Identification of The Potential Signature Genes Between Lung Adenocarcinoma and Lung Squamous Cell Carcinoma

Xue Li
Fudan University
zhenni wang (412680778@qq.com)
The Third People's Hospital of Qingdao
Jun Xie
Fudan University
Naishuo Zhu
Fudan University

Research article

Keywords: lung squamous cell carcinoma, lung adenocarcinoma, differentially expressed genes, bioinformatical analysis

Posted Date: April 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-386986/v1

Identification of the potential signature genes between lung adenocarcinoma and lung squamous cell carcinoma

Xue Li[1], Zhenni Wang[2], Jun Xie[1*], Naishuo Zhu[1*]

1 School of Life Sciences, Fudan University, Shanghai, 200433, China
2 The Third People's Hospital of Qingdao, Qingdao, 266041, China

*Correspondence should be addressed to Naishuo Zhu: nzhu@fudan.edu.cn and Jun Xie: xiejun@fudan.edu.cn

Acknowledgment:

This work was supported by National Natural Science Foundation of China (31370927 and 30571650), Natural Science Foundation of Zhejiang Province (CN) (LY-18H160059), and Natural Science Foundation of Shanghai (13431900602).

Ethical Approval:

This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of Interest:

The authors have no competing interests to declare.

Abstract

Objectives: The purpose of this study was to identify and compare the potential signature genes along with the key pathways related to the pathogenesis of lung adenocarcinoma and lung squamous cell carcinoma through bioinformatics analysis.

Methods: The differential expressions of miRNAs (DEmiRNAs) and mRNAs (DEmRNAs) in lung adenocarcinoma (AD) and lung squamous cell carcinoma (SC) were identified using the microarray data of 2 miRNA and 6 mRNA from the Gene Expression Omnibus (GEO) database. The interaction between the DEmiRNAs and DEmRNAs was explored, followed by the KEGG pathway enrichment, Gene Ontology annotation (GO) enrichment analysis, DEmiRNA-gene interactive network, and protein-protein interaction (PPI) network to predict the hub genes and identify the key pathways. The sequencing data was then downloaded from The Cancer Genome Atlas (TCGA) to validate the hub genes, perform survival analysis, construct the ROC curve along with identifying the regulatory networks of mRNA-miRNA-lncRNA. Finally, the resulting signature genes and significant regulatory networks in AD and SC were filtrated across these analyses.

Results: The DEmiRNAs and DEmRNAs were found to be significantly enriched in the phagosomes, human T-cell leukemia virus 1 infection, and ECM-receptor interaction in the AD, whereas in SC, they were enriched in the cell cycle and p53 signaling pathways. Furthermore, we

found that the hsa-miR-21-5p, hsa-miR-200a-5p, and COL1A1 may act as potential meaningful biomarkers for the AD while hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-205-5p, and FRMD6 may play an important role in SC. The HMGA1/hsa-miR-424-5p/PVT1 may be a significant regulatory network of AD while KIF11/hsa-miR-30d-5p/DLEU2, CCNE2/hsa-miR-30d-5p/DLEU2, and SPTBN1/hsa-miR-218-5p/MAGI2-AS3 may be the useful networks in regulating the SC progression.

Conclusion: This study provided the signature targets and theoretical basis for further research on the biomarkers and molecular mechanisms of the SC, AD, and NSCLC.

Keywords: lung squamous cell carcinoma, lung adenocarcinoma, differentially expressed genes, bioinformatical analysis

Introduction

Non-small cell lung cancer (NSCLC) accounts for approximately 80-85% of lung malignancy cases, including squamous cell carcinoma (SC), adenocarcinoma (AD), and large cell carcinoma [1]. Lung adenocarcinoma comprises 40% of all the lung cancer (LC) cases, and it arises from small airway epithelial type II alveolar cells, which secrete mucus and other substances [2, 3]. It tends to occur in the periphery of the lung and remains the main histological subtype of lung cancer in both smokers and nonsmokers, irrespective of their age and gender [4]. Lung squamous cell carcinoma accounts for about 30% of lung cancer, and it develops from early versions of squamous cells or epithelial cells in the airway of the bronchial tubes found in the center of the lungs. A large amount of data indicates that smoking, gender, and age are the crucial factors that influence the risk of SC. For example, 80% of lung cancer deaths are attributed to cigarette smoking, where most of them are male and aged around 50 years [5, 6].

Compared to SCLC, the cells in NSCLC tend to grow and divide more slowly, leading to relatively late spreading and metastasizing of the tumor. However, more than 70% of these cases are diagnosed as unresectable advanced stage tumors [7]. Despite many intervention methods being proposed, the prognosis for the NSCLC patients remains poor, with a high recurrence rate even after the treatment. Recently, both targeted therapy and immunotherapy have appeared as genetic alteration-guided targeted therapies. These include TK inhibitors (TKIs) that target the EGFR and the antibody-directed therapies that target CTLA-4, PD-1, and PD-L1 to show remarkable early success in the management of advanced NSCLC [8-10]. To improve the clinical outcomes of lung cancer patients, biomarkers are urgently required to accurately subclassify lung cancer and also design precisely targeted therapy for lesions.

Technologies such as DNA microarray and sequencing can easily and quickly obtain a lot of data. However, due to the differences in the production methods, heterogeneity becomes a certain influencing factor between different technologies, which may hinder a high-throughput comprehensive analysis [11]. Therefore, to reduce the heterogeneity, we used 4 miRNA microarray and 6 mRNA microarray data to synthetically select hub genes for AD and SC, which was followed by downloading of the sequencing profiles from TCGA to validate the previously selected hub genes by microarray. After processing and strict analysis, the signature genes and key signaling pathways were screened out to explore an in-depth understanding of the molecular mechanism of lung cancer, which provided therapeutic targets for drug designing and molecular

markers for clinical diagnosis.

Materials and methods

Collection of Datasets and analyses of DEGs

The public gene expression profiles of AD and SC were obtained from the GEO database (http://www.ncbi.nlm.nih.gov/geo/), which included miRNA microarray datasets from GSE74190 and GSE51855 for both AD and SC along with mRNA microarray datasets from GSE32863, GSE75037, GSE116959 and GSE30219, GSE33479, GSE51852 for AD and SC, respectively. The data were collected, and all the samples of lung cancer tissues were compared with the normal samples. The detailed information of the microarray datasets is shown in Table 1. The workflow demonstrating the identification of core genes and pathways is presented in Figure 1.

The statistical analyses were carried out using the R (version 4.0.3) software. All the data normalization and differential gene expression analyses were performed using the limma package [12]. The non-paired t-test was used to calculate the p-values to screen differentially expressed miRNA and mRNA (DEmiRNA and DEmRNA), and the threshold was set as $|\log FC| \geq 1$ and adjusted P-values < 0.05.

Accession number	Type	Platform	Experimental group (patient group)	Control group (normal group)
GSE74190	miRNA	GPL19622	AD: 35, SC: 30	HC: 44
GSE51855	miRNA	GPL7341	AD: 76, SC: 29	HC: 5
GSE32863	mRNA	GPL6884	AD: 58	HC: 58
GSE75037	mRNA	GPL6884	AD: 83	HC: 83
GSE116959	mRNA	GPL17077	AD: 57	HC: 11
GSE30219	mRNA	GPL570	SC: 61	HC: 14
GSE33479	mRNA	GPL6480	SC: 14	HC: 13
GSE51852	mRNA	GPL6480	SC: 28	HC: 4

Table 1: Basic information about the microarray datasets used in this research, HC: healthy control group; AD: adenocarcinoma; SC: squamous cell carcinoma.

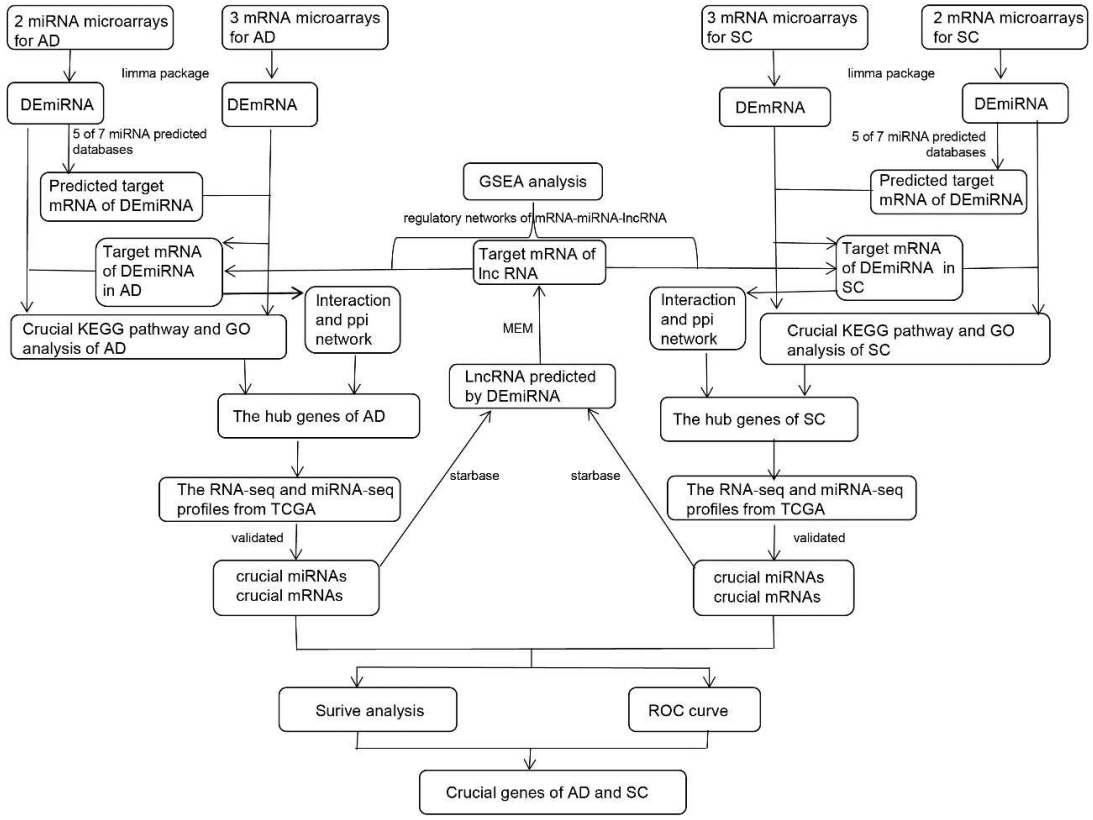


Figure 1: Workflow demonstrating the identification of core genes and pathways.

Relationship pairing between DEmiRNA and DEmRNA

The predicted target genes from the DEmiRNA were obtained by using at least five out of the seven commonly used databases, including miRWalk [13], starBase [14], miRDB [15], miRanda, Targetscan [16], Tarbase v.8 [17], and miTarbase [18]. The predicted target genes from DEmiRNA were matched with the DEmRNA selected by the microarray analysis to obtain the target gene from DEmiRNA.

KEGG pathway and GO analysis

The Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis and Gene Ontology annotation enrichment (GO) analysis of DEmiRNA and DEmRNA were performed using the R clusterprofiler package [19], and the threshold values were selected as gene count ≥ 2 and adjusted P-values < 0.05.

Constructing the DEmiRNA-gene interaction and PPI network

As per the interaction information obtained between the DEmiRNA and its target genes, the interaction network was constructed using Cytoscape 3.7.0 software to obtain a genetic relationship map [20].

The STRING database was used to predict the protein-protein interaction (PPI) network for the the target genes of DEmiRNA. The species used was homo sapiens and the parameter of PPI was set as 0.4 [21].

Validation of hub genes

The significant mRNA and miRNA were selected based on the KEGG analysis, GO analysis, interaction network, and PPI network analysis. Then, the RNA-seq and miRNA-seq data were downloaded from The Cancer Genome Atlas (TCGA) to validate the hub genes, which were previously selected by the microarray analysis from the GEO database. The resulting data were treated using the edgeR package [22]. The genes having a threshold false discovery rate (FDR) of < 0.05 and $|\log FC|$ of ≥ 1 were considered as statistically significant differential genes.

Survival analysis and ROC curve of the hub genes

The hub mRNA and miRNA obtained after the validation by TCGA were used to perform the overall survival analysis using the R package of survival. The HR and log-rank p-value was calculated and presented on the Kaplan-Meier curves.

