A mRNA-miRNA-lncRNA regulatory network related to potential pathogenesis and prognostic markers of Non- small cell lung cancer

Fei Wang
Hubei University of Chinese Medicine
Chong Yuan (1367147371@qq.com)
Hubei University of Chinese Medicine
Yanfang Yang
Hubei University of Chinese Medicine
Bo Liu
Hubei University of Chinese Medicine
Hezhen Wu
Hubei University of Chinese Medicine

Research Article

Keywords: Non-small cell lung cancer, mRNA-miRNA-lncRNA, pathogenesis, prognostic biomarkers

Posted Date: March 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-318131/v1

Abstract

Non-small cell lung cancer (NSCLC) is one of the most malignant tumors with the fastest increasing incidence and mortality rate, but the etiology of NSCLC is still not clear. Most of lncRNAs have some structural similarities with mRNAs, suggesting that miRNAs negatively regulate the expression of lncRNAs to affect the occurrence and development of tumor. Therefore, system bioinformatics was used to explore the potential biomarkers and possible pathogenesis of NSCLC in this study. Firstly, all the clinical information and transcriptome data were downloaded from GEO and TCGA databases. R language was used to analyze the differentially expressed genes (DEGs) in NSCLC and normal lung tissues. Then, 50 overlapped DEGs were obtained via Venn database, including 10 down-regulated mRNAs and 40 down-regulated mRNAs. Secondly, the top 20 DEGs were selected for KEGG pathway and GO enrichment analysis. After screening 4 HUB genes related to the survival and prognosis of NSCLC patients, their prognosis models were established. Meanwhile, HUB genes related miRNAs and lncRNAs were screened. Finally, a mRNA-miRNA-lncRNA network related to the survival and prognosis of NSCLC patients was established, including 4 up-regulated mRNAs, 3 up-regulated miRNAs, 10 down-regulated miRNAs, 6 up-regulated lncRNAs and 19 down-regulated lncRNAs.

Subject terms: Non-small cell lung cancer, mRNA-miRNA-lncRNA, pathogenesis, prognostic biomarkers.

Introduction

Lung cancer poses the greatest threat to human health and life activities. According to the data released by China National Cancer Center, there were about 780,000 new cases of lung cancer and 630,000 people died of it in China in 2015[1]. Epidemiological data shows that Non-small cell lung cancer (NSCLC) patients account for 85% of lung cancer patients[2]. So, it has become the focus of medical research. At present, the treatment of NSCLC patients includes traditional surgery, chemotherapy, radiotherapy, targeted therapy and immunotherapy. The 5-year survival rate is 10%-30%, and the recurrence rate is still high. Because the high degree of malignancy of NSCLC, leading to untimely medical treatment, most patients with NSCLC has missed the best operation opportunity when diagnosed, more than 50% of patients has been relatively advanced once diagnosed[3]. Even if surgical treatment, the overall effective rate of patients with NSCLC is lower than 25%. Therefore, it is very important to elaborate the pathogenesis and the biomarkers related to the survival and prognosis of NSCLC.

miRNAs negatively regulate the expression of lncRNAs through a mechanism similar to that of mRNAs, and then play a series of biological roles[4]. Many studies have confirmed that they play an important role in the development and occurrence of tumor. It is generally believed that lncRNAs indirectly inhibit the negative regulation of miRNAs on genes by competing with miRNAs to bind to the 3 '- UTR of mRNAs. All of them interact to form a ceRNA network which is the necessary process for human to understand the pathogenesis of tumor and develop new treatment methods[5]. However, there are few reports on the effect of ceRNA network on NSCLC. Therefore, it is necessary to systematically analyze the relationship

between miRNAs, lncRNAs, mRNAs related to the survival and prognosis of NSCLC patients, and to elaborate the regulatory mechanism of ceRNA network on NSCLC.

Bioinformatics is a new discipline based on molecular biology and multi disciplines. It uses biology, computer science and information technology to analyze a large number of complex biological data[6,7]. The emergence of bioinformatics breaks the limitations of traditional research, and provides a reliable reference for scientific research. At present, the application of bioinformatics in the study of proteins closely related to the occurrence and development of diseases has become a research hotspot.

In this study, the gene expression profiles of NSCLC patients were first screened from Gene Expression Synthesis (GEO) and The Cancer Genome Atlas (TCGA) databases, and Venn database was used to screen the intersection of them, so as to improve the accuracy of the study. Through the systematic analysis of DEGs, the top 20 DEGs were used for KEGG and GO enrichment. After obtaining the Hub genes related to the survival and prognosis of NSCLC, R software, GEPIA2.0 and the Human Protein Atlas databases were used to verify the expression of them. Meanwhile, the corresponding miRNAs and lncRNAs were also analyzed. Finally, the ceRNA network related to the survival and prognosis of NSCLC was constructed to analyze the pathogenesis of NSCLC and to provide new diagnostic biomarkers. The flow chart of this study is shown in Figure 1.

Results

Acquiring 68 DECs in NSCLC

From GSE6044 and GSE66759 in GEO database, 109 human NSCLC samples and 10 human normal lung samples were collected totally. Then, 1014 human NSCLC samples and 627 human normal lung samples were collected from TCGA and GTEx databases respectively. Limma package of R software (version: 3.40.2) was used to analyze the DEGs in different datasets. As shown in Figure S1, 245 DEGs were obtained from GSE6044, of which 89 were up-regulated and 156 were down-regulated. Then, 2852 DEGs were screened from GSE66759, of which 1701 were up-regulated and 1151 were down-regulated. The results are shown in Figure S2. At the same time, 4350 DEGs were got from TCGA database, including 1836 up-regulated genes and 2514 down-regulated genes, as Figure S3 showing. In order to improve the accuracy of this study, Venn database was used to analyze them. Then, 68 overlapped DEGs were obtained through it, including 25 down-regulated genes and 43 up-regulated genes. The results are shown in Figure 2A. Using STRING database, the interaction between DEGs was analyzed. And its TSV format was imported into Cytoscape 3.7.1 software for visualization, as shown in Figure 2B. The red node represented the up-regulated genes, the green node represented the down-regulated genes. Through its main plug-in cytoHubba, all DEGs were sorted by degree. The top 10 and top 20 DEGs were obtained as HUB genes in Figure 2C,D respectively.

Functional enrichment analysis of the top 20 HUB genes

Based on the DAVID 6.8 database, GO enrichment and KEGG pathway analyses of the top 20 HUB genes were obtained. At the same time, Omicshare database was used to demonstrate the above results with circos. The top 10 GO biological process indicated that these genes were mainly involved in the process of mitosis. And their top 10 GO cell components were mainly microtubule, kinetochore, extracellular matrix and chromosome. The top 10 GO molecular function indicated that these genes could bind to proteins, drugs, ubiquitin and microtubules et all (Figure 2E). KEGG analysis revealed that these genes were mainly enriched in ECM-receptor interaction, protein digestion and absorption, cell cycle, p53 signaling pathway, focal adhesion pathways and Amoebiasis (Figure 2F). Above results revealed that these genes regulated cell cycle by participating in cell mitosis. However, their binding relationship with drugs, proteins and microtubules was remained to be verified.

