Received: 12 January 2020 | Revised: 10 June 2020 | Accepted: 21 August 2020

DOI: 10.1002/mgg3.1512

ORIGINAL ARTICLE

Molecular Genetics & Genomic Medicine Open Access WILEY

Screening lncRNAs with diagnostic and prognostic value for human stomach adenocarcinoma based on machine learning and mRNA-lncRNA co-expression network analysis

Qun Li[1] | Xiaofeng Liu[1] | Jia Gu[2] | Jinming Zhu[3] | Zhi Wei[1] | Hua Huang[1]

[1]Department of Gastroenterology, The 960th Hospital of the PLA Joint Logistics Support Force, Jinan, China

[2]Department of Pathology, The 960th Hospital of the PLA Joint Logistics Support Force, Jinan, China

[3]Department of General surgery, The 960th Hospital of the PLA Joint Logistics Support Force, Jinan, China

Correspondence
Xiaofeng Liu, Department of Gastroenterology, The 960th Hospital of the PLA Joint Logistics Support Force, No. 25, Shi Fan, Tianqiao District, Jinan 250031, China.
Email: liuxiaofeng0531@126.com

Abstract

Background: Stomach adenocarcinoma (STAD), is one of the most lethal malignancies around the world. The aim of this study was to find the long noncoding RNAs (lncRNAs) acting as diagnostic and prognostic biomarker of STAD.

Methods: Base on TCGA dataset, the differentially expressed mRNAs (DEmRNAs) and lncRNAs (DElncRNAs) were identified between STAD and normal tissue. The machine learning and survival analysis were performed to evaluate the potential diagnostic and prognostic value of lncRNAs for STAD. We also build the co-expression network and functional annotation. The expression of selected candidate mRNAs and lncRNAs were validated by Quantitative real-time polymerase chain reaction (qRT-PCR) and GSE27342 dataset. GSE27342 dataset were also to perform gene set enrichment analysis.

Results: A total of 814 DEmRNAs and 106 DElncRNAs between STAD and normal tissue were obtained. FOXD2-AS1, LINC01235, and RP11-598F7.5 were defined as optimal diagnostic lncRNA biomarkers for STAD. The area under curve (AUC) of the decision tree model, random forests model, and support vector machine (SVM) model were 0.797, 0.981, and 0.983, and the specificity and sensitivity of the three model were 75.0% and 97.1%, 96.9% and 96%, and 96.9% and 97.1%, respectively. Among them, LINC01235 was not only an optimal diagnostic lncRNA biomarkers, but also related to survival time. The expression of three DEmRNAs (*ESM1*, *WNT2*, and *COL10A1*) and three optimal diagnostic lncRNAs biomarkers (FOXD2-AS1, RP11-598F7.5, and LINC01235) in qRT-PCR validation was were consistent with our integrated analysis. Except for FOXD2-AS1, *ESM1*, *WNT2*, *COL10A*1, and LINC01235 were upregulated in STAD, which was consistent with our integration results. Gene set enrichment analysis results indicated that DNA replication, Cell cycle, ECM-receptor interaction, and P53 signaling pathway were four significantly enriched pathways in STAD.

Conclusion: Our study identified three DElncRNAs as potential diagnostic biomarkers of STAD. Among them, LINC01235 also was a prognostic lncRNA biomarkers.

KEYWORDS

biomarker, long noncoding RNAs, machine learning, stomach adenocarcinoma

1 | INTRODUCTION

Stomach adenocarcinoma (STAD), the predominant subtype in stomach cancer, is one of the major malignancies in the world (Gu et al., 2017). At present, the main treatment of STAD has been gastrectomy accompanied by chemotherapy and radiation therapy. Due to the no symptoms or no specific symptoms of the disease in its early stages, 80% of patients are diagnosed at advanced stages (Cancer Genome Atlas Research Network, 2014). In spite of advancements have been made with treatment, the survival of STAD patient remains low (Cervantes et al., 2013; Siegel et al., 2014). Hence, searching for new diagnosis and prognosis biomarkers of STAD are urgent issues.