The receiver operating characteristic (ROC) curve was used to judge the diagnostic value of the hub genes previously validated by TCGA. The area under the curve (AUC) and the p-value were calculated and presented on the plot.

Prediction and construction of the regulatory networks of mRNA-miRNA-lncRNA

The DEmiRNAs obtained after the validation by TCGA were submitted to the starbase database (http://starbase.sysu.edu.cn/) to predict the lncRNA, with the chip dataset at medium stringency (>= 2) [14]. Then the selected lncRNAs were used to predict the interactive mRNA by Multi Experiment Matrix (MEM, https://biit.cs.ut.ee/mem/) [23] with the output limit set as the top 50. Finally, the target mRNA predicted by the lncRNA was matched with the target mRNA of miRNA to get the common target genes that might interact with both miRNA and lncRNA.

A correlation analysis was conducted among the selected mRNAs, miRNAs, and lncRNAs using the R software through Pearson's correlation analysis, and $p < 0.05$ was defined as the statistically significant threshold. The obtained statistically significant mRNAs, miRNAs, and lncRNAs were then used to create a network between them (mRNA-miRNA-lncRNA).

To explore the cancer-related pathways associated with the gene expression levels in these regulatory networks, the Gene Set Enrichment Analysis (GSEA) was performed. The reference gene set used in this study was the c2.cp.kegg.v7.2.symbols.gmt. A gene set was considered as significantly enriched if the normal p-value and FDR were both found to be less than 0.05.

Result

Differentially expressed genes

Table 2 shows differentially expressed genes obtained by microarray datasets after normalization. In AD, 13 DEmiRNAs were up-regulated while 11 DEmiRNAs were down-regulated whereas, in the SC, 19 DEmiRNAs were up-regulated while 23 DEmiRNAs were down-regulated. Similarly, in the AD, 186 DEmRNAs were up-regulated while 482 DEmRNAs were down-regulated whereas, in the SC, 337 DEmRNAs were up-regulated and 402 DEmRNAs were down-regulated. The predicted target genes obtained using the databases from the DEmiRNAs were found to be 3135 and 4364 in AD and SC, respectively. The predicted target genes from the DEmiRNAs were matched with DEmRNAs, which resulted in 29 genes showing significant up-regulation while 82 genes showed significant down-regulation in the AD whereas, in the SC, 97 genes showed significant up-regulation while 113 genes showed significant

down-regulation. The volcano of these DEGs and the principal component analysis (PCA) are shown in Supplementary Figure S1 and Figure S2, respectively.

Microarray datasets	Up	Down
GSE74190	AD: 32, SC: 55	AD: 35, SC: 65
GSE51855	AD: 16, SC: 29	AD: 22, SC: 39
GSE32863	AD: 548	AD: 779
GSE75037	AD: 1626	AD: 1811
GSE116959	AD: 742	AD: 1399
GSE30219	SC: 1321	SC: 1780
GSE33479	SC: 1505	SC: 1713
GSE51852	SC: 2410	SC:2555
GSE74190+GSE51855	AD: 13, SC: 19	AD: 11, SC: 23
GSE32863+GSE75037+GSE116959	AD:186	AD: 482
GSE30219+GSE33479+GSE51852	SC: 337	SC: 402
GSE32863+GSE75037+GSE116959+predicted_mRNA	AD: 29	AD: 82
GSE30219+GSE33479+GSE51852+predicted_mRNA	SC: 97	SC: 113

Table 2. Differentially expressed genes obtained from the microarray datasets predicting the target mRNA: the predicted target genes from DEmiRNAs obtained by the databases. HC: healthy control group; AD: adenocarcinoma; SC: squamous cell carcinoma; predicted_mRNA: the predicted target genes obtained using the databases from the DEmiRNAs.

KEGG pathway enrichment analysis

Figure 2 shows the KEGG pathways enriched by the DEmRNA and DEmiRNA. The human T-cell leukemia virus 1 infection, the ECM-receptor interaction, and phagosome were significantly enriched by these two prediction groups (DEmRNA and DEmiRNA) in the AD whereas, in the SC, both cell cycle and p53 signaling pathways were significantly enriched. 5 up-regulated DEmiRNAs (hsa-miR-21-5p, hsa-miR-130b-3p, hsa-miR-200a-5p, hsa-miR-200b-3p, hsa-miR-424-5p) along with 4 down-regulated DEmiRNAs (hsa-miR-101-3p, hsa-miR-143-3p, hsa-miR-195-5p, hsa-miR-497-5p) were significantly enriched in the AD KEGG pathways. Similarly, 3 up-regulated DEmiRNAs (hsa-miR-21-5p, hsa-miR-141-3p, hsa-miR-200c-3p) and 5 down-regulated DEmiRNAs (hsa-miR-1-3p, hsa-miR-29c-3p, hsa-miR-181c-5p, hsa-miR-195-5p, hsa-miR-497-5p) were significantly enriched in the SC KEGG pathways.

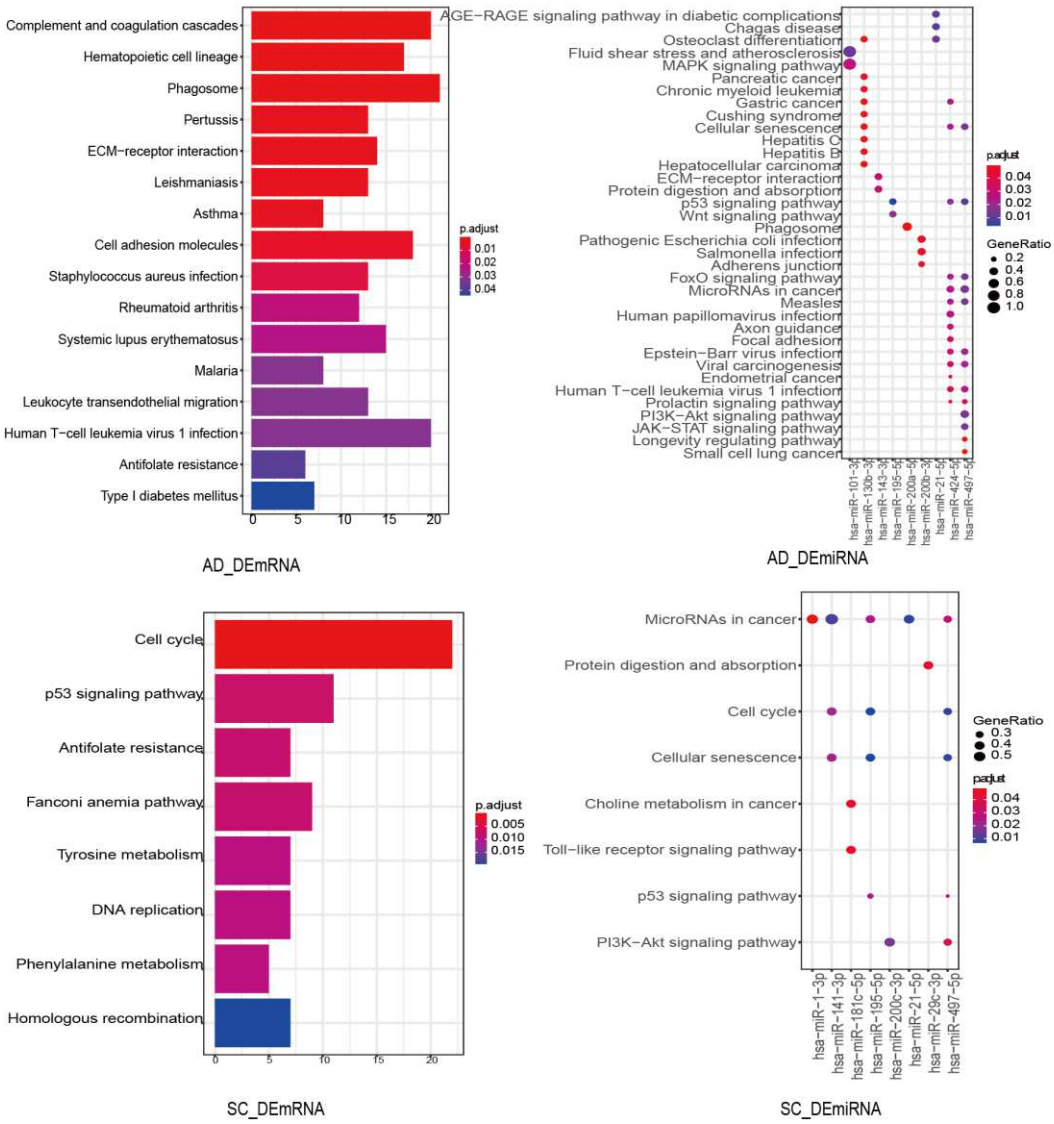


Figure 2. The KEGG pathways enriched by the DEmRNA and DEmiRNA of AD and SC.

GO analysis

Supplementary Table S1 presents the GO analysis of the AD and SC using DEmRNA and DEmiRNA, which is further divided into three functional groups, including biological process group (BP), cellular component group (CC), and molecular function group (MF). The table shows the coexisting terms among these three analytical DEG groups. Among these, 6 up-regulated DEmiRNAs (hsa-miR-21-5p, hsa-miR-130b-3p, hsa-miR-183-5p, hsa-miR-200a-5p, hsa-miR-429, hsa-miR-424-5p, hsa-miR-182-5p, hsa-miR-200b-3p) and 6 down-regulated DEmiRNAs (hsa-miR-101-3p, hsa-miR-143-3p, hsa-miR-145-5p, hsa-miR-218-5p, hsa-miR-195-5p, hsa-miR-497-5p) were significantly enriched in the GO analysis of AD while 13 up-regulated DEmiRNAs (hsa-miR-9-5p, hsa-miR-21-5p, hsa-miR-130b-3p, hsa-miR-149-5p, hsa-miR-205-5p, hsa-miR-224-5p, hsa-miR-429, hsa-miR-200b-3p, hsa-miR-96-5p, hsa-miR-141-3p, hsa-miR-183-5p, hsa-miR-196b-5p, hsa-miR-221-3p) and 12 down-regulated DEmiRNAs (hsa-miR-144-3p, hsa-miR-145-5p, hsa-miR-195-5p, hsa-miR-218-5p, hsa-miR-497-5p, hsa-miR-1-3p, hsa-miR-29c-3p, hsa-miR-30b-5p, hsa-miR-130a-3p, hsa-miR-101-3p,

hsa-miR-143-3p, hsa-miR-181a-5p) were significantly enriched in the GO analysis of SC. According to the GO analysis, cardiac septum morphogenesis, proteinaceous extracellular matrix, and the growth factor binding coexisted in the AD and SC.

The interaction network between DEmiRNAs and its target genes

The interaction network was constructed between the DEmiRNA and its target genes as shown in Figure 3, which included 131 nodes and 195 interactions in AD and 244 nodes and 481 interactions in SC. The node with high topological score was regarded as hub genes of the interaction network. In the AD, the top 10 high degree DEmiRNA were hsa-miR-182-5p, hsa-miR-195-5p, hsa-miR-218-5p, hsa-miR-497-5p, hsa-miR-424-5p, hsa-miR-21-5p, hsa-miR-200b-3p, hsa-miR-145-5p, hsa-miR-625-5p, and hsa-miR-130b-3p while the top 10 high degree DEmRNA included RECK, TGFBR3, SESN1, PSAT1, PIK3R1, CCND2, KLF9, PDIA6, TACC1, and EFNB2. In the SC, the top 10 high degree DEmiRNA were hsa-miR-9-5p, hsa-miR-182-5p, hsa-miR-181a-5p, hsa-miR-195-5p, hsa-miR-21-5p, hsa-miR-497-5p, hsa-miR-145-5p, hsa-miR-30b-5p, hsa-miR-200b-3p, and hsa-miR-30d-5p while the top 10 high degree DEmRNA included LIFR, FRMD6, FIGN, CCNE2, FOXP1, DDAH1, TGFBR3, PRUNE2, ENPP4, and NAA50.