Survival analysis of the top 20 HUB genes

Screening HUB genes which are significantly related to the survival and prognosis of NSCLC patients is a necessary prerequisite for the analysis of the pathogenesis and treatment of NSCLC. Therefore, the top 20 HUB genes were imported into the Kaplan Meier plotter database for analysis. By analyzing the clinical data of 999 patients, the 4 HUB genes (RRM2, TYMS, NCAPG, CDK1) related to the OS of NSCLC were obtained with a setting $P < 0.05$ (Figure 3).

In order to verify the expression level of the 4 HUB genes in NSCLC, R software (version 4.0.3), GEPIA 2.0 and the Human Protein Atlas databases were used with a setting $P < 0.05$. The results of GEPIA 2.0 database and R software showed that the 4 HUB genes were highly expressed in NSCLC (Figure 3). As shown in Figure 4, the Human Protein Atlas database recorded immunohistochemical data of CDK1, NCAPG and TYMS in NSCLC and normal lung tissues with two antibodies. But no RRM2, unfortunately. The evidence of the above data further confirmed the previous results in this study.

Establishment of prognosis model

It is necessary to establish the prognosis model after obtaining the HUB genes which were significantly related to the survival and prognosis of NSCLC. Before that, a significant correlation between the four genes through GEPIA2.0 database were established (Figure 5A). Then, Cox regression analysis and "forest plot" R package were used to display *P* value, HR and 95% CI of each variable. The results showed that NCAPG, RRM2, TYMS and CDK1 were associated with age, pTNM_stage and new tumor of NSCLC patients. Interestingly, they were not significantly correlated with gender and smoking (Figure 5B). The nomogram and its corresponding curves could help us predict the total recurrence rate of NSCLC patients in 1, 3, 5 years (Figure 5C,D). But it is not enough to understand the pathogenesis of NSCLC with these imformation.

Identifcation of 14 miRNAs related to the OS of NSCLC

Based on the miRTarBase and miRWalk database, 40 miRNAs which had been confirmed by western blot, real-time PCR and other experiments were obtained. 15 of them related to RRM2, 7 of them related to

TYMS, 7 of them related to NCAPG and 11 of them related to CDK1. Then, 14 miRNAs related to the OS of NSCLC were obtained via the Kaplan Meier plotter database with a setting log-rank $P < 0.05$ (as shown in Figure S4). 3 of them were up-regulated and the other 11 miRNAs were down-regulated. It is an important method to study the pathogenesis of NSCLC by analyzing the relationship between them and their corresponding mRNAs.

Identifcation of 25 lncRNAs related to the OS of NSCLC

In order to further study the regulation of lncRNA on mRNA and miRNA, 25 lncRNAs were obtained via the lnCAR, DIANA and the Kaplan Meier plotter databases (Figure S5). 6 of them were up-regulated and the other 19 lncRNAs were down-regulated. There is no doubt that all of them are potential biomarkers of NSCLC. But how they regulate the expression of miRNAs needs more data.

ceRNA network construction

By competing with miRNA to bind 3'-UTR of mRNA, lncRNA indirectly inhibits the negative regulation of miRNA on mRNA. Therefore, after integrating the mRNA-miRNA-lncRNA interactions, a ceRNA network was constructed with 4 mRNAs, 11 miRNAs and 25 lncRNAs via Cytoscape 3.7.1. As shown in Figure 6, nodes with different colors and shapes represents different targets. Through the analysis of the relationship between them by Cytoscape 3.7.1 software, we found that the degree of hsa-mir-142-5p, hsa-mir-30e-5p and hsa-mir-30d-5p were all 8. Therefore, they were the most critical miRNAs. NEAT1 and LMCD1-AS1 played the most important role in these lnRNAs, because their degree values were all 4 and higher than others. The degree of CDK1 and NCAPG were 5 and 4, while the degree of RRM2 and TYMS were 2 and 1, respectively. This suggested that CDK1 and NCAPG were the core starting points or turning points in the pathogenesis of NSCLC. However, the core nodes involved in the network needed to be verified by experiments in vivo and in vitro.

Discussion

NSCLC includes squamous cell carcinoma (SCC), adenocarcinoma and large cell carcinoma and it is one of the most common malignant tumors in the world. NSCLC accounts for about 80% of all lung cancer, and about 75% of the patients are in the advanced stage when they are found. So, the 5-year survival rate of patients is very low[8]. The current surgical treatment, chemotherapy and other means can't significantly improve the survival rate of patients, but greatly reduce the quality of life of patients at the same time[9]. It brings heavy economic burden to the patients' families, which makes countless families fragmented. We believe that the search for biomarkers related to prognosis and diagnostic of NSCLC is the necessary premise and the biggest problem for the development of therapeutic drugs for NSCLC. Therefore, this study aims to explore the DEGs in clinical samples, and to analyze their relationship with the survival and prognosis of NSCLC patients.

With the advent of the era of big data, bioinformatics has been greatly improved[10]. This is a great help for our scientific research. Through the collection of clinical data and transcriptome experiment data, it

provides a larger research platform for the researchers, and also provides a guarantee for researchers' data. Therefore, 1123 NSCLC clinical samples and 637 normal samples were downloaded from TCGA (the largest tumor database in the world) and GEO (the most comprehensive database). The purpose of comprehensive analysis of them was to further improve the accuracy of this study. After the analysis of DEGs in NSCLC and normal lung tissues by limma software package of R software (version: 3.40.2), STRING database and Cytoscape 3.7.1 software were used to analyze the interaction between the genes. The top 20 HUB genes were obtained by ranking the degree in cytoHubba. Their functional analysis and KEGG pathway enrichment analysis showed that they played an anti-NSCLC role mainly by participating in the mitosis of tumor cells and regulating cell cycle[11]. But the specific function should to be verified.