With the advances of microarray technology, bioinformatics have become most usually used tool to obtain potential biomarkers in multiple diseases (Wang et al., 2015, 2019; Yang & Li, 2019). Mounting evidence demonstrates that long noncoding RNA (lncRNA) is closely relevant to the biological processes in cancers, such as tumor occurrence, development, and metastasis (Dey et al., 2014; Gu et al., 2017). In recent years, many lncRNAs has been identified as novel candidate biomarkers for diagnostic and prognostic of various cancer (Pan et al., 2019; Wei et al., 2019; Zeng et al., 2017). However, to our knowledge, there are few study on lncRNA biomarkers in STAD is rare. Bioinformatic analysis of the Cancer Genome Atlas (TCGA) datasets has been proven to be a novel tool in seeking diagnostic and prognostic markers for a variety of malignancies (Ding et al., 2017; Tsai et al., 2016). Machine learning is considered to be one of the most accurate prediction methods, with the ability to determine the importance of variables and to model complex interactions between independent variables (Cutler et al., 2007).

In this study, aiming to identify the diagnostic and prognostic lncRNAs biomarkers in STAD patients, we applied the bioinformatics analysis according to the lncRNA and mRNA expression profiles derived from TCGA dataset. We performed the machine learning and survival analysis to evaluate the potential diagnostic and prognostic value of lncRNAs for STAD. The DElncRNA-DEmRNA co-expression network was structured by Pearson correlation coefficient. The functions of the DEmRNAs co-expressed with the identified optimal diagnostic lncRNAs in STAD was analyzed by functional annotation. The expression levels of *ESM1* (MIM#: 601521), *WNT2* (MIM#: 147870), *COL10A1* (MIM#: 120110), FOXD2-AS1, RP11-598F7.5, and LINC01235 were verified by qRT-PCR. To our knowledge, this is first time to seek diagnostic and prognostic lncRNAs biomarkers in STAD by using machine learning.

2 | MATERIALS AND METHODS

2.1 | Integrated profiles in TCGA

The lncRNA expression profiles (Level 3-IlluminaHiSeq-lncRNASeq data) and mRNA expression profiles (Level 3-IlluminaHiSeq-mRNASeq data) and correlated clinical information were download from the Cancer Genome Atlas (TCGA) (http://tcga-data.nci.nih.gov/) through Genomic Data Commons tool. The present study included only patients who were histologically diagnosed as STAD. Finally, 375 STAD tissues and 32 normal adjacent samples from patients with STAD were included in this study.

2.2 | Identification of DEmRNAs and DElncRNAs

The RNA-Seq expression datasets were downloaded, and then, transformed from Fragments Per Kilobase Million (FPKM) data into Transcripts Per Kilobase Million (TPM) data. TPM has been considered to be more comparable than FPKM and reads per kilobase of transcript per million mapped reads (RPKM) (Li et al., 2010). Log2 of its TPM value was used as the measure of mRNAs and lncRNAs expression level here. The DElncRNAs and DEmRNAs in STAD compared to adjacent normal tissues were calculated using the R package DESeq2. Benjamini and Hochberg multiple testing method was applied to acquire the false discovery rate (FDR). FDR <.05 and |Log2fold change|>2 were used to define DElncRNAs and DEmRNAs. Hierarchical clustering analysis of DElncRNAs and DEmRNAs were further performed by using R package v3.3.3 (https://www.r-project.org/).

2.3 | Identification of the optimal diagnostic lncRNAs biomarkers for STAD

LASSO algorithm was conducted by the glmnet package (https://cran.r-project.org/web/packages/glmnet/) to decrease dimensions of the data. We also performed Elastic net to decrease dimensions of the data. We performed single 10-fold cross-validation cycles with the coordinate descent algorithm for each fold and found regularization parameters that result in the smallest average mean squared errors across all folds. The optimal DElncRNAs were selected in STAD and normal tissue.

To further identify the optimal diagnostic lncRNA biomarkers for STAD, we performed feature selection procedures as follows. (1) The importance value of each lncRNA was ranked according to mean decrease in accuracy from large to small by random forest algorithm. (2) The optimum number of features was found by adding a DElncRNA at a time in the top-down forward-wrapper packaging method. (3) By using support vector machine (SVM) at each increment and the optimal diagnostic lncRNA biomarkers were identified for STAD.

The "random Forests" packet (https://cran.r-project.org/web/packages/randomForest/) was used to establish the random forest model. The "rpart" packet (https://cran.r-project.org/web/packages/rpart/) was used to build the decision tree model. The e1071 package (https://cran.r-project.org/web/packages/e1071/index.html) in R was used to establish the SVM model. Diagnostic ability of these three models, and each lncRNA biomarker was evaluated by acquiring the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, and specificity. By using pROC package in R language, we performed the ROC analyses to assess the diagnostic value of lncRNA biomarker. The AUC under binomial exact confidence interval was calculated and ROC curve was produced.