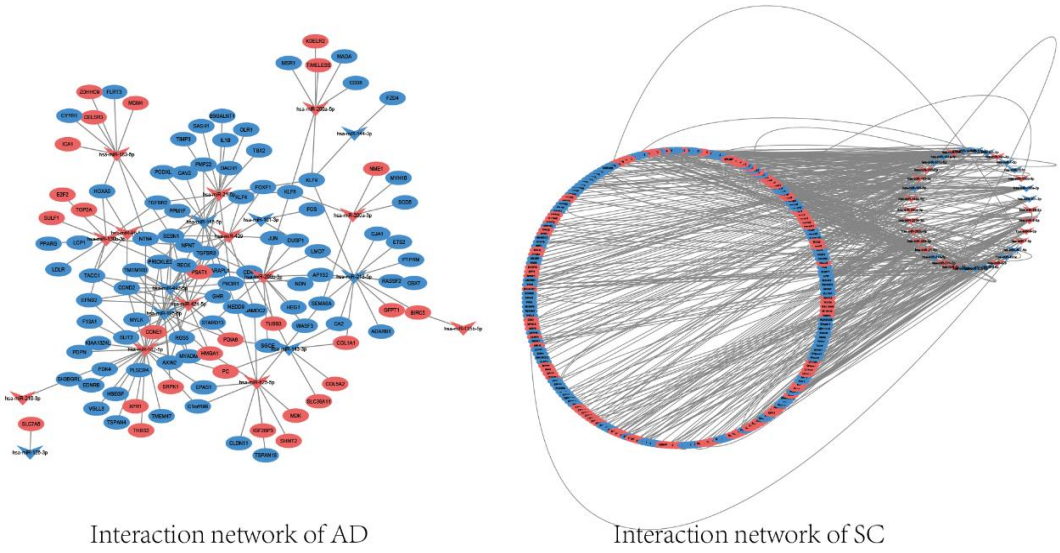
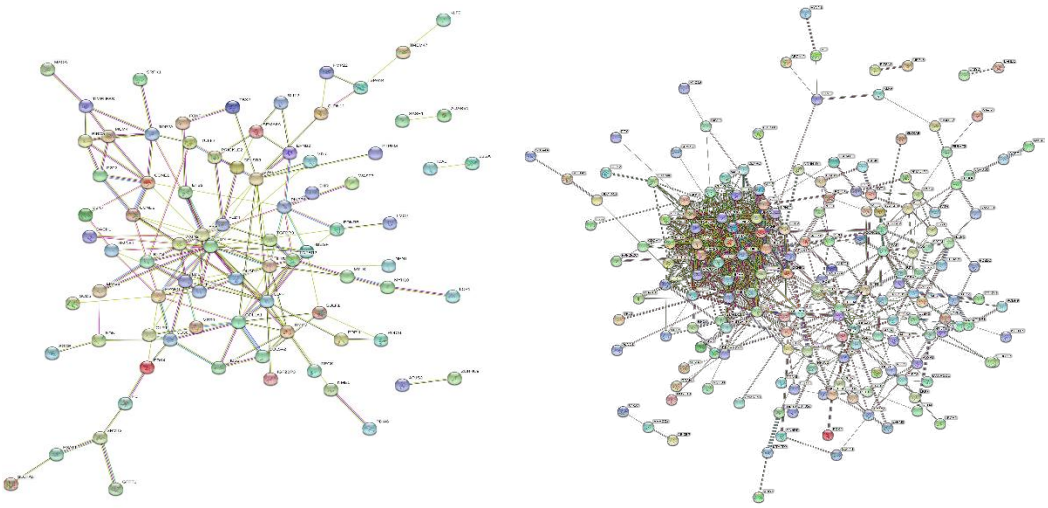


Figure 3. Interaction networks between the DEmiRNA and its target genes in the AD and SC. Red represents up-regulated expression, while blue represents down-regulated expression. The arrow represents miRNA, while the ellipse represents mRNA.

PPI network

Figure 4 presents the PPI network of the target genes obtained from the DEmiRNAs. It consisted of 82 nodes and 164 edges in AD, and 171 nodes and 707 edges in SC. In the AD, the top 10 high degree mRNA included JUN, CD44, FOS, IL1B, COL1A1, PPARG, GJA1, LDLR, PIK3R1, and CCNE1 while in the SC, the top 10 high degree mRNA included CDK1, CCNA2, CCNB1, EZH2, KIF11, CHEK1, BIRC5, TOP2A, KIF23, and NCAPG.



PPI network of AD PPI network of SC

Figure 4. The PPI networks in the AD and SC.

Validation of the hub genes

The significant genes in the AD and SC, previously obtained from the microarray data were further selected by combining the KEGG analysis, GO analysis, interaction network, and PPI network analysis. Table 3 shows these hub mRNAs and miRNAs after the validation by TCGA. To select the mRNAs in AD, 510 patients and 58 healthy controls were analyzed while for SC, 496 patients and 49 healthy controls were analyzed. For the selection of miRNAs in AD, 510 patients with 45 healthy controls were analyzed, whereas for miRNA in SC, 473 patients with 45 healthy control were analyzed. The miRNAs and mRNAs with higher differential expression levels in AD compared to SC were observed in 3 miRNAs, including hsa-miR-21-5p, hsa-miR-625-5p, and hsa-miR-200a-5p, and 5 mRNAs, including PIK3R1, CCND2, EFNB2, GJA1, and COL1A1. Similarly, the miRNAs and mRNAs with higher differential expression levels in SC compared to AD were observed in 15 miRNAs, including hsa-miR-149-5p, hsa-miR-200c-3p, hsa-miR-29c-3p, hsa-miR-30b-5p, hsa-miR-145-5p, hsa-miR-181a-5p, hsa-miR-181c-5p, hsa-miR-497-5p, hsa-miR-30d-5p, etc., and 6 mRNAs, including FRMD6, FOXP1, DDAH1, PRUNE2, ENPP4 and NAA50. Supplementary Figure S3 shows the expression plots of the selected hub genes in both AD and SC, where the data was derived from TCGA.

	Gene expression: just in AD[a]	Gene expression: AD > SC[b]	Gene expression: AD = SC[c]	Gene expression: AD < SC[d]	Gene expression: just in SC[e]
Up-regulated miRNA	hsa-miR-21-5p, hsa-miR-625-5p	hsa-miR-200a-5p	hsa-miR-183-5p, hsa-miR-182-5p, hsa-miR-424-5p, hsa-miR-96-5p, hsa-miR-224-5p	hsa-miR-9-5p, hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-196b-5p, hsa-miR-205-5p	hsa-miR-149-5p, hsa-miR-200c-3p
Down-regulated miRNA	-	-	hsa-miR-101-3p, hsa-miR-143-3p, hsa-miR-195-5p, hsa-miR-218-5p, hsa-miR-144-3p, hsa-miR-1-3p	-	hsa-miR-29c-3p, hsa-miR-30b-5p, hsa-miR-145-5p,hsa-miR-181a-5p, hsa-miR-181c-5p, hsa-miR-497-5p, hsa-miR-30d-5p

Up-regulated mRNA	-	COL1A1	PSAT1, PDIA6, CCNE1, CCNE2, CDK1, CCNA2, CCNB1, EZH2, KIF11, CHEK1, BIRC5, TOP2A, KIF23, NCAPG	-	FRMD6, NAA50
Down-regulated mRNA	PIK3R1, CCND2, EFNB2, GJA1	-	RECK, TGFBR3, SESN1, KLF9, TACC1, JUN, FOS, IL1B, LDLR, PPARG, LIFR	-	FOXP1, DDAH1, PRUNE2, ENPP4

Table 3. Differentially expressed genes obtained after the validation by TCGA. [a] the hub genes with FDR < 0.05 and |logFC| ≥1 expressed differentially just in the AD. [b] the genes expressed both in AD and SC but the difference in the expression levels of hub genes between the patients and healthy control in AD was higher than SC. [c] the genes that had the same difference in the expression levels of hub genes in AD and SC. [d] the genes expressed both in AD and SC with the difference in the expression levels of hub genes in SC being higher than the AD. [e] the hub genes with FDR < 0.05 and |logFC| ≥1 expressed differentially just in the SC.

Survival analysis

The hub genes validated by TCGA were used to perform the overall survival analysis. Table 4 shows the meaningful hub genes in the overall survival of the patients with AD and SC. The genes with the expression in AD > SC including hsa-miR-21-5p, hsa-miR-200a-5p, COL1A1, PIK3R1, and CCND2 are closely related to the overall survival in AD, while the genes with the gene expression in AD < SC including hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-205-5p, FRMD6, FOXP1, DDAH1, and ENPP4 are associated with the overall survival in SC. Figure 5 presents the Kaplan-Meier curves of these hub genes with either gene expression being AD > SC or AD < SC.

	OS analysis in AD, p < 0.05	p < 0.05 and AD > SC	OS analysis in SC, p < 0.05	p < 0.05 and AD < SC
miRNA	hsa-miR-21-5p, hsa-miR-200a-5p, hsa-miR-429, hsa-miR-101-3p, hsa-miR-195-5p	hsa-miR-21-5p, hsa-miR-200a-5p	hsa-miR-182-5p, hsa-miR-183-5p, hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-205-5p, hsa-miR-144-3p, hsa-miR-195-5p, hsa-miR-218-5p, hsa-miR-1-3p,	hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-205-5p
mRNA	COL1A1, CCNE1, PIK3R1, CCND2, PPARG, RECK, SESN1, FOS	COL1A1, PIK3R1, CCND2	CDK1, CCNA2, EZH2, KIF11, CHEK1, TOP2A, FRMD6, FOXP1, DDAH1, ENPP4	FRMD6, FOXP1, DDAH1, ENPP4

Table 4. The meaningful hub genes for overall survival in AD and SC.

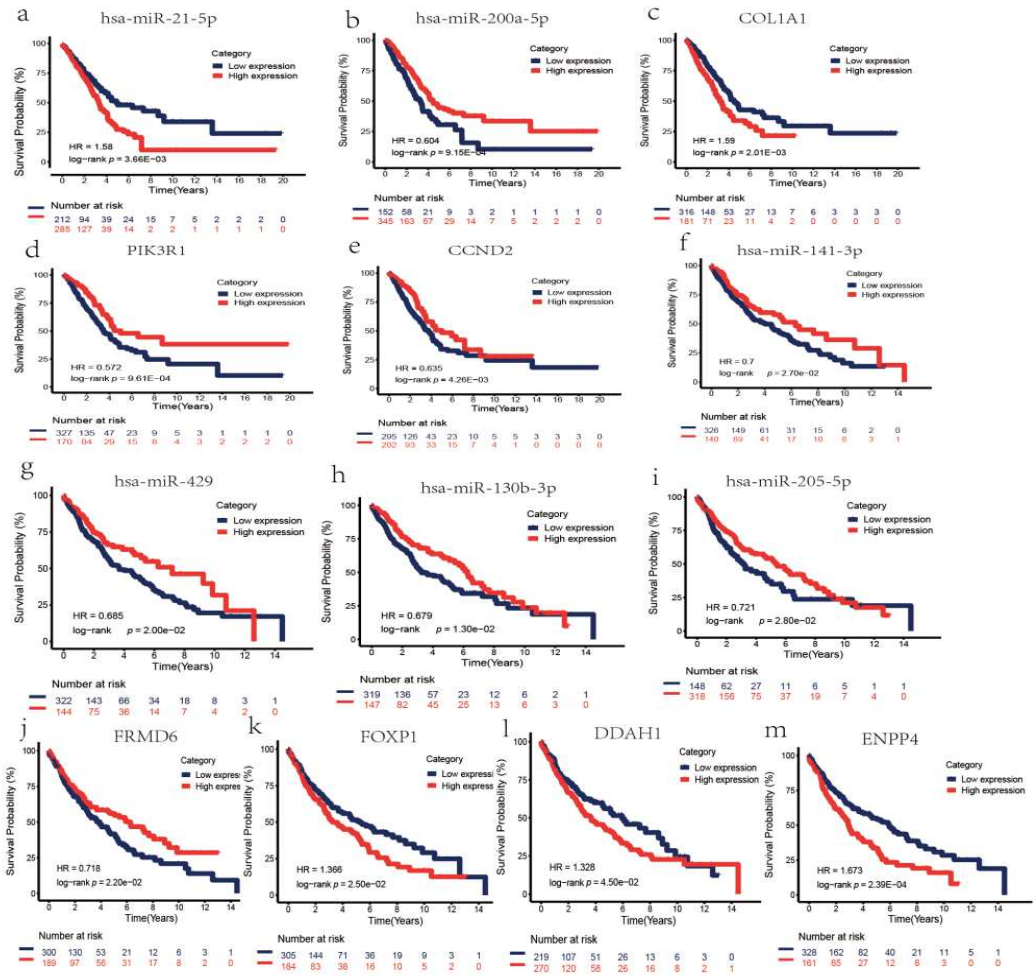


Figure 5. The overall survival analysis using the hub genes, where either the gene expression in AD > SC or AD < SC. a-e: genes in the AD; f-m: genes in the SC.

