Machine Learning Identifies the miR-196b and miR-34c-5p as the Chemotherapy Response Biomarkers of Lung Adenocarcinoma

Yuanmin Lan

Shenzhen Longgang People's Hospital

Jiqing Xue

Shenzhen Longgang People's Hospital

Liang Chen

Shenzhen Longgang People's Hospital

Dongwei Wu

Shenzhen Longgang People's Hospital

Wei Wang

Shenzhen Longgang People's Hospital

Kezhen Shen

Shenzhen Longgang People's Hospital

Piji Chen (✉ piji.chen@outlook.com)

Shenzhen Yantian People's Hospital <https://orcid.org/0000-0002-7632-7654>

Research article

Keywords: Machine Learning, miRNA, chemotherapy, Platinum, lung adenocarcinoma, liquid biopsy

Posted Date: August 19th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-54016/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Lung adenocarcinoma is one of the most common malignant lung cancers. Although platinum-based chemotherapy is the first-line adjuvant treatment for middle and late stage lung adenocarcinoma, the response of chemotherapy varies between patients. Moreover, there are no effective biomarkers that could predict chemotherapy response in clinical practice. MiRNAs that are stable in all types of body fluid have demonstrated their diagnostic and prognostic capacity in variety of cancers. Here, we utilized three different machine learning algorithms to identify miRNA signatures specific to chemotherapy response in lung cancer.

Methods

Through a public dataset, a panel of miRNAs for response to chemotherapy was identified by Machine Learning. The predictive capacity was determined by the receiver operation curve. A cohort involving 30 patients with lung adenocarcinoma was utilized for validate the miRNA panel.

Results

Machine Learnings identified five chemotherapy response featured miRNAs (miR-196b, miR-34c-5p, miR-181b, miR-27b and miR-26a). The putative targets of these miRNA signatures are enriched in the biosynthesis. Two of these miRNA signatures (miR-196b and miR-34c-5p) were validated for their chemotherapy response prediction in our 30 serum samples of lung adenocarcinoma with the accuracy of 0.90 and 0.93, respectively.

Conclusions

Our study demonstrates circulating miRNA could potentially be predictive biomarker for chemotherapy response in lung adenocarcinoma.

Background

Lung cancer is one of the most common malignant tumors in the world with the top morbidity and mortality of all malignant tumors (1). Lung adenocarcinoma accounts for 80–85% of all lung cancers (2). Since most lung adenocarcinoma patients are diagnosed as advanced, chemotherapy and radiation therapy have become the main treatment methods for such patients. Platinum chemotherapy is the standard first-line chemotherapy regimen for lung adenocarcinoma (3). However, only a few patients can benefit from the platinum-based chemotherapy and there are very few biomarkers that can predict a patient's response to conventional chemotherapy drugs (4). Early treatment of lung adenocarcinoma involves multiple predictive response methods, including imaging studies performed alone or in combination with specific biomarkers (5–7). Due to the lack of sufficient specificity and sensitivity, the predictive biomarkers of lung adenocarcinoma chemotherapy response are not ideal in clinical practice

(8). Therefore, new biomarkers of lung adenocarcinoma progression in chemotherapy patients need to be investigated.

Although most biomarker research in cancer research utilized differential expression methods for biomarker identification, the application of predictive models and data mining in machine learning methods are increasing for biomarker discovery (9). Specifically, tree-based classification and regression have been widely used for the diagnosis and prognosis of cancer. Although it is still difficult to explain how their predictive models work and interpret how the features are relevant to the biology in diseases, the use of tree-based machine learning algorithms has the advantages of rule extraction and clinical interpretability (10).

Previous studies have confirmed that anti-tumor chemotherapeutic drugs leads to the changes of microRNA (miRNA) expression profiles in various malignant tumor cells, suggesting that miRNA expression is associated with the response to the chemotherapy (11). This possibility could contribute to new ways to treat lung cancer using miRNA expression features. MiRNA is one of the most studied molecules with the potent to be biomarkers. Dysregulated miRNA is a common feature of all types of cancer and is associated with its development, progression, and response to treatment. In addition, miRNAs are stable in human body, making it easy to detect their expression in tissues and body fluids. These characteristics of miRNAs make them potentially useful diagnostic, prognostic, and predictive biomarkers for cancer research (12). A few biomarkers studies have demonstrated early diagnosis and progression could be predicted by tissue or blood miRNAs expression (13).