2.4 | Survival analysis of optimal diagnostic lncRNAs biomarkers for STAD

To determine the potential association between the identified DElncRNAs and survival in STAD patients, survival analysis (https://cran.r-project.org/web/packages/survival/index.html) in R was performed.

2.5 | DEmRNAs co-expressed with the identified optimal diagnostic lncRNAs

The correlation between the optimal diagnostic lncRNAs and DEmRNAs were analyzed by the Pearson correlation coefficient. The threshold for DElncRNA-DEmRNA co-expression pairs was $p < 0.05$ and $R > 0.5$. We used the Cytoscape software (http://www.cytoscape.org/) to build the DElncRNA-DEmRNA co-expression network.

2.6 | Functional annotation of DEmRNAs co-expressed with the identified optimal diagnostic lncRNAs

Gene Ontology (GO) classification and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were performed using Metascape (http://metascape.org/gp/index.html). $p < 0.05$ was considered to indicate a statistically significant difference.

TABLE 1 Primer sequences used for qRT-PCR

Name	Sequence (5′ to 3′)
ACTB-F	CATGTACGTTGCTATCCAGGC
ACTB-R	CTCCTTAATGTCACGCACGAT
ESM1-F	CAGTGAGTGCAAAAGCAGCC
ESM1-R	TCCTCCCCATTAGAAGGCTGA
WNT2-F	TCTCGGTGGAATCTGGCTCTGG
WNT2-R	TGGCTAATGGCACGCATCACATC
COL10A1-F	CAGGAAAACCAGGCTACGGA
COL10A1-R	CCAGCTGGTCCAACZTCTCC
FOXD2-AS1-F	TGCATCCTGTGTCCTGTGTC
FOXD2-AS1-R	CCACTAGGGTCTCGCTGTTG
RP11-598F7.5-F	GCATGTCTGTCTCAAGCTGC
RP11-598F7.5-R	TGCAGAAGTTCGTGGAGGAC
LINC01235-F	CGAGACCAGCCTGACCAACATG
LINC01235-R	CTCCTGCCTTAGCCTCCTGAGTAG

2.7 | Confirmation by qRT-PCR

Base on the results of TCGA integration analysis, three DEmRNAs (*ESM1*, *WNT2*, and COL10A1) and three optimal diagnostic lncRNAs biomarkers (FOXD2-AS1, RP11-598F7.5, and LINC01235) were screened as candidate mRNAs and lncRNAs. Twelve tissues samples of STAD patients (n = 6) and normal adjacent (n = 6) were obtained. This study was approved by the Institutional Ethics Committees of the 960th Hospital of the PLA Joint Logistics Support Force and complied with the Declaration of Helsinki.

All the participants had signed a written informed. Total RNA was extracted from samples using a RNA simple total RNA kit (Tiangen, China). RNA was reverse-transcribed using a Fast Quant RT Kit (Tiangen, China) according to the manufacturer's instructions. Quantitative real-time PCR were conducted using the Super Real PreMix Plus SYBR Green (Tiangen, China) on ABI 7500 real-time PCR system. The $2^{-\Delta\Delta Ct}$ method was used to analyze the relative quantification of mRNA and lncRNA levels. Each sample was analyzed in triplicate. The PCR primers used are listed in Table 1. The human ACTB were used as endogenous controls for mRNA and lncRNA expression in analysis.

2.8 | Validation in the Gene Expression Omnibus (GEO) dataset

GSE27342 dataset was obtained from the GEO (https://www.ncbi.nlm.nih.gov/geo/), which consisted of 80 patients with STAD and 80 normal controls. The GEO dataset GSE27342 was performed to confirm the expression of some DEmRNAs and DElncRNAs.

2.8.1 | Gene set enrichment analysis

The samples in the GSE27342 dataset were divided into two groups including 80 patients with STAD and 80 normal controls. Gene set enrichment analysis (http://www.broadinstitute.org/gsea/index.jsp) was performed to understand the meaningful KEGG pathway in the two groups. The annotated gene sets of version 6.0 were downloaded from the Molecular Signatures Database (MSigDB; http://software.broadinstitute.org/gsea/msigdb/index.jsp). The inclusion criteria were normalized $p < 0.05$ and false discovery rate (FDR) <25%.