The ROC curve

The diagnostic value of the hub genes, previously validated by TCGA was confirmed by the ROC curves. Table 5 shows the meaningful hub genes under a p-value of < 0.05. The hsa-miR-21-5p, hsa-miR-625-5p, hsa-miR-200a-5p, and COL1A1 with high AUC suggested a significant diagnostic value in AD while hsa-miR-149-5p, hsa-miR-200c-3p, hsa-miR-9-5p, hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-196b-5p, hsa-miR-205-5p, FRMD6, and NAA50 revealed great diagnostic value in SC. The ROC curves of the meaningful hub genes are shown in Figure 6.

	The p of ROC < 0.05, AD > SC	The p of ROC < 0.05, AD < SC
miRNA	hsa-miR-21-5p, hsa-miR-625-5p, hsa-miR-200a-5p	hsa-miR-149-5p, hsa-miR-200c-3p, hsa-miR-9-5p, hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-196b-5p, hsa-miR-205-5p,
mRNA	COL1A1	FRMD6, NAA50

Table 5. The hub genes for AD and SC under the meaningful p-values. AD > SC: the gene expression in AD is higher than SC; AD < SC: the gene expression in AD is less than SC;

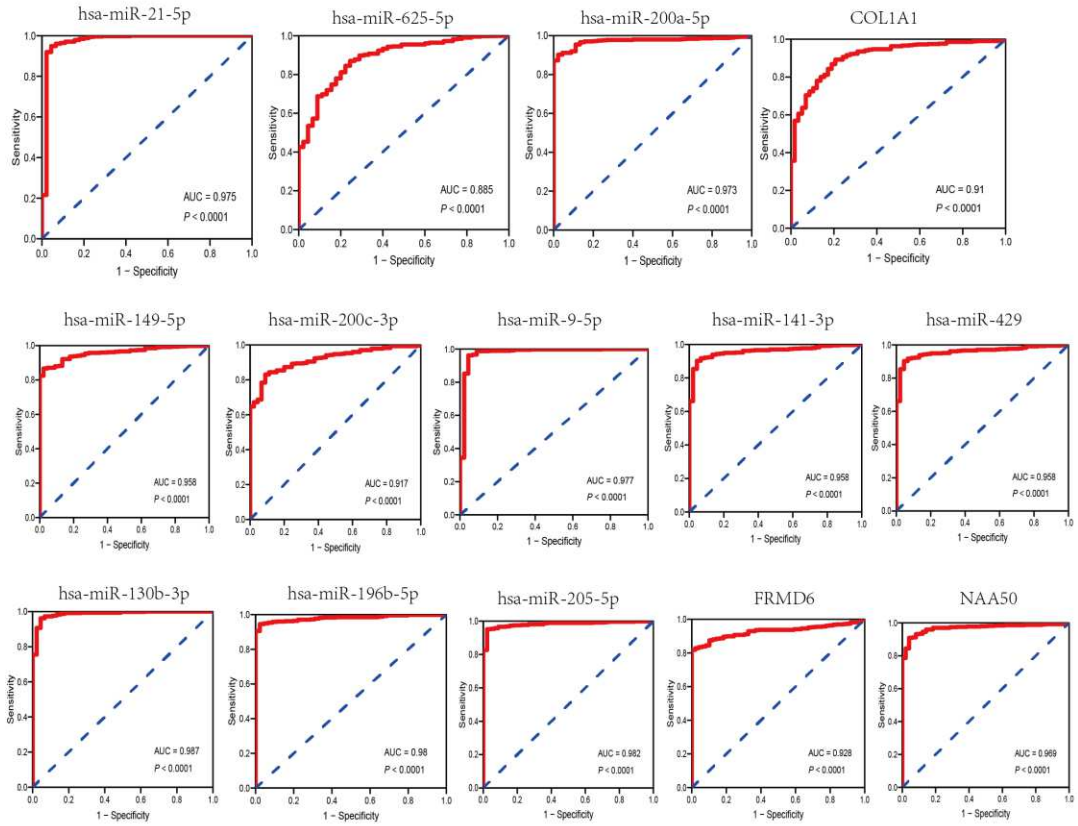


Figure 6. The ROC curves of the hub genes.

Construction of regulatory networks for mRNA-miRNA-lncRNA

Based on Pearson's correlation analysis, the probable interactive relationship among the mRNAs, miRNAs, and lncRNAs are shown in Table 6 and Figure 7. In AD, the correlation among HMGA1, hsa-miR-424-5p, and PVT1 mostly revealed a probable significant regulatory network while in the SC, the correlations among KIF11, hsa-miR-30d-5p, and DLEU2; CCNE2, hsa-miR-30d-5p, and DLEU2; SPTBN1, hsa-miR-218-5p, and MAGI2-AS3 were more obvious than the others, suggesting that all of them might function as regulators in SC. Figure 8 shows the expression and correlation plots of the regulatory network in AD, while Figure 9 shows the expression and correlation plots of regulatory networks in SC. Categorizing each regulatory network as one group, the KEGG pathway was found to be enriched in all the mRNA, miRNA, and lncRNA of each group. The related GSEA of each regulatory network is presented in Table 7.

	mRNA	miRNA	lncRNA
AD	HMGA1	hsa-miR-424-5p	PVT1
	PSAT1	hsa-miR-424-5p	LINC00662
	TGFBR2	hsa-miR-130b-3p	NEAT1
SC	KIF11	hsa-miR-30d-5p	DLEU2
	CCNE2	hsa-miR-30d-5p	DLEU2
	SPTBN1	hsa-miR-218-5p	MAGI2-AS3
	ADARB1	hsa-miR-218-5p	KCNQ1OT1
	ADARB1	hsa-miR-181a-5p	KCNQ1OT1
	PPP1R9A	hsa-miR-181a-5p	SNHG14

	ADARB1	hsa-miR-181c-5p	KCNQ1OT1
	PPP1R9A	hsa-miR-181c-5p	SNHG14
	COL1A1	hsa-miR-143-3p	MEG3
	SOBP	hsa-miR-143-3p	MEG3
	RIT1	hsa-miR-143-3p	LINC00662
	HELLS	hsa-miR-205-5p	MCM3AP-AS1
	ZBTB20	hsa-miR-182-5p	NEAT1
	TGFBR2	hsa-miR-130b-3p	NEAT1
	TGFBR2	hsa-miR-9-5p	NEAT1

Table 6. The probable regulatory networks of mRNA-miRNA-lncRNAs.

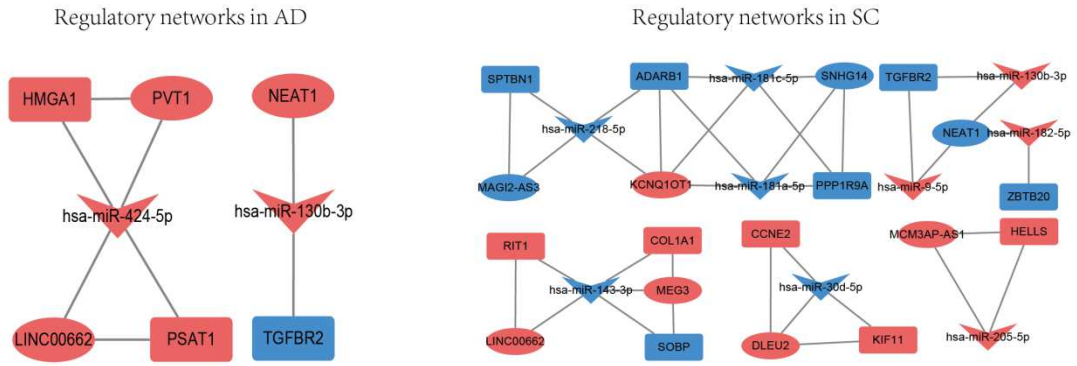


Figure 7. The probable regulatory networks of mRNA-miRNA-lncRNA. Red represents up-regulated expression, while blue represents down-regulated expression. The arrow represents miRNA, the rectangle represents mRNA, while the ellipse represents lncRNA.

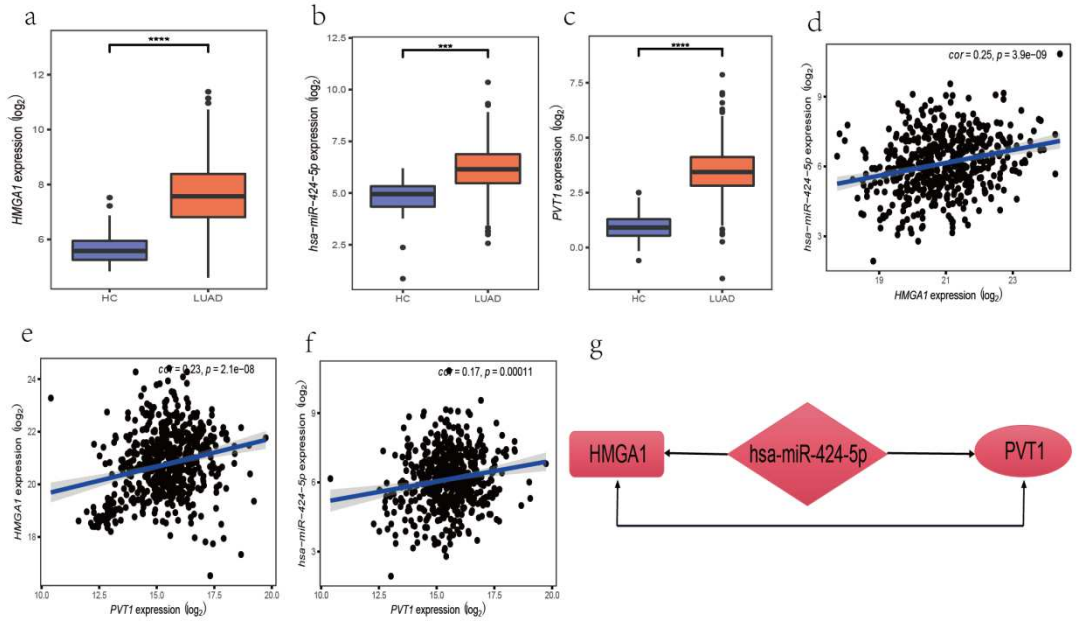


Figure 8. The expression and correlation plots of the competitive endogenous RNA (ceRNA) network in the AD. a-c: the expression plots of HMGA1, hsa-miR-424-5p, and PVT1 in AD. d-f: the correlations found in the AD. h: the regulatory network in the AD. The arrow direction

represents the predicted direction. The rhombus represents miRNA, the rectangle represents mRNA, while the ellipse represents lncRNA. Red represents up-regulated expression, while blue represents down-regulated expression.