Here, we used a tree-based machine learning algorithm on a public dataset for lung adenocarcinoma patients to discover miRNA biomarkers specific to the chemotherapy response. In addition, we verified the chemotherapy response miRNA features in our independent blood cohort and confirmed circulating miR-196b and miR-34c-5p are potentially the chemotherapy response biomarkers.

Methods

All calculations and plotting with regard to data retrieval, processing, filtering, normalization, differential expression, and applying and evaluating machine learning algorithms were performed in R environment (version 3.5.3) with the supports of R/bioconductor packages.

Data retrieval and exploratory data analysis

We queried “lung adenocarcinoma” and “platinum chemotherapy” in Gene Expression Omnibus (GEO) and downloaded the dataset GSE56264 for analysis through “GEOquery” (14). A total of 40 samples from lung adenocarcinoma patients undergoing the platinum-based chemotherapy were included in this dataset. According to the manifest table provided by the original researcher, twenty-four samples were categorized into chemotherapy non-responders while 16 samples were chemotherapy responders. It contained a total of 1205 miRNA probes information.

For investigating the batch effect, we performed the boxplots. Principal Component Analysis (PCA) was conducted to test the variation of samples presented by the miRNA expression. We removed the miRNA probes containing 0 or missing values that presented in more than 10 samples. Quantile normalization was performed to remove the batch effect.

Differential expression analysis was performed to investigate the feature miRNAs by “limma” package (15). Volcano plot was utilized to visualize the differentially expressed miRNAs. The absolute value of log fold change larger than 2 and FDR less than 0.05 were regarded as the threshold. Violin plots were utilized to visualize the single miRNA expression of each sample.

Machine learning model training, evaluation and plotting

Tree-based machine learning algorithms were applied using caret and randomForest packages (Liaw and Wiener, 2002). Importance of variables from random forest model was extracted and plotted by randomForest packages.

Another three machine learning models (Naïve Bayes, Latent Dirichlet Allocation and Supportive Vector Machine) were utilized to evaluate the miRNA features determined from random forest model. The input parameters were the miRNA signatures determined by random forest model. The prediction accuracy of models was determined by the comparison of predicted cases and the true cases in the GSE56264 dataset. The best number of miRNA signatures was selected as all three models showed the highest prediction accuracy.

The putative target pathway of miRNA biomarkers

The putative target pathway of miRNA signatures were determined by DIANA mirPath v.3 (<http://snf-515788.vm.okeanos.grnet.gr/>) (16). The putative target algorithm was set as TargetScan and the p value threshold was 0.05. Three terms (Biological Processes, Cellular Components and Molecular Function) of Gene Ontology and KEGG were queried, respectively.

Sample isolation

For patients with lung cancer, blood specimens were collected at 1 to 3 days before the first cycle of chemotherapy. Ten millilitres of peripheral venous blood of each patient were collected in an anticoagulant tube with ethylenediaminetetraacetic acid (EDTA). Subsequently, the blood samples were centrifuged in blood collection tubes for 10 minutes at 1900*g (3000 rpm) at 4 °C to remove leukocyte and red blood cells. The supernatant (serum) was then transferred to a clean 1.5 mL centrifuge tube, centrifuged at 16,000 *g at 4 °C for 10 minutes. These serum samples were stored at – 80 °C until analysis.

Serum RNA extraction

Total RNA was extracted and purified from 10 mL of serum aliquots using a QIAGEN miRNeasy Serum/Serum kit (catalog no.217184; QIAGEN, Germany). RNA quantification and quality determination were performed using an MD2000 Series spectrophotometer (Biofuture, Britain, UK). To eliminate

variance among the samples, U6 was utilized as a endogenous control. The primers for U6 were listed as followed: Forward, 5'-CTCGCTTCGGCAGCACATATACT-3'; Reverse, 5'- ACGCTTCACGAATTTGCGTGTC-3' (17).