3 | RESULTS

3.1 | DEmRNAs and DElncRNAs in STAD

The detailed characteristics of 375 STAD tissues and 32 normal adjacent samples from patients with STAD are listed in Table S1. A total of 814 DEmRNAs (550 downregulated and 264 upregulated mRNAs) and 106 DElncRNAs (55 downregulated and 51 upregulated lncRNAs) between STAD and normal tissue were identified with FDR <0.05 and |Log2fold change|>2. All DEmRNAs and DElncRNAs between STAD and normal are displayed in Tables S2 and S3, respectively. Hierarchical clustering analysis of the top 100 DEmRNAs and all of DElncRNAs between STAD and normal tissue are demonstrated in Figures 1a,b, respectively.

3.2 | Identification of optimal diagnostic lncRNAs biomarkers for STAD

According to reduced dimensions of the data, we obtained 28 and 49 DElncRNAs between STAD and normal tissues by using LASSO algorithm and Elastic net, respectively (Tables 2 and 3). Hierarchical clustering analysis of the 28 DElncRNAs are shown in Figure 2a. The random forest analysis was used to rank the 28 DElncRNAs according to the mean decrease in accuracy (Figure 2b). Ten-fold cross-validation result suggested that the average accuracy rate of three DElncRNAs (FOXD2-AS1, LINC01235, and RP11-598F7.5) reached the higher score for the first time (Figure 2c). Base on the Elastic net, 10-fold cross-validation result also suggested that the average accuracy rate of three DElncRNAs (FOXD2-AS1, LINC01235, and RP11-598F7.5) reached the higher score for the first time (Figure S1). Hierarchical clustering analysis of the three DElncRNAs (FOXD2-AS1, LINC01235, and RP11-598F7.5) are displayed in Figure 2d. Therefore, these three DElncRNAs were determined as the optimal diagnostic lncRNA biomarkers for STAD which were used to establish the random forests, decision tree, and SVM models. Hierarchical clustering analysis of these three DElncRNAs between STAD and normal tissue are displayed in Figure 2d. Box-plot uncovered the expression levels of these three DElncRNAs between STAD and normal tissues (Figure 2e–g). The AUC of the decision tree model was 0.797 and the specificity and sensitivity of this model were 75.0% and 97.1%, respectively (Figure 3a). The AUC of the random forests model was 0.981 and the specificity and sensitivity of this model were 96.9% and 96%, respectively (Figure 3b). The AUC of the SVM model was 0.983 and the specificity and sensitivity of this model were 96.9% and 97.1% (Figure 3c). The AUC of all these three lncRNAs (FOXD2-AS1, LINC01235, and RP11-598F7.5) were also above 0.916 (Figure 3d–f). Taken together, the AUC of all these three lncRNAs and their combination were all greater than 0.79 which indicated the FOXD2-AS1, LINC01235, and RP11-598F7.5 and their