In the Kaplan Meier plotter database, we found that RRM2, TYMS, NCAPG and CDK1 were significantly correlated with the survival and prognosis of NSCLC. Through comprehensive analysis of the clinical data recorded in GEPIA 2.0 database and the clinical samples we downloaded before, these four genes were found significantly up-regulated in NSCLC and had a strong correlation, which confirmed the results of previous survival analysis. What surprised us most was that the Human Protein Atlas database contained the expression results of CDK1, NCARG and TYMS in the pathological tissues of NSCLC patients and normal human lung tissues. Although RRM2 was not recorded, combined with the above results, the prognostic model of these 4 genes and NSCLC were established. The results showed that NCAPG, RRM2, TYMS and CDK1 were associated with age, pTNM_ stage and new tumor of NSCLC patients. Interestingly, they were not significantly correlated with gender or smoking. But many studies had shown that smoking was one of the main causes of lung cancer, so this conclusion needed to be further proved[12,13]. Although the prognostic model and related parameters were established via the results of comprehensive analysis of clinical samples in TCGA and GEO, but it was reliable and could well predict the 1-, 3-, 5- year survival rate and recurrence rate of NSCLC patients.

miRNAs are endogenous small RNA with a length of about 20-24 nucleotides[14]. Due to the universality and diversity of miRNA, it is suggested that miRNA may have a wide range of biological functions. Many studies have shown that miRNA is involved in a series of important processes in life, including cell proliferation, apoptosis, cell death, fat metabolism and cell differentiation[15]. What is most noteworthy is that many studies have shown that there is a close relationship between the differential expression of miRNA and tumor, which provides a new idea for us to treat tumor. lncRNA is a non-coding RNA with a length of more than 200 nucleotides[16]. Previous studies have shown that it is involved in many important regulatory processes, such as X chromosome silencing, genomic imprinting, chromatin modification, transcriptional activation, transcriptional interference, and intranuclear transport[17]. Its abnormal expression or function is closely related to the occurrence of human diseases, including cancer, degenerative neurological diseases and many other major diseases that seriously endanger human health. Because the structure of most lncRNA is similar to that of mRNA, it suggests that lnRNA can indirectly inhibit the negative regulation of gene by competing with miRNA to bind 3'- UTR of mRNA, and then play a series of biological roles. The ceRNA network formed by the interaction of these three is of great value for the study of tumor pathogenesis. Through miRTarBase and miRWalk databases, 40

miRNAs that had been verified to be significantly associated with the 4 HUB genes were screened. Among them, 11 down-regulated miRNAs and 3 up-regulated miRNAs were significantly associated with NSCLC. Literature review found that the above miRNAs were significantly correlated with the occurrence and development of tumor, which was consistent with our previous research. The analysis of lnCRA, DIANA and the Kaplan Meier plotter databases showed that 6 up-regulated lncRNAs and 19 down-regulated lncRNAs were significantly associated with NSCLC. Based on the above results, a ceRNA network was established, which was significantly related to the survival and prognosis of NSCLC. However, lncRNA could form the precursor of miRNA through intracellular splicing, and then process into specific miRNA to regulate the expression of genes. It could also play the role of endogenous miRNA sponge and inhibit the expression of miRNA. Therefore, the specific relationship between them needed to be further explained by more experiments.

Taken together, through system bioinformatics, a ceRNA network with 4 miRNAs, 14 miRNAs and 25 lncRNAs was constructed. And it was significantly related to the survival and prognosis of NSCLC from the data of patients with NSCLC in TCGA and GEO databases. However, our experimental conclusion was obtained through analysis, more experiments were needed to verified these results.

Conclusion

In conclusion, we identified a ceRNA network consisting of 4 mRNAs, 11 miRNAs and 25 lncRNAs in NSCLC based on TCGA and GEO databases. The construction of this network provided new ideas and insights for the pathogenesis, clinical diagnosis and treatment of NSCLC.

Materials And Methods

Collection of clinical information

GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi) is a gene expression database created and maintained by the National Biotechnology Information Center of the United States[18]. It contains tens of thousands of genomic data, and provides tools to help users query and download experiments and gene expression profiles. As the largest cancer gene information database (covering 33 cancer types, more than 30000 tumor samples and more than 20000 genes' expression information), TCGA (https://www.cancer.gov/ tcga) aims to apply high-throughput genomic analysis technology to help people have a better understanding of cancer[19].

Identification of DEGs

In order to ensure the accuracy of this study, data from GEO and TCGA databases were used for comprehensive analysis. Limma package of R software (version 4.0.3) was used to analyze the DEGs[20]. The adjusted P-value was analyzed to correct for false positive results in GEO datasets. "Adjusted P < 0.05 and |Log FC| >2" were defined as the thresholds for the screening of differential expression of

mRNAs. The ggplot2, ggord and pheatmap packages were used to visualize the above results. Then, Venn database was used to obtain the intersection genes.

Identification of HUB genes

Analysis of the interaction between proteins is an important means to obtain HUB genes. The Search Tool for the Retrieval of Interacting Genes database (STRING, https://string-db.org/) contains more than 1100 kinds of proteins and more than 52 million kinds of interactions[21]. By importing the official name of DEGs into it, the TSV format of “protein-protein interaction” (PPI) network was obtained, and the relevant results were visualized via the plug-in cytoHubba in Cytoscape 3.7.1 software[22]. Finally, the HUB genes were screened according to the degree.

KEGG pathway and GO enrichment analyses

The analysis of the biological function of HUB genes and the pathways involved in them is an essential link in the study of the occurrence and development of diseases. The database for annotation, visualization, and integrated discovery (DAVID 6.8, https://david.abcc.ncifcrf.gov/) is a database which provides a comprehensive set of functional annotation tools for researchers to study genes[23]. Therefore, DAVID 6.8 was used to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment and Gene Ontology (GO) analysis about the HUB genes with a setting $P < 0.05$ in this study.

Survival analysis of HUB genes

Biomarkers related to the survival and prognosis of NSCLC can provide new ideas and insights for us to against NSCLC. The Kaplan Meier plotter is a database which includes 54,000 genes and 21 cancer types from GEO and TCGA[24]. It is usually used to discover survival biomarkers via meta-analysis. Through analyzing the clinical data of 999 patients, the HUB genes related to the overall survival (OS) of NSCLC were obtained with a setting log-rank $P < 0.05$. At the same time, GEPIA2.0 database (http://gepia.cancer-pku.cn/), the Human Protein Atlas database (https://www.proteinatlas.org/) and R software (version 4.0.3) were used to analyze the expression of these HUB genes in NSCLC and normal lung tissues, $P < 0.05$ was considered to be statistically significant[25].

Establishment of prognosis model

The "forest plot" R package was used to display the *P* value, HR and 95% CI of each variable. Based on the first mock exam results of multivariate Cox proportional hazards analysis, the nomogram was established using the R package "RMS" to predict the total recurrence rate of 1, 3, 5 years. This model was helpful for us to calculate the prognosis risk of a single patient by the point at which each risk factor was related[26].

Identification of miRNAs related to the survival and prognosis of NSCLC

As a database, the experimentally validated microRNA-target interactions database (miRTarBase, http://mirtarbase.cuhk.edu.cn/php/index.php) has accumulated more than 360 thousand miRNA-target interactions (MTIs) which were verified by reporter gene analysis, Western blot, microarray and next generation sequencing experiments[27]. miRWalk database contains information such as miRNA-mRNA binding experiment data. In order to avoid the inaccuracy of selectivity in this study, both of them were used to predict miRNAs related to the biomarkers of NSCLC[28]. Then, the Kaplan Meier plotter

database was used to obtain the miRNAs related to the OS of NSCLC with a setting log-rank $P < 0.05$.