Reverse transcription and real-time PCR analysis

We performed our qRT-PCR according to the manufacturer's instructions with minor adaption. Mir-X miRNA First-Strand Synthesis and SYBR Premix Ex Taq kits (Tli RNaseH Plus) (Takara Bio, China) were used for microRNA reverse transcription and RT-PCR, respectively. Specific and universal primer pairs (Takara Bio, China) were used to amplify microRNA. The qPCR reactions were incubated in 96-well plates at 95 °C for 3 minutes, followed by 45 cycles of amplification (95 °C for 5 seconds and 64 °C for 30 seconds). The primers sequence for miR-196b (Forward: 5'- TAGGTACCACTTTATCCCGTTCACCA - 3' and Reverse: 5'-ATCTCGAGGCAGGGAGAGAGGAATAA-3') were obtained from previous study (18). The primers sequence for miR-34c-5p (Forward: 5'- GCCTGTCACAACGTGTTGGGGT-3' and Reverse: 5'- ACCTGGCCGTGTGGTTAGTGA - 3') were also adapted from another study (19). MicroRNA expression was calculated using the $2^{-\Delta\Delta Ct}$ method and was subjected to statistical analysis. Double distilled H₂O was used as a negative control, and each sample was assayed in triplicate.

Statistical analysis

All statistical analysis was performed under R environment (version 3.5.3). Student's t test was conducted for two group comparison. Receiver Operation Curve (ROC) was plotted for the miRNA biomarker discrimination ability. Area under the curve (AUC) was used as the accuracy of ROC. A P value of < 0.05 or FDR < 0.05 was considered to be statistically significant.

Results

Differentially expressed miRNAs were identified by Random Forest algorithm.

For predicting the chemotherapy effectiveness by serum miRNAs, we utilized the GSE56264 datasets to filter out the chemotherapy feature miRNAs. After we got the miRNA candidate, we validated these miRNAs in an independent cohort containing 15 chemotherapy responders and 15 non-responders (Fig. 1).

To characterize the miRNA expression between platinum-based chemotherapy responders and non-responders, we first performed the differentially expressed miRNA analysis of the GSE56264 datasets. First, we checked the batch effect between samples by boxplot (**Supplementary Fig. 1**). Due to the large batch effect between samples, we filtered the miRNA probes containing 0 or missing value that presented in more than 10 samples. After filtration of no expression miRNA probes, we included 249 miRNAs in the quantile normalization. After cleaning and normalizing the dataset, we performed principal component analysis (PCA) of all samples in GSE56264 to explore the heterogeneity between responders and non-responders of chemotherapy with regard to miRNA expression. PCA demonstrated chemotherapy responders had different miRNA expression pattern as compared with non-responders (Fig. 2A,

Supplementary Fig. 2A). Differential expression analysis revealed eight miRNAs were significantly differently expressed between responders and non-responders (Fig. 2B). Notably, we failed to find the difference of miRNAs between responders and non-responders are significant to be predictive biomarkers in violin plots even though the log fold change is larger than the cut-off we set because the expression of differentially expressed miRNAs in most cases of both groups were zero.

To further investigate the chemotherapy miRNA signatures and overcome the limitations of conventional differential expression analysis, we applied the forest-based machine learning algorithm on the dataset. All 249 miRNAs after data cleaning were subjected to the random forest model. The top 20 important chemotherapy feature miRNAs were identified by random forest model (Fig. 3A). Interestingly, there was no overlap of miRNAs signatures identified between random forest and differential expression analysis. PCA plot also demonstrated a slightly bigger separation in the first principal component than that in differential expression analysis, suggesting machine learning algorithm could find the features more accurate than the conventional differential expression analysis (Fig. 3B).

The prediction of chemotherapy miRNA signatures were confirmed by three different machine learning algorithms.

Although the importance of each miRNAs was identified by random forest, we still needed to determine how many miRNAs we should involve as the chemotherapy response signatures. To predict the chemotherapy response by miRNAs, we validated the miRNAs predictive capacity by naïve Bayes (NB), the latent Dirichlet allocation (LDA) and Supportive Vector Machine (SVM). The prediction accuracy of miRNAs did not increase when the number of miRNAs involved was larger than 5 in all three models, suggesting the top 5 importance miRNAs could achieve the best predictive accuracy (Fig. 4). Therefore, we determined the top five miRNAs (miR-196b, miR-34c-5p, miR-181b, miR-27b and miR-26a) as the chemotherapy predictive miRNA signatures. Heatmap of these five miRNA signatures revealed there was a difference between chemotherapy responders and non-responders (Fig. 5A). We also found these five miRNAs were detectable in microarray (Fig. 5B), suggesting the five miRNAs were ideal for biomarker design.