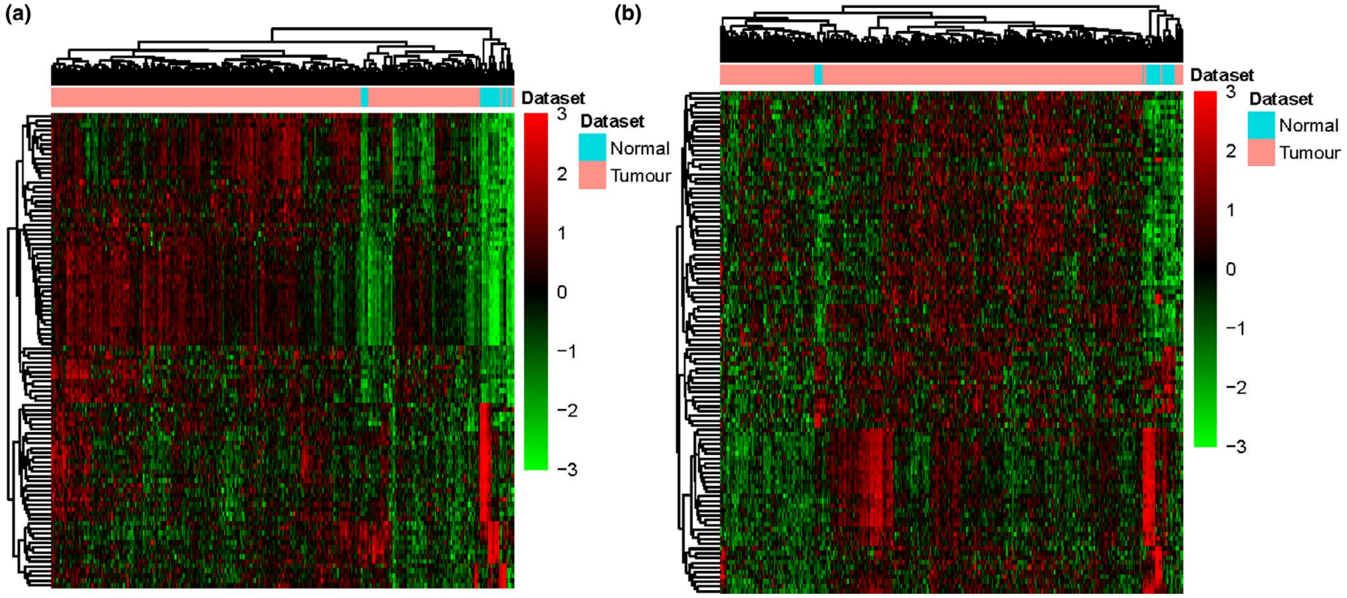

FIGURE 1 Hierarchical clustering analysis of DElncRNAs and top 100 DEmRNAs between STAD and normal tissues. (a) DEmRNAs. (b) DElncRNAs. Row and column represented DElncRNAs/DEmRNAs and tissue samples, respectively. The color scale represented the expression levels

TABLE 2 DElncRNAs between STAD and normal tissues after reduced dimensions of data by LASSO algorithm

Symbol	log2FoldChange	*p*-value	FDR	Updown
FOXD2-AS1	2.139327751	2.28E-31	2.63E-28	Up
IL12A-AS1	−3.340180888	5.18E-31	4.78E-28	Down
AC090616.2	−2.265896134	1.14E-29	8.77E-27	Down
RP11-598F7.5	2.374331116	2.37E-27	1.37E-24	Up
ADAMTS9-AS1	−2.750538461	3.62E-27	1.86E-24	Down
DLGAP1-AS2	2.043503544	2.23E-25	6.44E-23	Up
RP11-613D13.8	−2.423609575	2.96E-24	7.19E-22	Down
LINC02158	−2.245429764	3.21E-24	7.39E-22	Down
AL928768.3	−3.703197788	1.06E-22	1.96E-20	Down
LINC01235	2.53991133	5.63E-22	8.95E-20	Up
LINC01336	−2.549764194	5.00E-20	5.12E-18	Down
AC015849.16	2.256425866	9.54E-20	9.36E-18	Up
RNF144A-AS1	2.139212352	7.70E-19	6.96E-17	Up
C5orf66-AS1	−3.187219896	9.81E-19	8.53E-17	Down
CDKN2B-AS1	−2.587107347	2.91E-18	2.27E-16	Down
LINC01697	−2.747733567	9.67E-18	6.73E-16	Down
AC073283.7	2.081655386	2.20E-16	1.18E-14	Up
LINC00982	−2.306850193	1.49E-15	6.46E-14	Down
RP11-963H4.3	−2.022400723	7.67E-15	3.00E-13	Down
RP11-641D5.2	−2.040949727	1.08E-14	3.99E-13	Down
PGM5-AS1	−3.192676148	1.26E-14	4.58E-13	Down
RP11-7 K24.3	−2.041496455	9.89E-14	2.90E-12	Down
AC096579.15	−2.46634156	1.27E-13	3.67E-12	Down
AC104024.1	−2.111937683	1.75E-13	4.82E-12	Down
PART1	−2.31587256	8.27E-12	1.64E-10	Down
C20orf166-AS1	−2.491364692	2.57E-11	4.40E-10	Down
CASC9	2.276494335	2.79E-11	4.69E-10	Up
AF001548.6	−2.081376039	1.45E-08	1.29E-07	Down

combination were related to STAD and could predict the occurrence of STAD.

3.3 | Survival analysis of optimal diagnostic lncRNAs biomarkers for STAD

The association between three lncRNAs (FOXD2-AS1, LINC01235, and RP11-598F7.5) and survival in patients with STAD (Figure 3g–i). Only LINC01235 was significantly associated with the prognosis of patients with STAD.

3.4 | DEmRNAs co-expressed with the identified optimal diagnostic lncRNAs

A total of three optimal DElncRNA biomarkers for STAD were co-expressed with 87 DEmRNAs, accounting for 122 DElncRNA-DEmRNA co-expression pairs. FOXD2-AS1, LINC01235, and RP11-598F7.5 were co-expressed with 55, 50, and 17 DEmRNAs, respectively (Figure 4).