Identification of lncRNAs related to the survival and prognosis of NSCLC

lnCAR is a database including 10 cancer types and more than 54000 samples[29]. By analyzing real-time PCR and northern hybridization data, it allows users to explore the functions of lncRNAs. DIANA database provides data from algorithms, databases and software to analyze the expression regulation of deep sequencing data, miRNA regulatory elements and the role of lncRNAs in various diseases and pathways[30]. Through comprehensive analysis, the intersection lncRNAs which had been verified by experiments in the two databases were obtained. Finally, the lncRNAs related to the OS of NSCLC were screened via the Kaplan Meier plotter database with a setting log-rank $P < 0.05$.

Construction of mRNA–miRNA–lncRNA network

A mRNA–miRNA–lncRNA regulatory network related to the OS of NSCLC was constructed via Cytoscape 3.7.1 software. Its main purpose was to explore the pathogenesis of NSCLC and provide new biomarkers for the diagnosis and treatment of NSCLC.

Statistical analysis

GO entry and KEGG pathway enrichment were screened according to $P < 0.05$ with students' t-test. ANOVA was used to analyze the expression of HUB genes related to the OS of NSCLC in GEPIA2.0 database, |Log2fc | cutoff < 1 and Q-value < 0.05 were considered to be significant. OS was analyzed based on the Mantel–Cox test and log-rank $P < 0.05$ was the threshold. R software (version 4.0.3) was used for the rest of the analysis, and the threshold value was log-rank $P < 0.05$.

Declarations

Conflicts of interest

The authors declare no conflicts of interest.

Author contribution

F.W. and C.Y. designed the study. Material preparation, data collection and analysis were performed by all of them. F.W. and C.Y. wrote the manuscript. B.L., Y.-F.Y., and H.-Z.W. revised the manuscript. All authors

read and approved the final manuscript.

Funding

This work was supported by National key R & D plan key project of TCM Modernization Research (2017YFC1701000).

Notes

[#]These authors contribute equally to this work.

Data availability statement

All data utilized in this study are included in this article, and all data supporting the findings of this study are available on reasonable request from the corresponding author.

References

1 Shi, Y. J. *et al.* Quality assessment of global lung cancer screening guidelines and consensus. *Zhong hua liu xing bing xue za zhi***42**, 241-247, doi:10.3760/cma.j.cn112338-20200806-01035 (2021).

2 Schegoleva, A. A. *et al.* Prognosis of Different Types of Non-Small Cell Lung Cancer Progression: Current State and Perspectives. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology***55**, 29-48, doi:10.33594/000000340 (2021).

3 Wang, W. *et al.* Efficacy of docetaxel plus carboplatin combination chemotherapy for advanced non-small cell lung cancer. *Zhong guo fei ai za zhi* **10**, 316-319, doi:10.3779/j.issn.1009-3419.2007.04.13 (2007).

4 Pothipor, C., Aroonyadet, N., Bamrungsap, S., Jakmunee, J. & Ounnunkad, K. A highly sensitive electrochemical microRNA-21 biosensor based on intercalating methylene blue signal amplification and a highly dispersed gold nanoparticles/graphene/polypyrrole composite. *The Analyst*, doi:10.1039/d1an00116g (2021).

5 Miao, H. *et al.* LncRNA TP73-AS1 enhances the malignant properties of pancreatic ductal adenocarcinoma by increasing MMP14 expression through miRNA -200a sponging. *Journal of cellular and molecular medicine*, doi:10.1111/jcmm.16425 (2021).

6 Yan, Y. *et al.* CCMAlnc Promotes the Malignance of Colorectal Cancer by Modulating the Interaction Between miR-5001-5p and Its Target mRNA. *Frontiers in cell and developmental biology***8**, 566932, doi:10.3389/fcell.2020.566932 (2020).

7 Yuan, C. *et al.* Network pharmacology and molecular docking reveal the mechanism of Scopoletin against non-small cell lung cancer. *Life sciences***270**, 119105, doi:10.1016/j.lfs.2021.119105 (2021).

8 Zang, X. *et al.* Dual-targeting tumor cells and tumor associated macrophages with lipid coated calcium zoledronate for enhanced lung cancer chemoimmunotherapy. *International journal of pharmaceutics***594**, 120174, doi:10.1016/j.ijpharm.2020.120174 (2021).

9 Weissferdt, A. *et al.* Agreement on Major Pathological Response in NSCLC Patients Receiving Neoadjuvant Chemotherapy. *Clinical lung cancer***21**, 341-348, doi:10.1016/j.cllc.2019.11.003 (2020).

10 Liu, J. & Wang, X. Early recognition of tomato gray leaf spot disease based on MobileNetv2-YOLOv3 model. *Plant methods***16**, 83, doi:10.1186/s13007-020-00624-2 (2020).

11 Marquis, C. *et al.* Chromosomally unstable tumor cells specifically require KIF18A for proliferation. *Nature communications***12**, 1213, doi:10.1038/s41467-021-21447-2 (2021).

12 Passaro, A. *et al.* Clinical features affecting survival in metastatic NSCLC treated with immunotherapy: A critical review of published data. *Cancer treatment reviews***89**, 102085, doi:10.1016/j.ctrv.2020.102085 (2020).

13 Chen, S. & Wu, S. Identifying Lung Cancer Risk Factors in the Elderly Using Deep Neural Networks: Quantitative Analysis of Web-Based Survey Data. *Journal of medical Internet research***22**, e17695, doi:10.2196/17695 (2020).

14 Werner, A. *et al.* Contribution of natural antisense transcription to an endogenous siRNA signature in human cells. *BMC genomics***15**, 19, doi:10.1186/1471-2164-15-19 (2014).

15 Ayati, S. H. *et al.* Regulatory effects of berberine on microRNome in Cancer and other conditions. *Critical reviews in oncology/hematology***116**, 147-158, doi:10.1016/j.critrevonc.2017.05.008 (2017).

16 Cardillo, N. *et al.* Identification of Novel lncRNAs in Ovarian Cancer and Their Impact on Overall Survival. *International journal of molecular sciences***22**, doi:10.3390/ijms22031079 (2021).

17 Duan, A. *et al.* Chromatin architecture reveals cell type-specific target genes for kidney disease risk variants. *BMC biology***19**, 38, doi:10.1186/s12915-021-00977-7 (2021).

18 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research***41**, D991-995, doi:10.1093/nar/gks1193 (2013).