Function annotation analysis identified the five miRNAs were enriched in biosynthesis.

To test the function pathway enriched by the five miRNAs, we performed function annotation analysis on the KEGG and GO by DIANA. Biological Processes (BP) of GO of the five miRNAs involved cellular nitrogen compound metabolic process, biosynthetic process, Fc-epsilon receptor signalling pathway, gene expression, symbiosis, encompassing mutualism through parasitism, L-fucose catabolic process. Cellular Components (CC) of GO involved organelle. Molecular Function (MF) of GO involved ion binding, nucleic acid binding transcription factor activity, molecular function. Pathways in KEGG involved Glycosphingolipid biosynthesis, Hematopoietic cell lineage, Synaptic vesicle cycle, Glycosphingolipid biosynthesis, Mucin type O-Glycan biosynthesis, MicroRNAs in cancer, Thyroid cancer (Fig. 6, **Supplementary Fig. 3).**

The chemotherapy prediction of miR-196b and miR-34c-5p were validated in patient serum.

Although the five miRNA signatures demonstrated the excellent predictive capacity, only miR-196b and miR-34c-5p showed large effect sizes between responders and non-responders. Considering both miR-196b and miR-34c-5p played critical roles in lung carcinogenesis, we only included these two miRNAs into clinical validation. In addition, due to blood-(serum-) based liquid biopsy has revealed its potentially clinical utility, we hypothesis serum miR-196b and miR-34c-5p level could predict the chemotherapy response for lung adenocarcinoma patients. A total of 30 patients who had been pathologically diagnosed as lung adenocarcinoma included in this study (Supplementary Table 1). As expected, we observed a significant difference of miR-196b expression between responders and non-responders ($p = 2.365 \times 10^{-5}$). The expression of miR-34c-5p in non-responders was significantly higher than that in responders (Fig. 7A). More importantly, both miR-196b and miR-34c-5p demonstrated the high predictive accuracy (0.898 and 0.931, respectively) in our cohort (Fig. 7B). Taken together, these results uncovered serum miR-196b and miR-34c-5p were differentially expressed between chemotherapy responders and non-responders, and potentially the chemotherapy predictive biomarkers for lung adenocarcinoma patients.

Discussion

In order to accurately investigate miRNA expression levels, it is typical to normalize miRNA expression before analysis. For this reason, unexpected system variation should be removed from the data (20). Elimination of batch effects is an important aspect of standardization but is often ignored in previous studies (20, 21). Batch effects are another confounding factors when the biological group is not divided into processing groups such as laboratories and technicians. The advantage of high-throughput technologies such as miRNA sequencing is that it provides enough data to detect and eliminate batch effects. Random forest classification algorithms were applied to miRNA and mRNA expression data in the TCGA breast cancer dataset, suggesting miRNA expression data is suitable for machine learning iteration (22). In this study, we demonstrated there was a large variation between samples regardless these samples came from one GEO dataset. Indeed, further investigation of missing value probes or no reaction probe is also needed to be considered in the normalization and batch effect removal.

Machine learning rises from data mining and rapidly develops in medical studies, especially for biomarkers discovery, radiological images analysis and disease stratification (23, 24). Compared with conventional differential expression analysis, machine learning has more capacity to deal with complex traits and information. Our results have revealed random forest model had more capacity and accuracy than conventionally differential expression analysis. Linear regression based differential expression analysis could be skewed by extreme values (outliers) in some samples, therefore, the differentially expressed genes or miRNAs could not represent the samples features. Moreover, tree-based model also identifies the importance of inputted parameters while conventional feature analysis could not indicate the top important parameters.