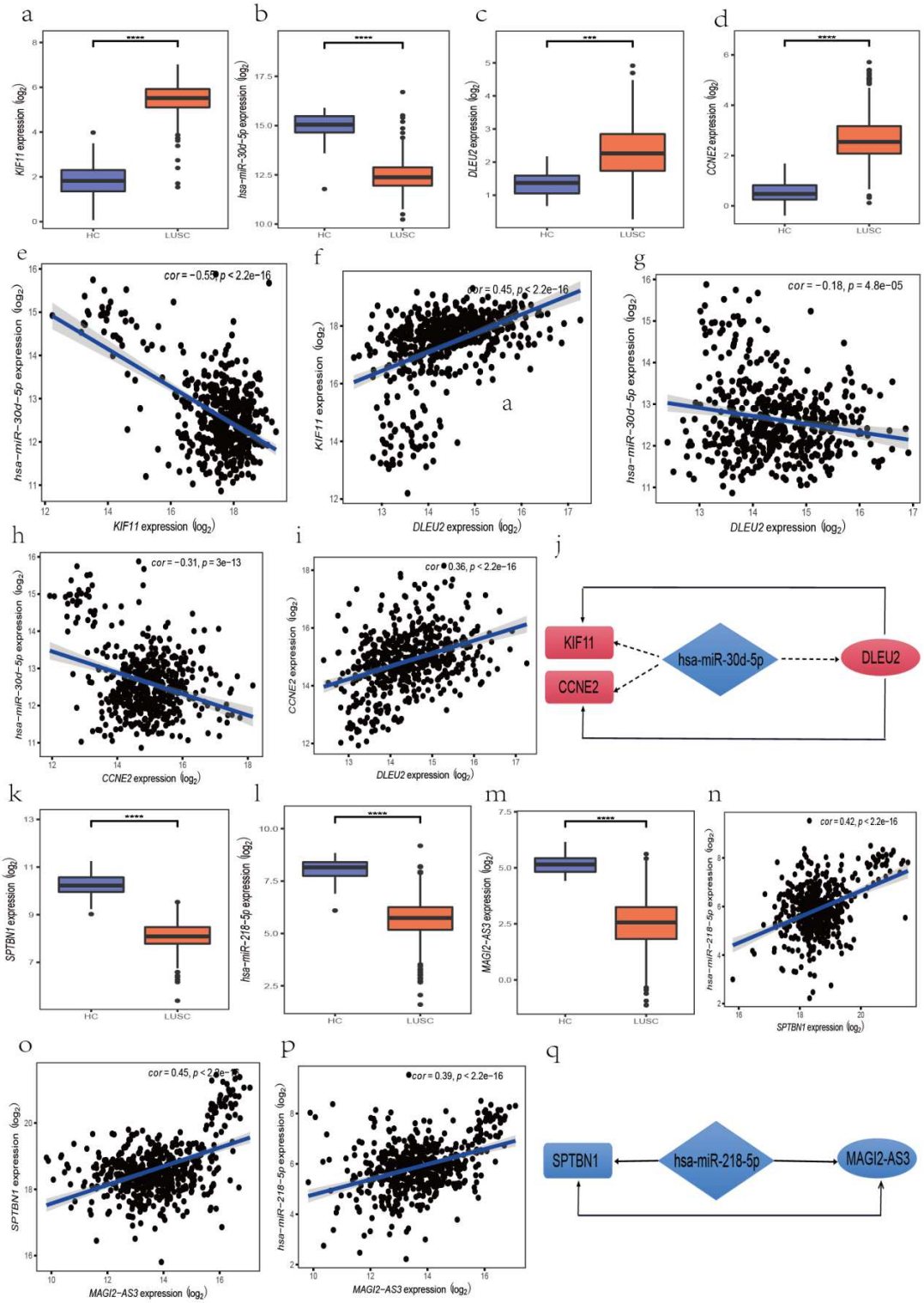


Figure 9. The expression and correlation plots of the ceRNA network in the SC. a-b, k-m: the expression plots of KIF11, hsa-miR-30d-5p, DLEU2, CCNE2, SPTBN1, hsa-miR-218-5p, and

MAGI2-AS3 in the SC. e-i, n-p: the correlations found in the SC. j-q: the regulatory networks in the SC. The arrow direction represents the predicted direction. The rhombus represents miRNA, the rectangle represents mRNA, while the ellipse represents lncRNA. Red represents up-regulated expression, while blue represents down-regulated expression.

HMGA1/hsa-miR-424-5p/PVT1	KIF11/hsa-miR-30d-5p/DLEU2	CCNE2/hsa-miR-30d-5p/DLEU2	SPTBN1/hsa-miR-218-5p/MAGI2-AS3
RIBOSOME	SPLICEOSOME	SPLICEOSOME	SPLICEOSOME
OXIDATIVE PHOSPHORYLATION	RIBOSOME	CELL CYCLE	OXIDATIVE PHOSPHORYLATION
SPLICEOSOME	CELL CYCLE	DNA REPLICATION	DNA REPLICATION
PROTEASOME	DNA REPLICATION	MISMATCH REPAIR	PARKINSONS DISEASE
PARKINSONS DISEASE	MISMATCH REPAIR	RNA DEGRADATION	RIBOSOME
PYRIMIDINE_METABOLISM	RNA DEGRADATION	PROTEASOME	BASE EXCISION REPAIR
AMINOACYL TRNA BIOSYNTHESIS	PROTEASOME	NUCLEOTIDE EXCISION REPAIR	NUCLEOTIDE EXCISION REPAIR
BASE EXCISION REPAIR	NUCLEOTIDE EXCISION REPAIR	BASAL TRANSCRIPTION FACTORS	PROTEASOME
HOMOLOGOUS RECOMBINATION	HUNTINGTONS DISEASE	PYRIMIDINE METABOLISM	RNA POLYMERASE
RNA POLYMERASE	BASAL TRANSCRIPTION FACTORS	OOCYTE MEIOSIS	CITRATE CYCLE TCA CYCLE
HUNTINGTONS_DISEASE	PYRIMIDINE METABOLISM	P53 SIGNALING PATHWAY	RNA DEGRADATION
DNA REPLICATION	OOCYTE MEIOSIS	PROTEIN EXPORT	HUNTINGTONS DISEASE
RNA DEGRADATION	P53 SIGNALING PATHWAY		PYRIMIDINE METABOLISM
PURINE METABOLISM			ALZHEIMERS DISEASE
GLYOXYLATE AND DICARBOXYLATE METABOLISM			PORPHYRIN AND CHLOROPHYLL METABOLISM
CELL CYCLE			KBASAL TRANSCRIPTION FACTORS
ALZHEIMERS DISEASE			

Table 7. The related GSEA of each regulatory network (KEGG pathway).

Discussion

The KEGG pathway analysis was enriched by the DEmiRNA and DEmRNA. Phagosome, Human T-cell leukemia virus 1 infection, and ECM-receptor interaction were significantly enriched in the AD, which was closely associated with the immune microenvironment and the maintenance of cell structure and function [24, 25]. The cell cycle and p53 signaling pathways were significantly enriched in the SC, which was closely connected to the basic biological processes and signaling pathways. According to the GO analysis, cardiac septum morphogenesis (related to BP), proteinaceous extracellular matrix (related to CC), and the growth factor binding (related to MF) coexisted in the AD and SC, suggesting that they are essential points for the formation of NSCLC.

The hub mRNAs and miRNAs in the AD and SC obtained from the microarray data were selected according to the KEGG analysis, GO analysis, interaction network, and PPI network analysis. 2 miRNA (hsa-miR-21-5p, hsa-miR-200a-5p) and 3 mRNA (COL1A1, PIK3R1, and CCND2) were closely related to the overall survival in the AD while 4 miRNA (hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, and hsa-miR-205-5p) and 4 mRNA (FRMD6, FOXP1, DDAH1and ENPP4) were related to the overall survival in the SC. 3 miRNA (hsa-miR-21-5p, hsa-miR-625-5p, hsa-miR-200a-5p) and 1 mRNA (COL1A1) with high AUC suggested a significant diagnostic value for the AD while 8 miRNA (hsa-miR-149-5p, hsa-miR-200c-3p, hsa-miR-9-5p, hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-196b-5p, hsa-miR-205-5p) and 2 mRNA (FRMD6, NAA50) presented a great diagnostic value for the SC.

Thus, hsa-miR-21-5p, hsa-miR-200a-5p, COL1A1 might be signature regulators for AD while hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-205-5p, and FRMD6 might be important for SC. Several articles have proved that miR-21-5p and miR-200a regulate the progression of AD. The miR-21-5p targets SET/TAF-Iα, WWC2, etc [26, 27] while the miR-200a correlates with MALAT1, DNMT1, GOLM1, etc. [28, 29]. Collagen 1A1 (COL1A1) is an extracellular matrix protein [30], which participates in the process of focal adhesion and may influence the metastatic ability of the cells [31]. The abnormal expression of COL1A1 has been reported in several cancers, including AD [31-33]. The miR-429 gene can bind to SNHG22 and SESN3 in the SC cells [34], and several reports have even proved that the expression of miR-205 can distinguish between squamous cell carcinoma and adenocarcinoma in lung cancer biopsies [35, 36]. To summarize, the selected differentially expressed genes between the AD and SC, especially with statistically significant ROC curve and association with overall survival, may act as significant biomarkers to subclassify NSCLC or important targets helping in designing a precise therapy for lung cancer.

Besides, miR-424-5p was filtrated in the previous research since it coexpressed with lncRNA PVT1 and was associated with NSCLC radiosensitivity [37]. Similarly, we also found that the regulatory network of HMGA1/hsa-miR-424-5p/PVT1 might play important biological functions in regulating the development of AD. Meanwhile, the regulatory networks of KIF11/hsa-miR-30d-5p/DLEU2, CCNE2/hsa-miR-30d-5p/DLEU2, and SPTBN1/hsa-miR-218-5p/MAGI2-AS3 might also be closely related to the tumor formation and progression of SC, where the values of cor > 0.3 and p < 0.05 in the SC regulatory groups indicated much more powerful connections. However, these regulatory networks were derived from one certain lung cancer (AD or SC), and we did not prove if it contributed to another cancer. Thus, these regulatory networks may not be able to distinguish between the two cancers, but these could be one of the obstacles during the treatment of lung cancer.

Conclusion

In terms of gene expression and molecular pathways in both AD and SC, there are some similarities with some differences as well, which may be significant to accurately subclassify the tumors morphologically and help in genetic alteration-guided targeted therapy. In summary, hsa-miR-21-5p, hsa-miR-200a-5p, and COL1A1 may be the potential signature biomarkers for AD while hsa-miR-141-3p, hsa-miR-429, hsa-miR-130b-3p, hsa-miR-205-5p, and FRMD6 may play an important role in the SC. The HMGA1/hsa-miR-424-5p/PVT1 may be a significant regulatory network found in the AD, whereas to regulate the SC progression, KIF11/hsa-miR-30d-5p/DLEU2, CCNE2/hsa-miR-30d-5p/DLEU2, and SPTBN1/hsa-miR-218-5p/MAGI2-AS3 may be the useful networks.

References

[1] Nawaz K, Webster RM. The non-small-cell lung cancer drug market. Nat Rev Drug Discov 2016;15:229-30.

[2] Noguchi M, Morikawa A, Kawasaki M, Matsuno Y, Yamada T, Hirohashi S, et al. Small

adenocarcinoma of the lung. Histologic characteristics and prognosis. Cancer 1995;75:2844-52.
[3] Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. Transl Lung Cancer Res 2016;5:288-300.
[4] Couraud S, Zalcman G, Milleron B, Morin F, Souquet PJ. Lung cancer in never smokers--a review. Eur J Cancer 2012;48:1299-311.
[5] Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin 2017;67:7-30.
[6] Kenfield SA, Wei EK, Stampfer MJ, Rosner BA, Colditz GA. Comparison of aspects of smoking among the four histological types of lung cancer. Tob Control 2008;17:198-204.
[7] Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer 2014;14:535-46.
[8] Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004;304:1497-500.
[9] Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448:561-6.
[10] Topalian SL, Weiner GJ, Pardoll DM. Cancer immunotherapy comes of age. J Clin Oncol 2011;29:4828-36.
[11] Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol 2014;32:926-32.
[12] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
[13] Sticht C, De La Torre C, Parveen A, Gretz N. miRWalk: An online resource for prediction of microRNA binding sites. PLoS One 2018;13:e0206239.
[14] Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 2014;42:D92-7.
[15] Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res 2015;43:D146-52.
[16] Park K, Kim KB. miRTar Hunter: a prediction system for identifying human microRNA target sites. Mol Cells 2013;35:195-201.
[17] Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. Nucleic Acids Res 2018;46:D239-D45.
[18] Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res 2014;42:D78-85.
[19] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284-7.
[20] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498-504.
[21] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011;39:D561-8.