19 Zhang, X. *et al.* Identification of functional lncRNAs in gastric cancer by integrative analysis of GEO and TCGA data. *Journal of cellular biochemistry***120**, 17898-17911, doi:10.1002/jcb.29058 (2019).

20 Li, L. *et al.* Upregulation of circular RNA circ_0001721 predicts unfavorable prognosis in osteosarcoma and facilitates cell progression via sponging miR-569 and miR-599. *Biomedicine & pharmacotherapy =*

*Biomedecine & pharmacotherapie***109**, 226-232, doi:10.1016/j.biopha.2018.10.072 (2019).

21 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research***47**, D607-d613, doi:10.1093/nar/gky1131 (2019).

22 Sun, Q. *et al.* Differential Expression and Bioinformatics Analysis of circRNA in Non-small Cell Lung Cancer. *Frontiers in genetics***11**, 586814, doi:10.3389/fgene.2020.586814 (2020).

23 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols***4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

24 Nagy, Á., Lánczky, A., Menyhárt, O. & Győrffy, B. Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Scientific reports***8**, 9227, doi:10.1038/s41598-018-27521-y (2018).

25 Tang, Z. *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic acids research***45**, W98-w102, doi:10.1093/nar/gkx247 (2017).

26 Liu, Z. *et al.* A lncRNA prognostic signature associated with immune infiltration and tumour mutation burden in breast cancer. *Journal of cellular and molecular medicine***24**, 12444-12456, doi:10.1111/jcmm.15762 (2020).

27 Huang, H. Y. *et al.* miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic acids research***48**, D148-d154, doi:10.1093/nar/gkz896 (2020).

28 Sticht, C., De La Torre, C., Parveen, A. & Gretz, N. miRWalk: An online resource for prediction of microRNA binding sites. *PloS one***13**, e0206239, doi:10.1371/journal.pone.0206239 (2018).

29 Zheng, Y. *et al.* lnCAR: A Comprehensive Resource for lncRNAs from Cancer Arrays. *Cancer research***79**, 2076-2083, doi:10.1158/0008-5472.can-18-2169 (2019).

30 Ren, L., Guo, D., Wan, X. & Qu, R. EYA2 upregulates miR-93 to promote tumorigenesis of breast cancer by targeting and inhibiting the STING signaling pathway. *Carcinogenesis*, doi:10.1093/carcin/bgab001 (2021).

Figures

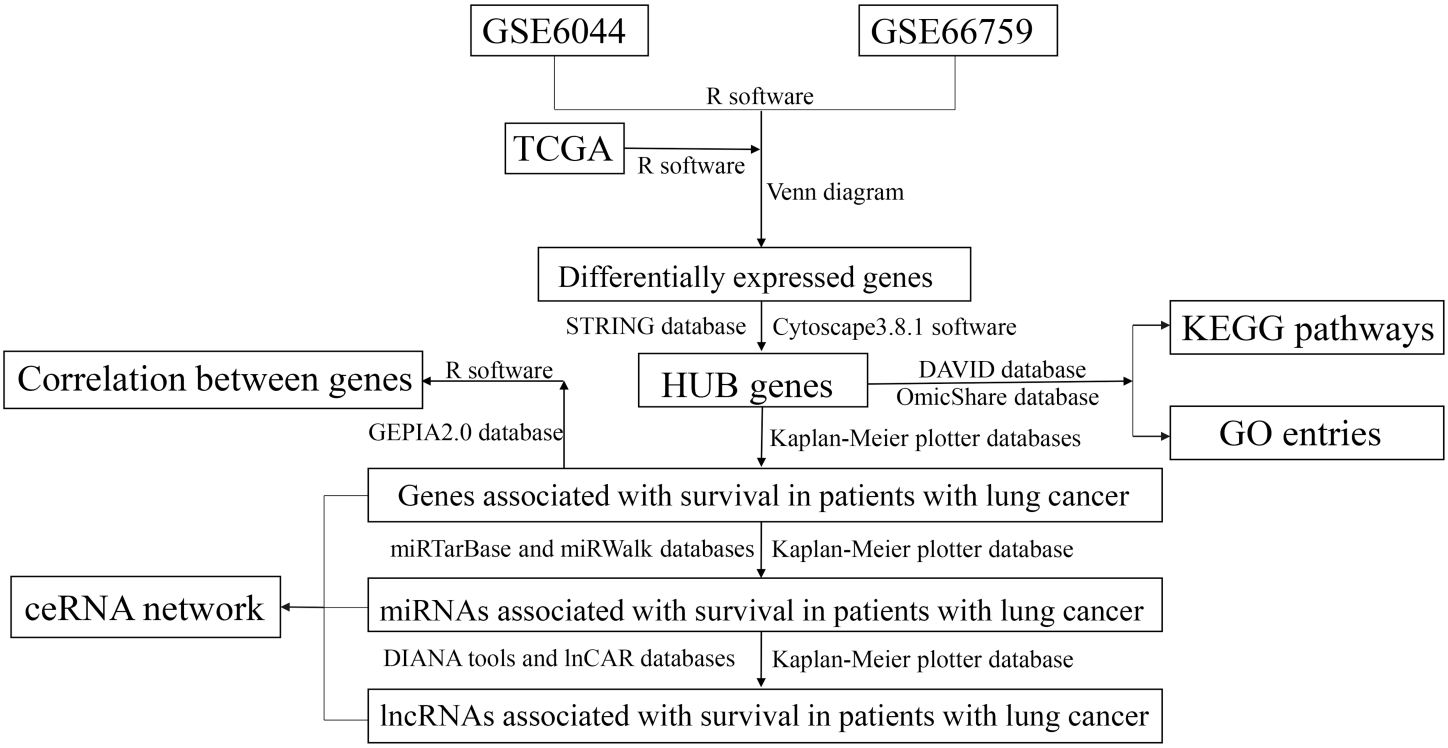

Figure 1

The flow chart of this study.