Lack of clinical and biological biomarker that can predict the prognosis of patients is the most critical issue reducing the chemotherapy effectiveness in the first-line chemotherapy for lung adenocarcinoma. In this study, we identified five miRNA expressions are strongly associated with chemotherapy response through random forest. Of them, we selected miR-196b and miR-34c-5p as the miRNA biomarker candidates to validate in our serum samples. We investigated the role of serum miR-196b and miR-34c-5p levels in chemosensitivity and prognosis of lung adenocarcinoma patients receiving platinum-based chemotherapy. To validate the miRNA signature for chemotherapy response prediction, we included 30 serum samples prior to the chemotherapy treatment. Our data indicates that serum miR-196b and miR-34c-5p levels can distinguish patients with response from patients with non-response after standard platinum-based chemotherapy treatment. For patients with lung adenocarcinoma who are sensitive to chemotherapy, the higher serum miR-196b and lower miR-34c-5p levels before chemotherapy are more pronounced than those who are resistant to chemotherapy. Meanwhile, we also assessed the prediction accuracy of serum miR-196b and miR-34c-5p detection. The results showed that both of them showed significant prediction accurate for chemotherapy response.

MiRNA is a small non-coding RNA molecule with a length of approximately 22 nucleotides (25, 26) and plays a vital role in the diagnosis, prognosis and recurrence of lung cancer (27–29). Many studies have reported the potential clinical application of miRNA in lung cancer chemotherapy. In lung cancer, miRNAs enhance anticancer treatments by regulating cell cycle, cell proliferation and apoptosis, DNA damage repair and tumor angiogenesis (30–32). For instance, overexpression of miR-1 increases the chemosensitivity of non-small cell lung cancer cells by inhibiting autophagy (33). Furthermore, overexpression of miR-202 promotes chemosensitivity by suppressing the KRAS gene in lung adenocarcinoma (34). Notably, there also is increasing evidences that circulating miRNAs may be potential biomarkers for early diagnosis and prognosis of lung cancer (35, 36). In addition, up-regulation of miR-200c has been observed to inhibit cell migration and enhance chemosensitivity of breast cancer (37). However, the role of serum miRNAs in predicting chemotherapy sensitivity in patients with lung adenocarcinoma has been rarely reported before. Our proof-of-concept study demonstrated circulating miR-196b and miR-34c-5p could be potent chemotherapy response predictive biomarkers. We included the pathway analysis is because we would like to validate the miRNAs we identified (from Machine Learning) is biologically relevant to the chemotherapy. As expected, the putative targets of these featured miRNAs are enriched in biosynthesis, which is one of the critical pathways in chemotherapy (38). The pathway of putative targets genes of the top five chemotherapy response miRNA signatures is enriched in the biosynthesis, suggesting miR-196b and miR-34c-5p that regulated epigenetically in lung adenocarcinoma have an interaction on the DNA damage induced by platinum (39).

Although our proof-of-concept study revealed circulating miR-196b and miR-34c-5p level are predictive biomarker for chemotherapy response in lung adenocarcinoma, there are still some limitations. We applied forest-based model to screen miRNA signature and compared its accuracy with conventional analysis. However, we did not test other models like neural network and deep learning yet in this study. Further investigation will be performed to test the utility of other models and select the best predictive model for chemotherapy response. Due to the blood samples scarcity, we only selected miR-196b and

miR-34c-5p as our assay target and we haven't validated other miRNA signatures in serum sample. A perspective cohort will be proposed to investigate whether other circulating miRNA has same predictive capacity.

Conclusion

In conclusion, we performed a forest-based machine learning algorithm to identify platinum-based chemotherapy response miRNA signatures. These miRNA signatures have an effect on DNA damage by participating in biosynthesis. We verified two miRNAs (miR-196b and miR-34c-5p) in our cohort with 30 serum samples of lung adenocarcinoma and these two circulating miRNAs demonstrated high prediction accuracy for chemotherapy response, highlighting the clinical utility of circulating miRNAs in adjuvant management of lung cancer.

Abbreviations

Receiver Operation Curve ROC

Area under the curve AUC

naïve Bayes NB

latent Dirichlet allocation LDA

Supportive Vector Machine SVM

Biological Processes BP

Cellular Components CC

Molecular Function MF

Declarations

There is no conflict of interest between authors.

Funding information is not applicable / No funding was received.

This study has been approved by the ethics committee of Shenzhen Longgang District People's Hospital. The ethics number: LDPH-2019005. All patients participated in this study have signed the consents.