[22] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139-40.
[23] Adler P, Kolde R, Kull M, Tkachenko A, Peterson H, Reimand J, et al. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. Genome Biol 2009;10:R139.
[24] Zhang M, Zhu K, Pu H, Wang Z, Zhao H, Zhang J, et al. An Immune-Related Signature Predicts Survival in Patients With Lung Adenocarcinoma. Front Oncol 2019;9:1314.
[25] Liang J, Li H, Han J, Jiang J, Wang J, Li Y, et al. Mex3a interacts with LAMA2 to promote lung adenocarcinoma metastasis via PI3K/AKT pathway. Cell Death Dis 2020;11:614.
[26] Zhong J, Ren X, Chen Z, Zhang H, Zhou L, Yuan J, et al. miR-21-5p promotes lung adenocarcinoma progression partially through targeting SET/TAF-Ialpha. Life Sci 2019;231:116539.
[27] Wang G, Zhou Y, Chen W, Yang Y, Ye J, Ou H, et al. miR-21-5p promotes lung adenocarcinoma cell proliferation, migration and invasion via targeting WWC2. Cancer Biomark 2020;28:549-59.
[28] Feng C, Zhao Y, Li Y, Zhang T, Ma Y, Liu Y. LncRNA MALAT1 Promotes Lung Cancer Proliferation and Gefitinib Resistance by Acting as a miR-200a Sponge. Arch Bronconeumol 2019;55:627-33.
[29] Yang L, Luo P, Song Q, Fei X. DNMT1/miR-200a/GOLM1 signaling pathway regulates lung adenocarcinoma cells proliferation. Biomed Pharmacother 2018;99:839-47.
[30] Exposito JY, Valcourt U, Cluzel C, Lethias C. The fibrillar collagen family. Int J Mol Sci 2010;11:407-26.
[31] Tian ZQ, Li ZH, Wen SW, Zhang YF, Li Y, Cheng JG, et al. Identification of Commonly Dysregulated Genes in Non-small-cell Lung Cancer by Integrated Analysis of Microarray Data and qRT-PCR Validation. Lung 2015;193:583-92.
[32] Hayashi M, Nomoto S, Hishida M, Inokawa Y, Kanda M, Okamura Y, et al. Identification of the collagen type 1 alpha 1 gene (COL1A1) as a candidate survival-related factor associated with hepatocellular carcinoma. BMC Cancer 2014;14:108.
[33] Yang Z, Liu B, Lin T, Zhang Y, Zhang L, Wang M. Multiomics analysis on DNA methylation and the expression of both messenger RNA and microRNA in lung adenocarcinoma. J Cell Physiol 2019;234:7579-86.
[34] Li ZW, Zhang TY, Yue GJ, Tian X, Wu JZ, Feng GY, et al. Small nucleolar RNA host gene 22 (SNHG22) promotes the progression of esophageal squamous cell carcinoma by miR-429/SESN3 axis. Ann Transl Med 2020;8:1007.
[35] Patnaik S, Mallick R, Kannisto E, Sharma R, Bshara W, Yendamuri S, et al. MiR-205 and MiR-375 microRNA assays to distinguish squamous cell carcinoma from adenocarcinoma in lung cancer biopsies. J Thorac Oncol 2015;10:446-53.
[36] Lebanony D, Benjamin H, Gilad S, Ezagouri M, Dov A, Ashkenazi K, et al. Diagnostic assay based on hsa-miR-205 expression distinguishes squamous from nonsquamous non-small-cell lung carcinoma. J Clin Oncol 2009;27:2030-7.
[37] Wang D, Hu Y. Long Non-coding RNA PVT1 Competitively Binds MicroRNA-424-5p to Regulate CARM1 in Radiosensitivity of Non-Small-Cell Lung Cancer. Mol Ther Nucleic Acids 2019;16:130-40.

Figures

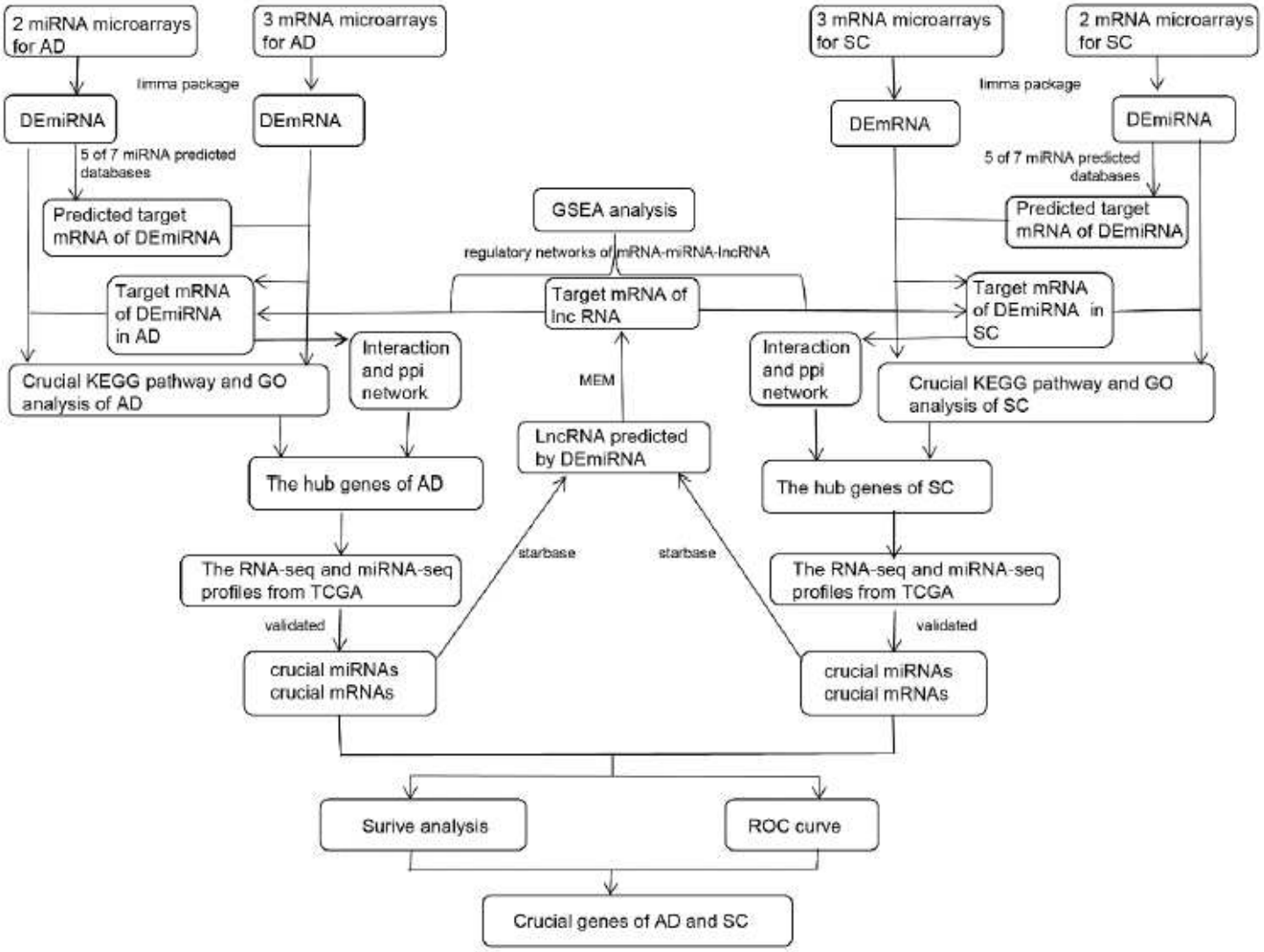


Figure 1

Workflow demonstrating the identification of core genes and pathways.