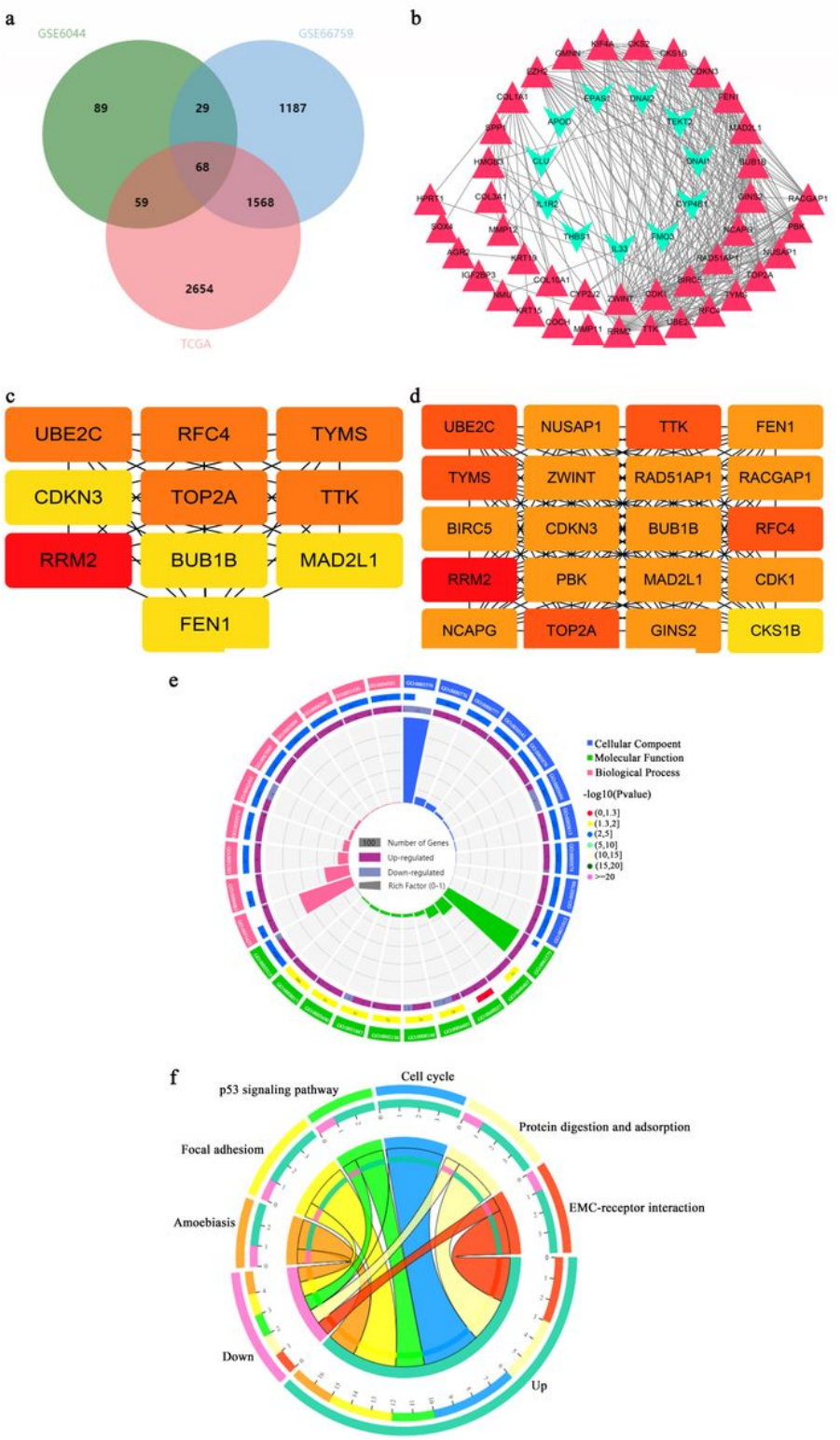

Figure 2

(a) Intersection DEGs of three datasets. (b) The interaction between DEGs. Red represented up-regulated DEGs and green represented down-regulated DEGs. (c) The top 10 HUB genes from cytoHubba. (d) The top 20 HUB genes from cytoHubba. (e) Go enrichment analysis of the top 20 HUB genes. (f) KEGG pathway enrichment analysis of the top 20 HUB genes.

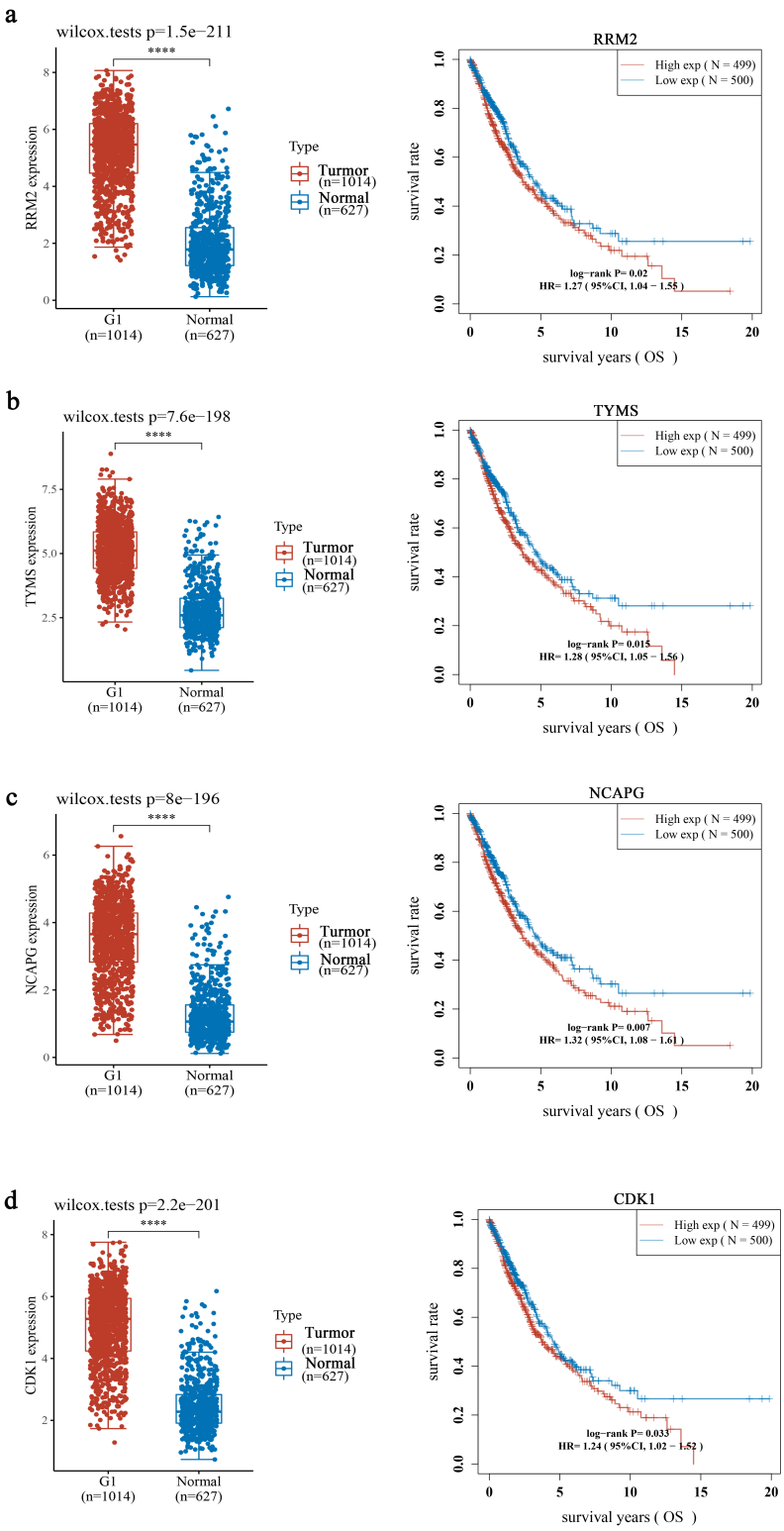

Figure 3

Kaplan–Meier survival and expression analysis of the top 20 HUB genes based on these three datasets. (a) RRM2. (b) TYMS. (c) NCAPG. (d) CDK1.