Availability of data and material

N/A

Authorship contribution: YL, JX, LC write this manuscript and performed experiments. DW, WW perform the analytical analyses. KS, PC proof-read this manuscript. all authors have read and approved the manuscript.

There is no conflict of interests between authors.

Acknowledgements

N/A

References

1. Soh J, Okumura N, Nakata M, et al. Randomized feasibility study of S-1 for adjuvant chemotherapy in completely resected Stage IA non-small-cell lung cancer: results of the Setouchi Lung Cancer Group Study 0701. *Jpn J Clin Oncol*. 2016;46:741–7.
2. Yang P. Epidemiology of lung cancer prognosis: quantity and quality of life. *Methods Mol Biol*. 2009;471:469–86.
3. Rossi A, Chiodini P, Sun JM, et al. Six versus fewer planned cycles of first-line platinum-based chemotherapy for non-small-cell lung cancer: a systematic review and meta-analysis of individual patient data. *Lancet Oncol*. 2014;15:1254–62.
4. Sirohi B, Ashley S, Norton A, et al. Early response to platinum-based first-line chemotherapy in non-small cell lung cancer may predict survival. *J Thorac Oncol*. 2007;2:735–40.
5. Bullard JT, Eberth JM, Arrington AK, Adams SA, Cheng X, Salloum RG. Timeliness of Treatment Initiation and Associated Survival Following Diagnosis of Non-Small-Cell Lung Cancer in South Carolina. *South Med J*. 2017;110:107–13.
6. Xu CH, Yang Y, Wang YC, Yan J, Qian LH. Prognostic significance of serum chemerin levels in patients with non-small cell lung cancer. *Oncotarget*. 2017;8:22483–9.
7. Yagishita S, Horinouchi H, Sunami KS, et al. Impact of KRAS mutation on response and outcome of patients with stage III non-squamous non-small cell lung cancer. *Cancer Sci*. 2015;106:1402–7.
8. Strimpakos AS, Banerji U, Thavasu P, Tsilimagou A, Psyrris A, Syrigos KN. Percentage Change in Plasma Cytokeratin 18 Is Associated with Clinical Outcomes in Patients Receiving Pemetrexed and Carboplatin for the Adenocarcinoma Subtype of NSCLC. *Oncology*. 2015;89:53–9.
9. Jagga Z, Gupta D. Machine learning for biomarker identification in cancer research - developments toward its clinical application. *Per Med*. 2015;12:371–87.
10. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform*. 2008;77:81–97.
11. Sun D, Li X, Ma M, et al. The predictive value and potential mechanisms of miRNA-328 and miRNA-378 for brain metastases in operable and advanced non-small-cell lung cancer. *Jpn J Clin Oncol*. 2015;45:464–73.

12. Iorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol Med*. 2017;9:852.
13. Ma C, Nguyen HPT, Luwor RB, et al. A comprehensive meta-analysis of circulation miRNAs in glioma as potential diagnostic biomarker. *PLoS One*. 2018;13:e0189452.
14. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23:1846–7.
15. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43:e47.
16. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res*. 2018;46:D239–45.
17. Lin X, Jia J, Du T, et al. Overexpression of miR-155 in the liver of transgenic mice alters the expression profiling of hepatic genes associated with lipid metabolism. *PLoS One*. 2015;10:e0118417.
18. Li H, Feng C, Shi S. miR-196b promotes lung cancer cell migration and invasion through the targeting of GATA6. *Oncol Lett*. 2018;16:247–52.
19. Treiber T, Treiber N, Plessmann U, et al.: A Compendium of RNA-Binding Proteins that Regulate MicroRNA Biogenesis. *Mol Cell* 66: 270–284 e213, 2017.
20. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*. 2014;32:896–902.
21. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11:733–9.
22. Cava C, Colaprico A, Bertoli G, Bontempi G, Mauri G, Castiglioni I. How interacting pathways are regulated by miRNAs in breast cancer subtypes. *BMC Bioinformatics*. 2016;17:348.
23. Adamson AS, Welch HG. Machine Learning and the Cancer-Diagnosis Problem - No Gold Standard. *N Engl J Med*. 2019;381:2285–7.
24. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18:500–10.
25. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009;136:215–33.
26. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. *Trends Mol Med*. 2014;20:460–9.
27. Skrzypski M, Czapiewski P, Goryca K, et al. Prognostic value of microRNA expression in operable non-small cell lung cancer patients. *Br J Cancer*. 2014;110:991–1000.
28. Yang J, Liu H, Wang H, Sun Y. Down-regulation of microRNA-181b is a potential prognostic marker of non-small cell lung cancer. *Pathol Res Pract*. 2013;209:490–4.
29. Zhu W, He J, Chen D, et al. Expression of miR-29c, miR-93, and miR-429 as potential biomarkers for detection of early stage non-small lung cancer. *PLoS One*. 2014;9:e87780.

