression value of a miRNA, respectively. ** represents $P < .01$, *** represents $P < .001$.

[3] Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ, and He J (2016). Cancer statistics in China, 2015. *CA Cancer J Clin* **66**(2), 115–132.

[4] Chen Z, Fillmore CM, Hammerman PS, Kim CF, and Wong K-K (2014). Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat Rev Cancer* **14**(8), 535–546.

[5] Grinberg-Rashi H, Ofek E, Perelman M, Skarda J, Yaron P, Hajdúch M, Jacob-Hirsch J, Amariglio N, Krupsky M, and Simansky DA (2009). The expression of three genes in primary non–small cell lung cancer is associated with metastatic spread to the brain. *Clin Cancer Res* **15**(5), 1755–1761.

[6] Laack NN and Brown PD (2004). Cognitive sequelae of brain radiation in adults. Seminars in oncology. Elsevier; 2004. p. 702–713.

[7] Oh Y, Taylor S, Bekele BN, Debnam JM, Allen PK, Suki D, Sawaya R, Komaki R, Stewart DJ, and Karp DD (2009). Number of metastatic sites is a strong predictor of survival in patients with nonsmall cell lung cancer with or without brain metastases. *Cancer* **115**(13), 2930–2938.

[8] Bartel DP (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**(2), 281–297.

[9] Lai EC (2002). Micro RNAs are complementary to 3 [variant prime] UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* **30**(4), 363.

[10] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, and Ferrando AA (2005). MicroRNA expression profiles classify human cancers nature. **435**(7043), 834–838.

[11] Liu C-G, Calin GA, Volinia S, and Croce CM (2008). MicroRNA expression profiling using microarrays. *Nat Protoc* **3**(4), 563–578.

[12] Daugaard I, Venø MT, Yan Y, Kjeldsen TE, Lamy P, Hager H, Kjems J, and Hansen LL (2017). Small RNA sequencing reveals metastasis-related microRNAs in lung adenocarcinoma. *Oncotarget* **8**(16), 27047.

[13] Long H, Wang Z, Chen J, Xiang T, Li Q, Diao X, and Zhu B (2015). microRNA-214 promotes epithelial-mesenchymal transition and metastasis in lung adenocarcinoma by targeting the suppressor-of-fused protein (Sufu). *Oncotarget* **6**, 38705–38736.

[14] Yang T, Thakur A, Chen T, Yang L, Lei G, Liang Y, Zhang S, Ren H, and Chen M (2015). MicroRNA-15a induces cell apoptosis and inhibits metastasis by targeting BCL2L2 in non-small cell lung cancer. *Tumor Biol* **36**(6), 4357–4365.

[15] Sørensen J, Hansen H, Hansen M, and Dombernowsky P (1988). Brain metastases in adenocarcinoma of the lung: frequency, risk groups, and prognosis. *J Clin Oncol* **6**(9), 1474–1480.

[16] Hwang SJ, Seol HJ, Park YM, Kim KH, Gorospe M, Nam D-H, and Kim HH (2012). MicroRNA-146a suppresses metastatic activity in brain metastasis. *Mol Cells* , 1–6.

[17] Singh M, Garg N, Venugopal C, Hallett R, Tokar T, McFarlane N, Mahendram S, Bakhshinyan D, Manoranjan B, and Vora P (2015). STAT3 pathway regulates lung-derived brain metastasis initiating cell capacity through miR-21 activation. *Oncotarget* **6**(29), 27461.

[18] Zhao C, Xu Y, Zhang Y, Tan W, Xue J, Yang Z, Zhang Y, Lu Y, and Hu X (2013). Downregulation of miR-145 contributes to lung adenocarcinoma cell growth to form brain metastases. *Oncol Rep* **30**(5), 2027–2034.

[19] L-t Chen, Xu S-d, Xu H, Zhang J-f, Ning J-f, and Wang S-f (2012). MicroRNA-378 is associated with non-small cell lung cancer brain metastasis by promoting cell migration, invasion and tumor angiogenesis. *Med Oncol* **29**(3), 1673–1680.

[20] Wei C-H, Wu G, Cai Q, Gao X-C, Tong F, Zhou R, Zhang R-G, Dong J-H, Hu Y, and Dong X-R (2017). MicroRNA-330-3p promotes cell invasion and metastasis in non-small cell lung cancer through GRIA3 by activating MAPK/ERK signaling pathway. *J Hematol Oncol* **10**(1), 125.

[21] Remon J, Alvarez-Berdugo D, Majem M, Moran T, Reguart N, and Lianes P (2016). miRNA-197 and miRNA-184 are associated with brain metastasis in EGFR-mutant lung cancers. *Clin Transl Oncol* **18**(2), 153–159.

[22] Kudinov AE, Deneka A, Nikonova AS, Beck TN, Ahn Y-H, Liu X, Martinez CF, Schultz FA, Reynolds S, and Yang D-H (2016). Musashi-2 (MSI2) supports TGF-β signaling and inhibits claudins to promote non-small cell lung cancer (NSCLC) metastasis. *Proc Natl Acad Sci* **113**(25), 6955–6960.

[23] Tang D, Yue L, Yao R, Zhou L, Yang Y, Lu L, and Gao W (2017). P53 prevent tumor invasion and metastasis by down-regulating IDO in lung cancer. *Oncotarget* **8**(33), 54548–54557.

[24] Garzon R, Marcucci G, and Croce CM (2010). Targeting microRNAs in cancer: rationale, strategies and challenges. *Nat Rev Drug Discov* **9**(10), 775–789.

[25] Calin GA and Croce CM (2006). MicroRNA signatures in human cancers. *Nat Rev Cancer* **6**(11), 857–866.

[26] Bracken CP, Scott HS, and Goodall GJ (2016). A network-biology perspective of microRNA function and dysfunction in cancer. *Nat Rev Genet* **17**(12), 719–732.