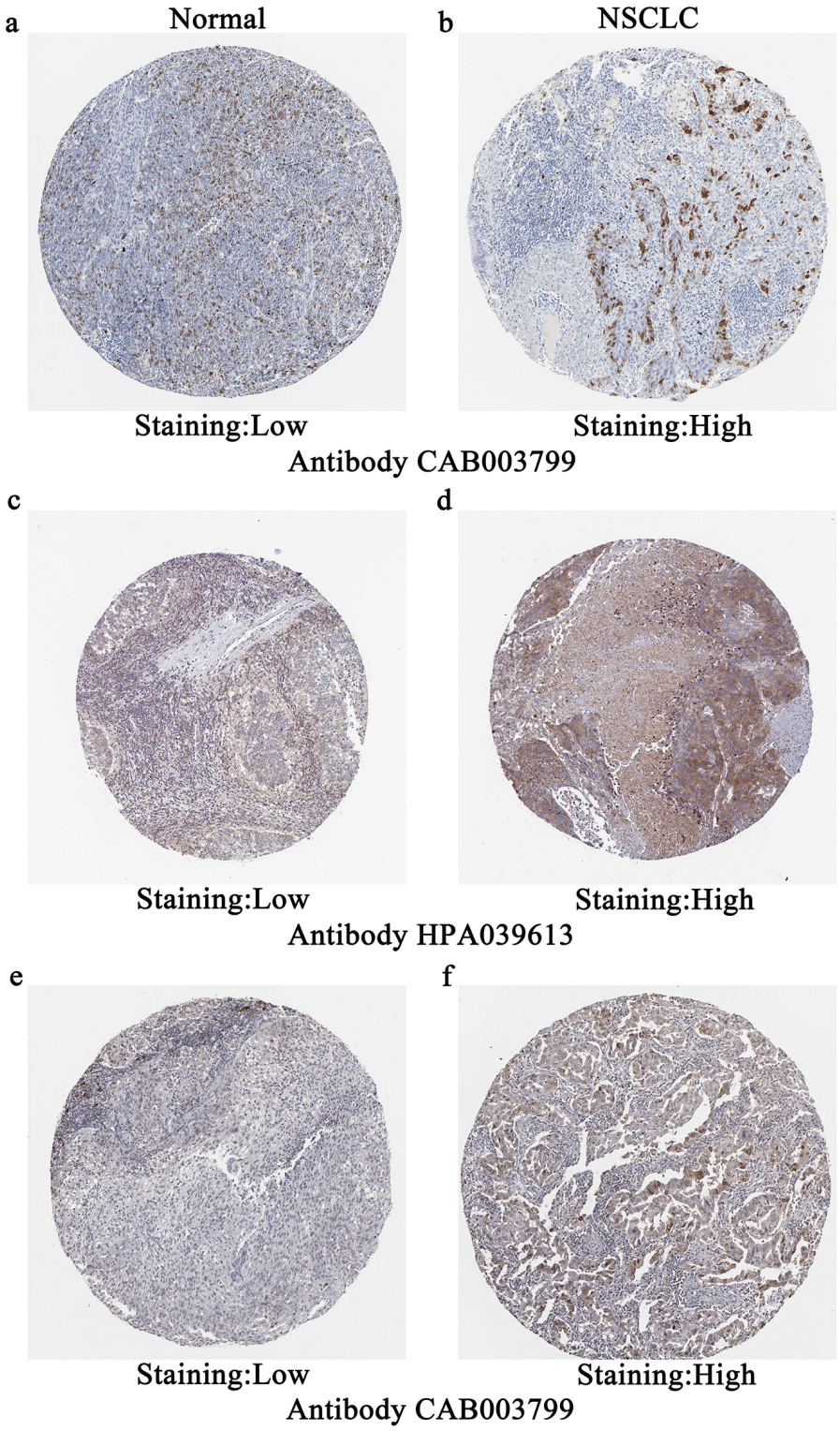

Figure 4

The expression of the 4 biomarkers in NSCLC patients and normal lung tissues based on the Human Protein Atlas database. (a-b) Expression of CDK1 in normal lung tissues and NSCLC tissues based on antibody CAB003799, respectively. (c-d) Expression of NCAPG in normal lung tissues and NSCLC tissues based on antibody HPA039613, respectively. (e-f) Expression of TYMS in normal lung tissues and NSCLC tissues based on antibody CAB003799, respectively.