30. Bottai G, Pasculli B, Calin GA, Santarpia L. Targeting the microRNA-regulating DNA damage/repair pathways in cancer. *Expert Opin Biol Ther.* 2014;14:1667–83.
31. Sullivan I, Salazar J, Majem M, et al. Pharmacogenetics of the DNA repair pathways in advanced non-small cell lung cancer patients treated with platinum-based chemotherapy. *Cancer Lett.* 2014;353:160–6.
32. Bonanno L, Favaretto A, Rosell R. Platinum drugs and DNA repair mechanisms in lung cancer. *Anticancer Res.* 2014;34:493–501.
33. Hua L, Zhu G, Wei J. MicroRNA-1 overexpression increases chemosensitivity of non-small cell lung cancer cells by inhibiting autophagy related 3-mediated autophagy. *Cell Biol Int.* 2018;42:1240–9.
34. Sun W, Ping W, Tian Y, Zou W, Liu J, Zu Y. miR-202 Enhances the Anti-Tumor Effect of Cisplatin on Non-Small Cell Lung Cancer by Targeting the Ras/MAPK Pathway. *Cell Physiol Biochem.* 2018;51:2160–71.
35. Hu Z, Chen X, Zhao Y, et al. Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer. *J Clin Oncol.* 2010;28:1721–6.
36. Shen J, Todd NW, Zhang H, et al. Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. *Lab Invest.* 2011;91:579–87.
37. Wang Q, Cheng Y, Wang Y, et al. Tamoxifen reverses epithelial-mesenchymal transition by demethylating miR-200c in triple-negative breast cancer cells. *BMC Cancer.* 2017;17:492.
38. Hagner N, Joerger M. Cancer chemotherapy: targeting folic acid synthesis. *Cancer Manag Res.* 2010;2:293–301.
39. Tellez CS, Juri DE, Do K, et al. miR-196b Is Epigenetically Silenced during the Premalignant Stage of Lung Carcinogenesis. *Cancer Res.* 2016;76:4741–51.

Figures

Figure 1

The flow chart of this study design.

Figure 2

conventional differential expression analysis identified no significant miRNAs between responder and non-responder. A, principal component plot of miRNA expression. Red, chemotherapy responder; Green, chemotherapy non-responders. B, Volcano plot of differentially expressed miRNA analysis. Red dot line, threshold of significance ($FDR < 0.05$ and $|\log \text{Fold change}| > 2$). Significant miRNAs were labelled as green.

Figure 3

Random forest revealed featured miRNA in chemotherapy responders. A, Mean decrease accuracy and mean decrease Gini of miRNAs obtained from random forest model trained on miRNA expression data to classify chemotherapy responses. The higher the Mean Decrease Accuracy and Mean Decrease Gini index for a miRNA, the more important that miRNA for construction of trees in random forest model.

Figure 5

Five miRNA signatures discriminate chemotherapy response. A, unsupervised clustering showing the five miRNA signatures were differentially expressed between responders and non-responders. B, the five miRNAs expression in each sample.

Figure 6

BP of GO and KEGG analysis of the top 5 miRNA signatures.

Figure 7

Serum miR-196b and miR-34c expression predict chemotherapy response. A, qRT-PCR showing miR-196b and miR-34c-5p in serum are differently expressed between responders and non-responders. B, the effect of chemotherapy can be predicted by serum miR-196b and miR-34c-5p expression.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigure3.tif](#)
- [SupplementaryFigure2.tif](#)
- [SupplementaryFigure1.tif](#)
- [SupplementaryTable1.xlsx](#)