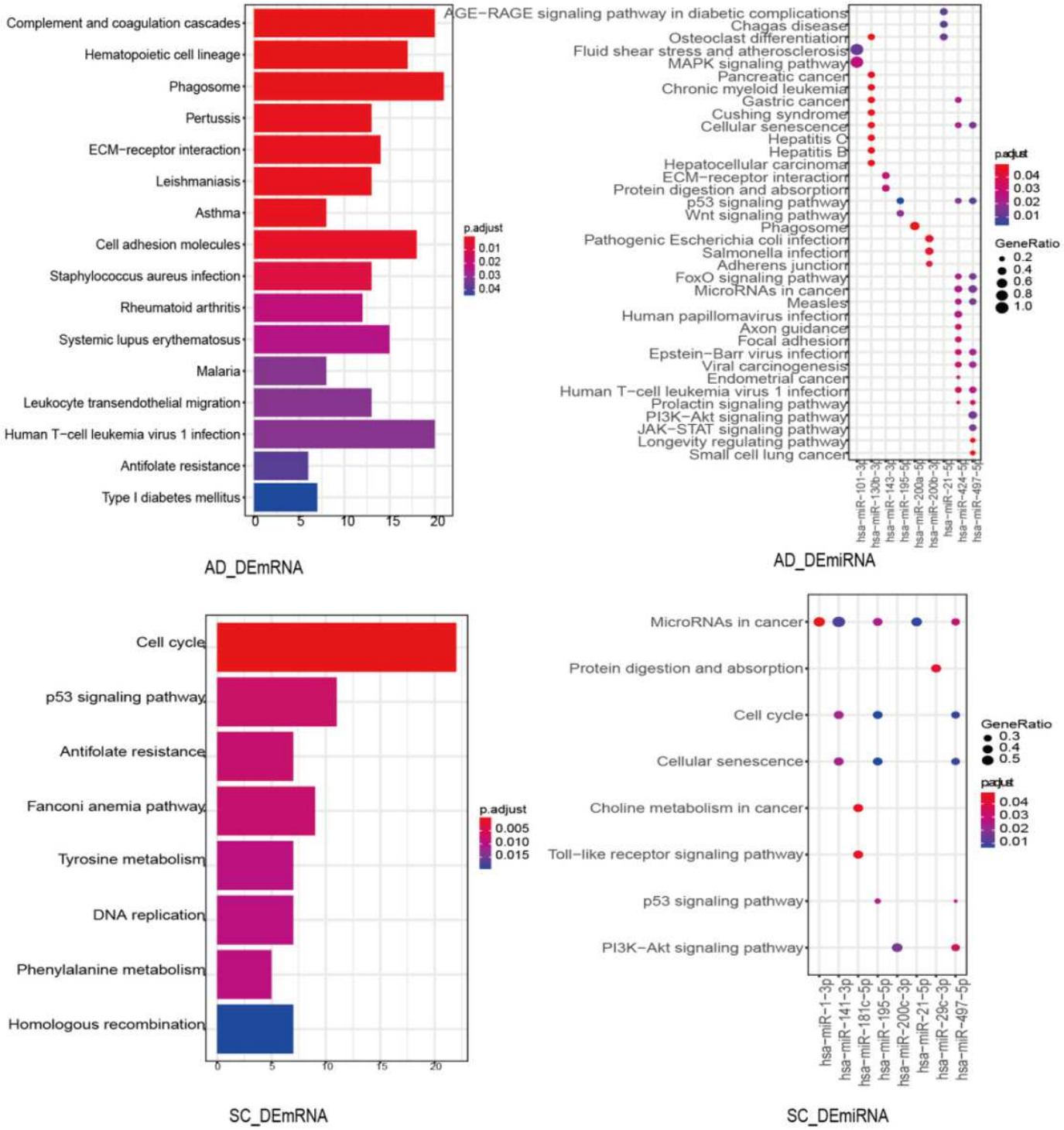


Figure 2

The KEGG pathways enriched by the DEmRNA and DEmiRNA of AD and SC

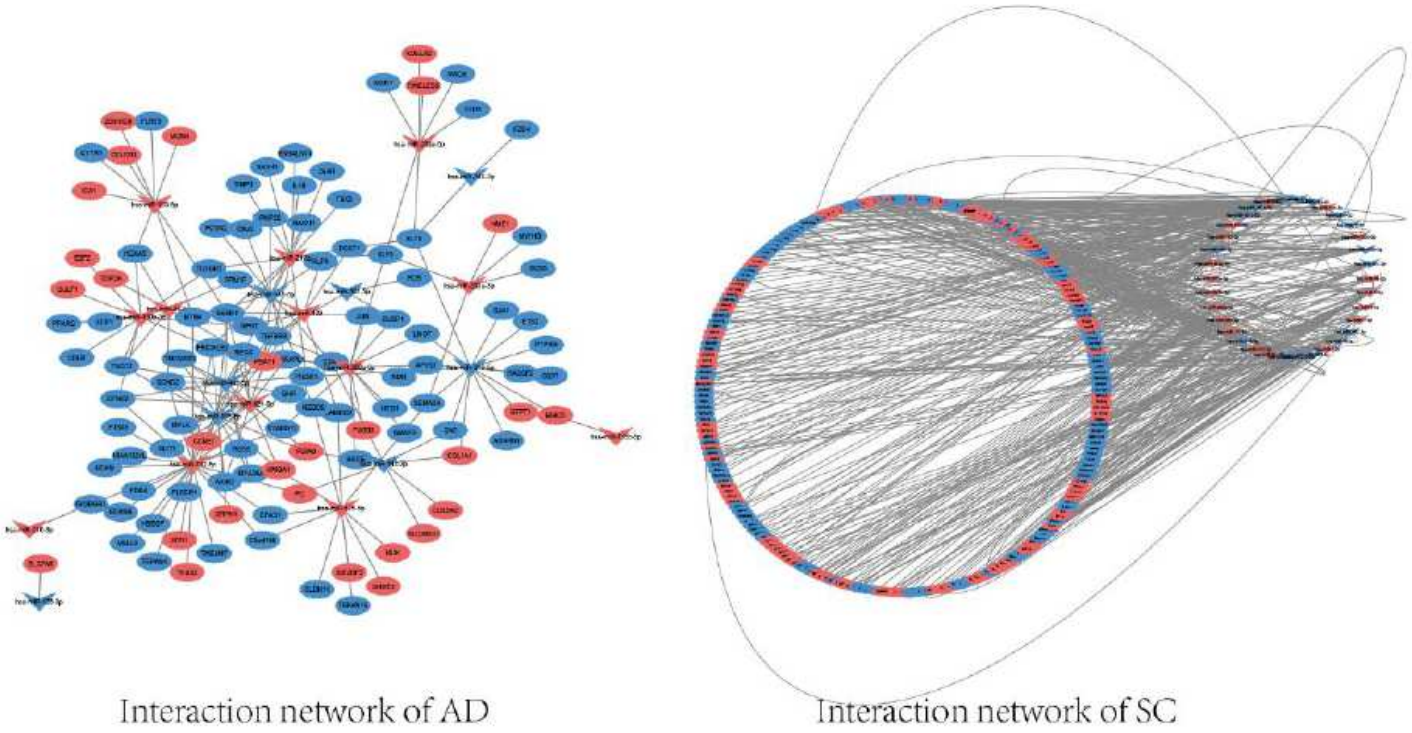


Figure 3

Interaction networks between the DEmiRNA and its target genes in the AD and Red represents up regulated expression, while blue represents down-regulated expression The arrow represents miRNA, while the ellipse represents mRNA.

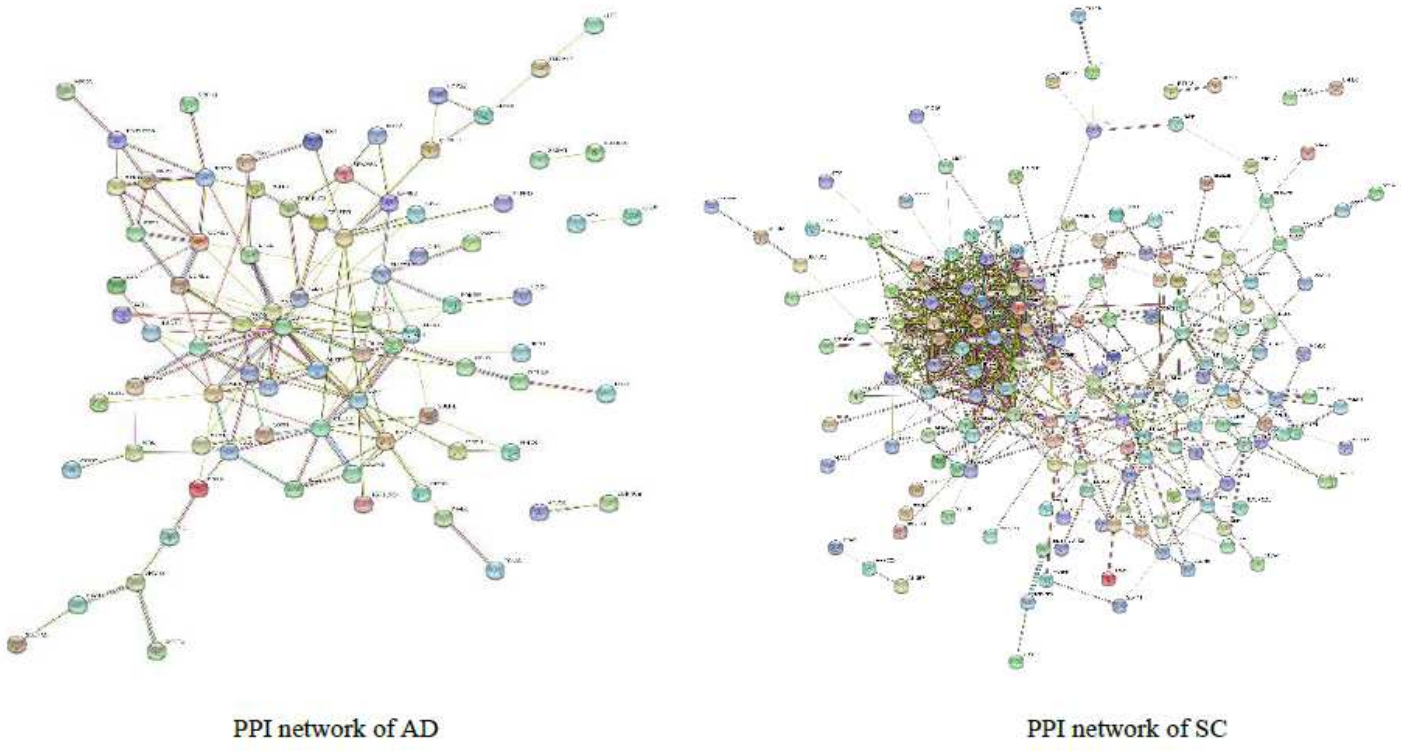


Figure 4

The PPI networks in the AD and SC.

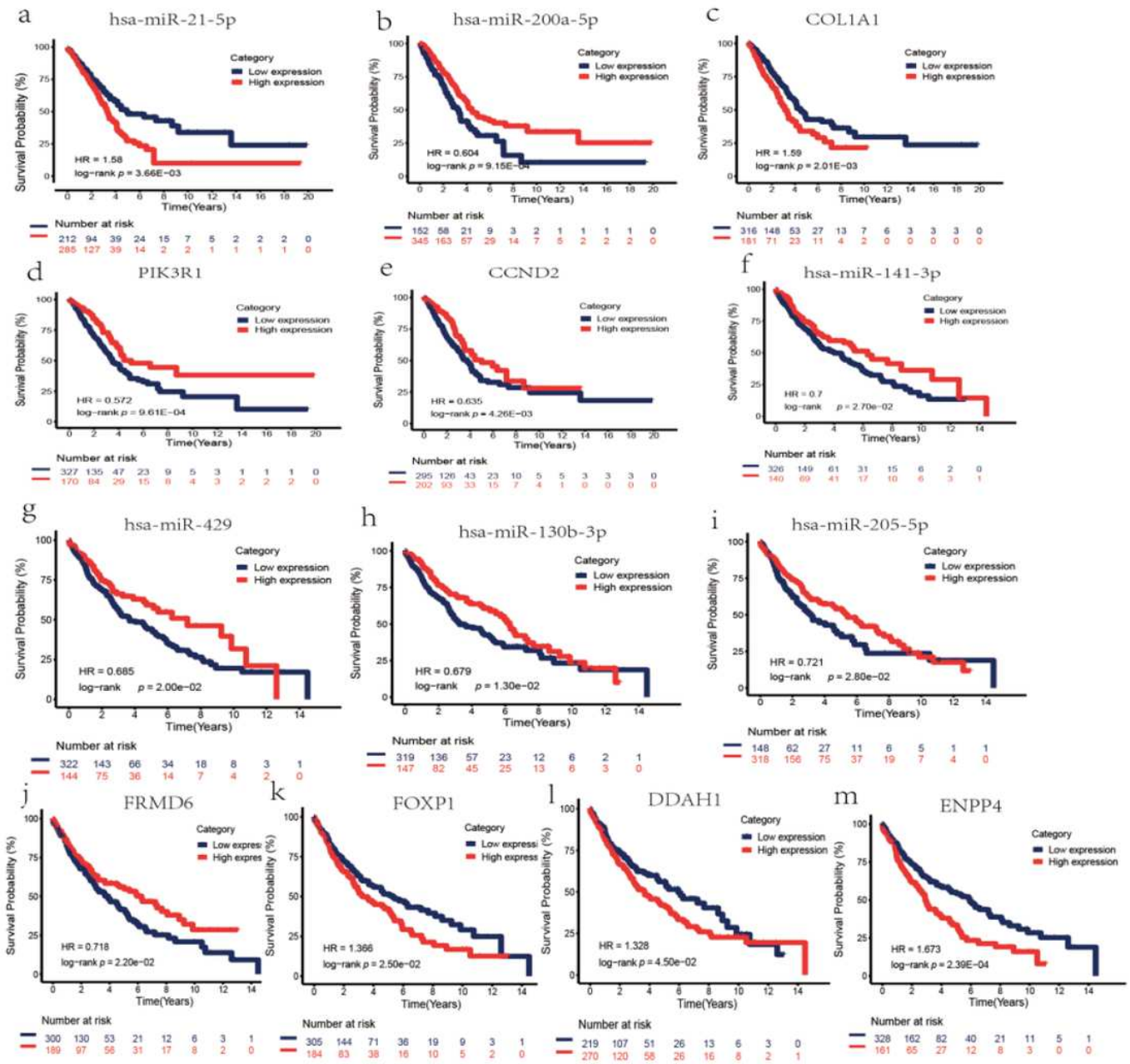


Figure 5

The overall survival analysis using the hub genes, where either the gene expression in AD > SC or AD < SC. a-e: genes in the AD; f-m: genes in the SC.

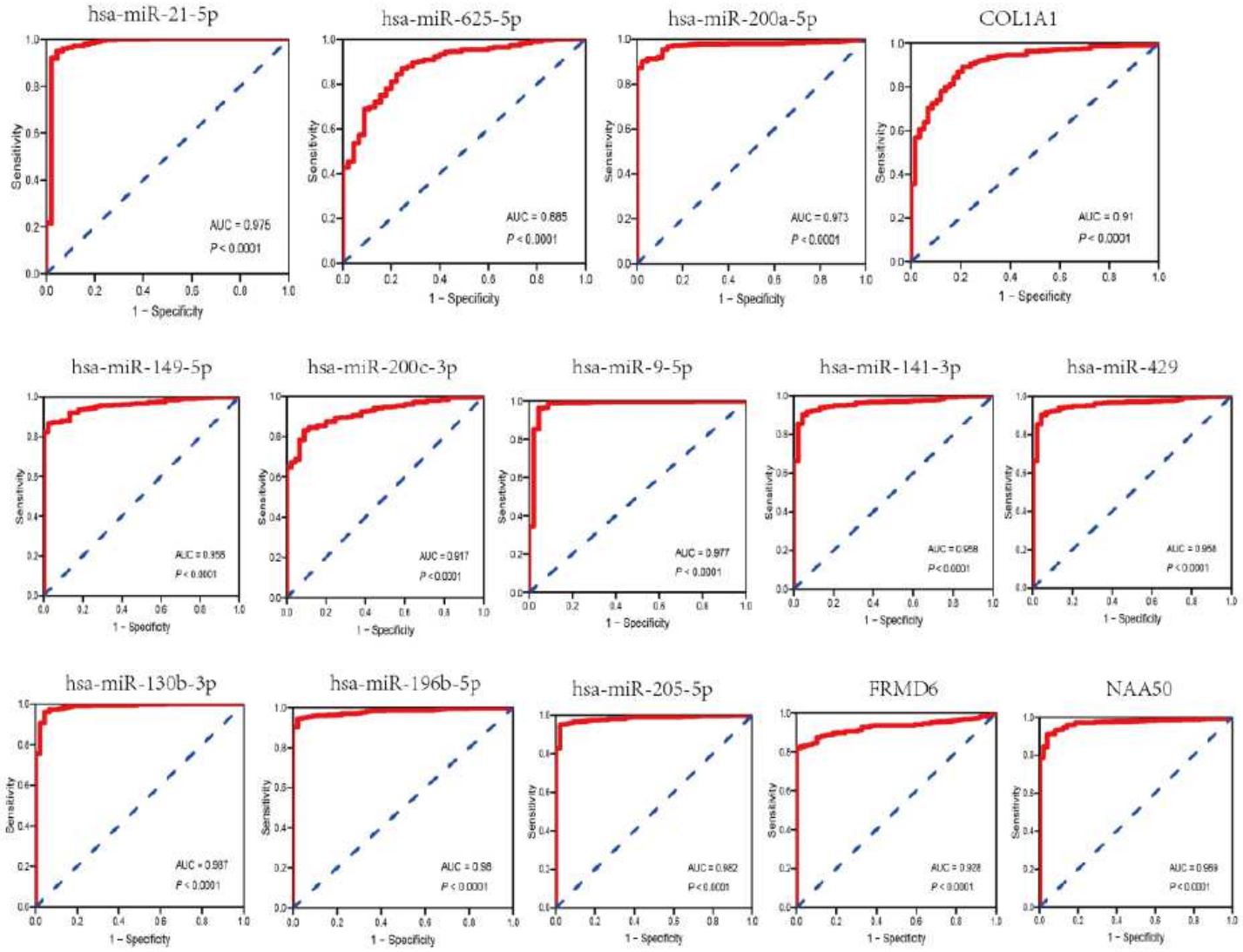


Figure 6

The ROC curves of the hub genes.