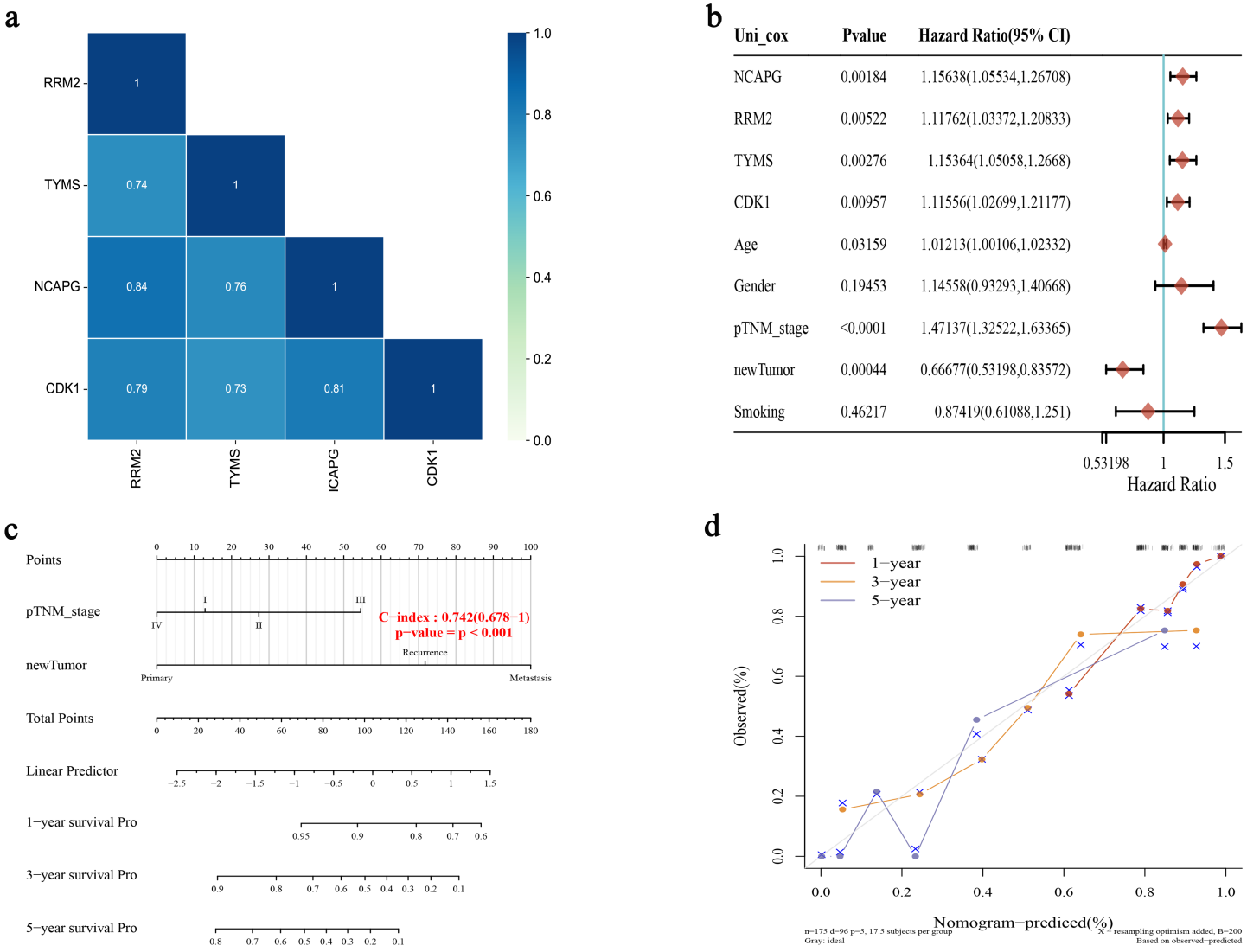

Figure 5

Analysis of the relationship among the 4 biomarkers and establishment of their prognostic models. (a) The correlation heat map of the 4 biomarkers. And the size of the number was proportional to the correlation between them. (b) Hazard ratio and P-value of constituents involved in univariate Cox regression and some parameters of the 4 biomarkers. (c) Nomogram to predict the 1-y□2-y and 3-y OS of NSCLC patients. (d) Calibration curve for the OS nomogram model in the discovery group. A dashed diagonal line represented the ideal nomogram, and the blue line□red line and orange line represented the 1-y□2-y and 3-y observed nomograms.

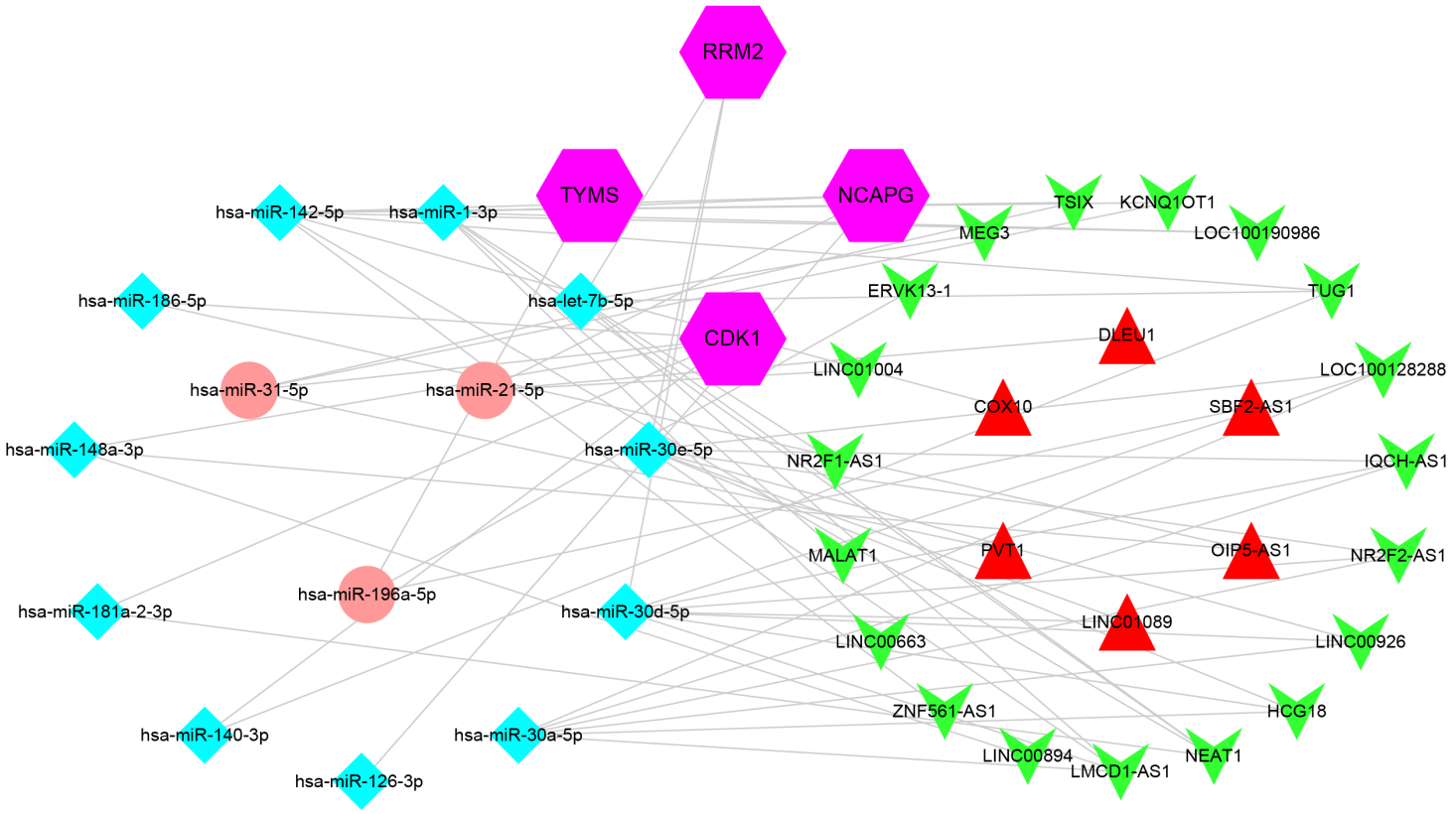

Figure 6

Construction of a ceRNA network significantly related to the OS of NSCLC. Different colors in the network represented different meanings. Among them, purple represented 4 up-regulated mRNAs, pink represented 3 up-regulated miRNAs, sky blue represents 11 down-regulated miRNAs, red represented 6 up-regulated lncRNAs, and green represented 19 down-regulated lncRNAs.

Supplementary Files

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

- Supplementaryinformation.pdf