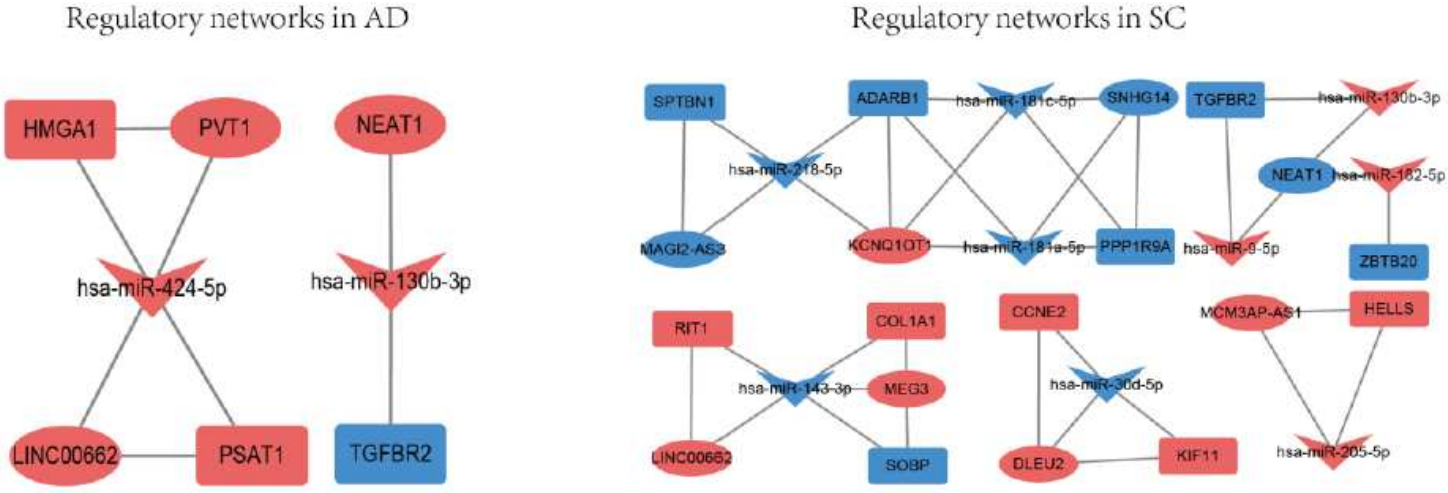


Figure 7

The probable regulatory networks of mRNA-miRNA-lncRNA. Red represents up-regulated expression, while blue represents down-regulated expression. The arrow represents miRNA, the rectangle represents mRNA, while the ellipse represents lncRNA.

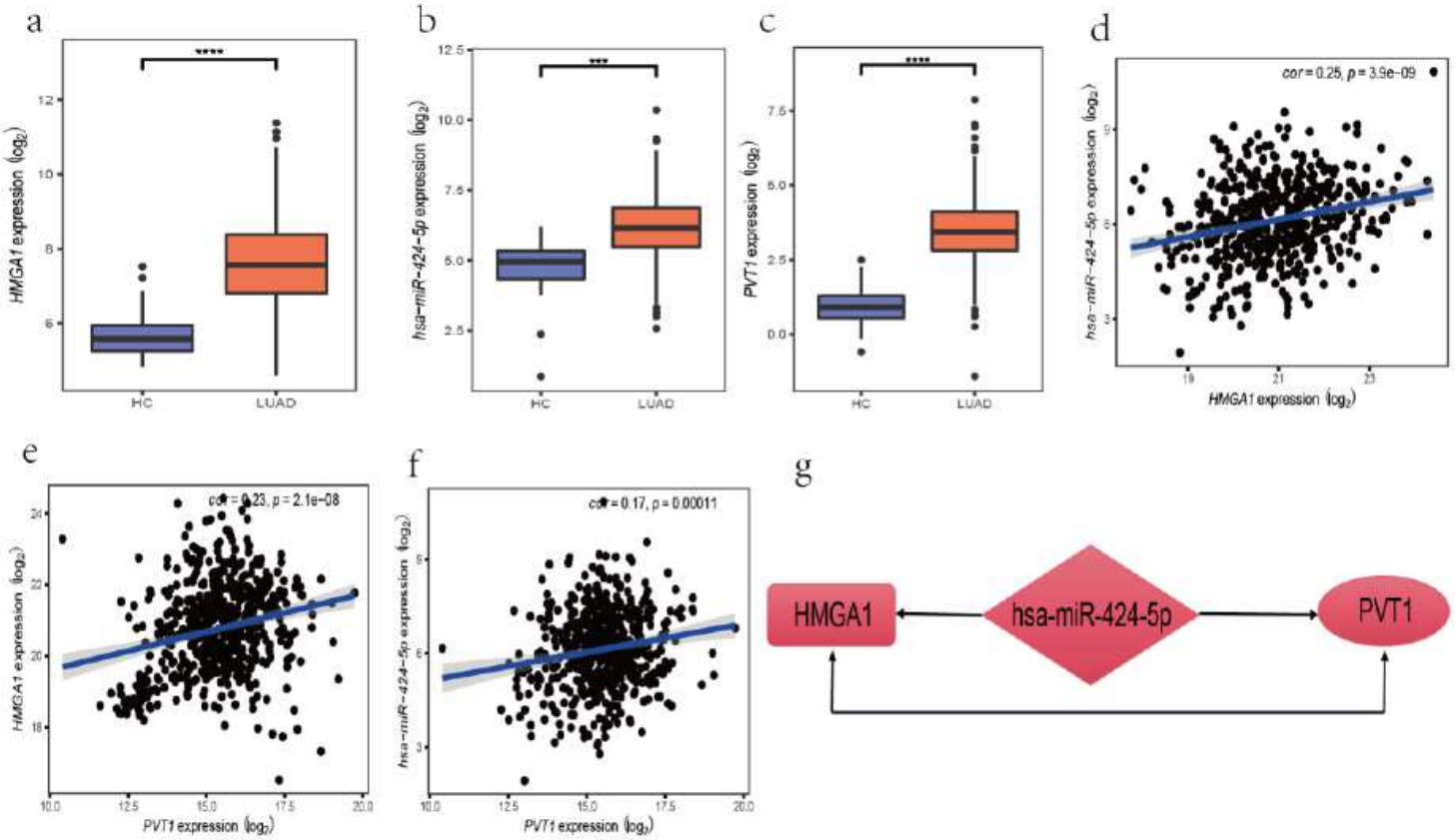


Figure 8

The expression and correlation plots of the competitive endogenous RNA (ceRNA) network in the AD. a-c: the expression plots of HMGA1, hsa-miR-424-5p, and PVT1 in AD. d-f: the correlations found in the AD. h: the regulatory network in the AD. The arrow direction represents the predicted direction. The rhombus repesents the predicted direction. The rhombus represents miRNA, the rectangle represents miRNA, the rectangle represents mRNA, while the ellipse represents lncRNA.mRNA, while the ellipse represents lncRNA. Red represents up-regulated expression, while blue represents down-regulated expression.

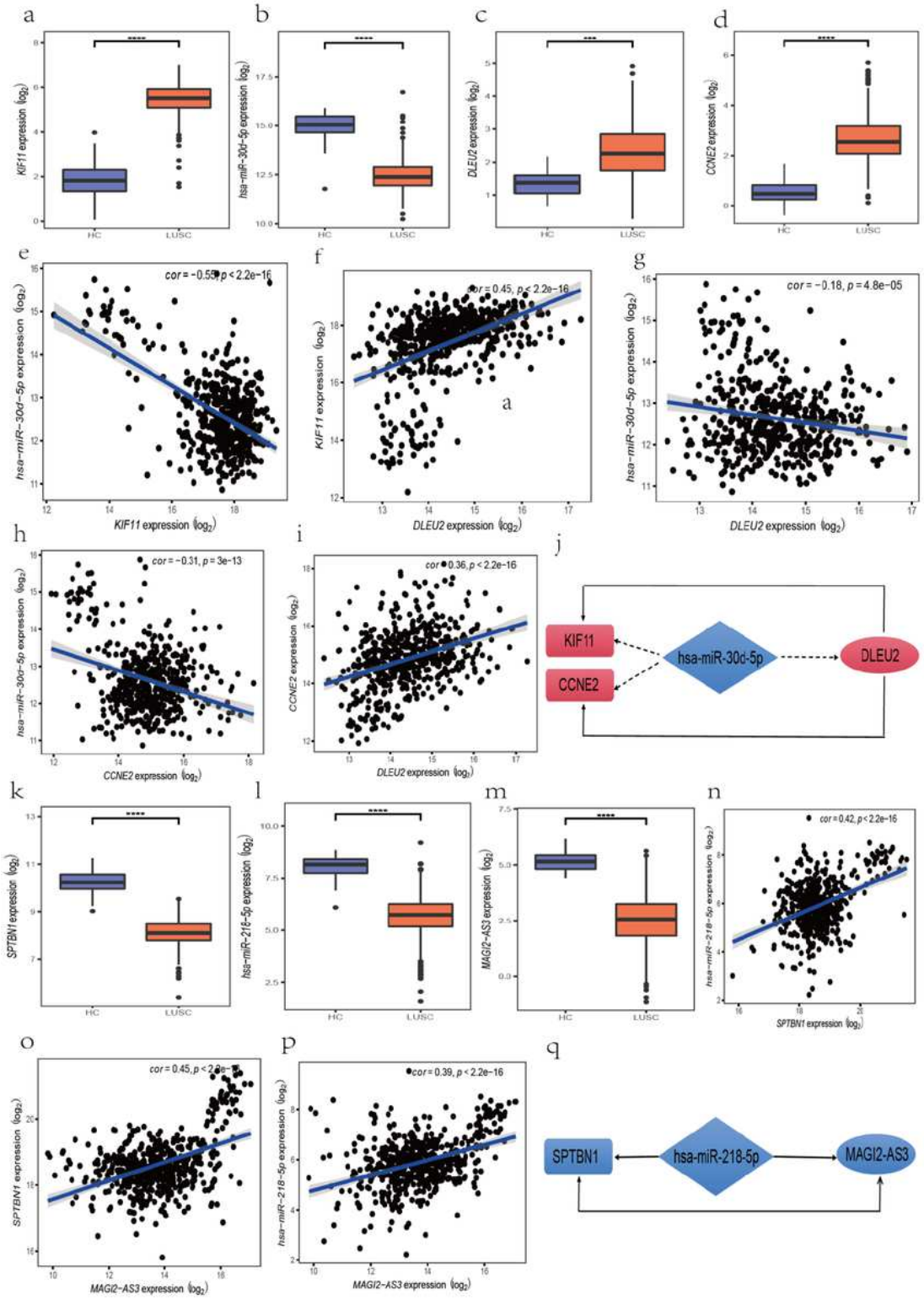


Figure 9

The expression and correlation plots of the ceRNA network in the SC. a-b, k-m: the expression plots of KIF11, hsa-miR-30d-5p, DLEU2, CCNE2, SPTBN1, hsa-miR-218-5p, and MAGI2-AS3 in the SC. e-i, n-p: the correlations found in the SC. j-q: the regulatory networks in the SC. The arrow direction represents the predicted direction The rhombus represents miRNA, the rectangle represents mRNA, while the ellipse

represents lncRNA. Red represents up-regulated expression, while blue represents down regulated expression.

Supplementary Files

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

- Supplementarymaterials.pdf