3.5 | Functional annotation of DEmRNAs co-expressed with the identified optimal diagnostic lncRNAs

Base on the functional annotation of 87 DEmRNAs co-expressed with the identified optimal diagnostic lncRNAs (Figure 5), nuclear chromosome segregation, extracellular matrix organization, extracellular matrix, and meiotic cell cycle were significantly enriched GO terms, and TNF signaling pathway, JAK-STAT signaling pathway, Transcriptional misregulation in cancer, and ECM-receptor interaction were four significantly enriched pathways.

TABLE 3 DElncRNAs between STAD and normal tissues after reduced dimensions of data by Elastic net

Symbol	log2FoldChange	*p*-value	FDR	Updown
FOXD2-AS1	2.139328	2.28E-31	2.63E-28	Up
RP11-598F7.5	2.374331	2.37E-27	1.37E-24	Up
LINC01235	2.539911	5.63E-22	8.95E-20	Up
AL928768.3	−3.7032	1.06E-22	1.96E-20	Down
GAPLINC	2.062859	1.75E-18	1.43E-16	Up
DUXAP8	3.316478	1.22E-35	2.80E-32	Up
LINC01336	−2.54976	5.00E-20	5.12E-18	Down
DLGAP1-AS2	2.043504	2.23E-25	6.44E-23	Up
LINC01697	−2.74773	9.67E-18	6.73E-16	Down
C5orf66-AS1	−3.18722	9.81E-19	8.53E-17	Down
RP1-60O19.1	2.739187	4.28E-14	1.32E-12	Up
RP11-867G23.10	−3.08974	1.47E-25	4.53E-23	Down
ADAMTS9-AS2	−2.15844	4.67E-16	2.24E-14	Down
AC015849.16	2.256426	9.54E-20	9.36E-18	Up
RNF144A-AS1	2.139212	7.70E-19	6.96E-17	Up
LINC02086	2.525563	2.62E-14	8.68E-13	Up
KRT7-AS	2.226896	3.66E-12	8.11E-11	Up
AC073283.7	2.081655	2.20E-16	1.18E-14	Up
LINC02158	−2.24543	3.21E-24	7.39E-22	Down
RP11-770 J1.3	−2.05015	1.84E-39	8.50E-36	Down
PGM5-AS1	−3.19268	1.26E-14	4.58E-13	Down
MEF2C-AS1	−2.09002	3.43E-16	1.76E-14	Down
RBMS3-AS3	−2.2516	2.02E-15	8.47E-14	Down
RP11-211G23.2	2.118864	2.16E-12	5.14E-11	Up
CTD-2540F13.2	2.141602	9.78E-18	6.73E-16	Up
AC108676.1	2.441715	4.71E-16	2.24E-14	Up
CASC9	2.276494	2.79E-11	4.69E-10	Up
PART1	−2.31587	8.27E-12	1.64E-10	Down
TRPM2-AS	2.748372	2.12E-20	2.39E-18	Up
CTD-2540F13.2	2.141602	9.78E-18	6.73E-16	Up
RP11-21A7A.2	−2.52023	9.31E-17	5.43E-15	Down
IL12A-AS1	−3.34018	5.18E-31	4.78E-28	Down
AC096579.15	−2.46634	1.27E-13	3.67E-12	Down
RP11-641D5.2	−2.04095	1.08E-14	3.99E-13	Down
RP11-963H4.3	−2.0224	7.67E-15	3.00E-13	Down
CDKN2B-AS1	−2.58711	2.91E-18	2.27E-16	Down
RP11-613D13.8	−2.42361	2.96E-24	7.19E-22	Down
RP11-626H12.2	2.19915	3.16E-17	1.94E-15	Up
AC090616.2	−2.2659	1.14E-29	8.77E-27	Down
C20orf166-AS1	−2.49136	2.57E-11	4.40E-10	Down
UBXN10-AS1	−2.77015	6.71E-20	6.73E-18	Down
LINC00671	−2.66457	3.07E-20	3.29E-18	Down
LINC00982	−2.30685	1.49E-15	6.46E-14	Down
AC104024.1	−2.11194	1.75E-13	4.82E-12	Down

(Continues)

TABLE 3 (Continued)

Symbol	log2FoldChange	*p*-value	FDR	Updown
SOX21-AS1	−2.18199	2.01E-07	1.34E-06	Down
RP11-351 J23.1	−2.7255	5.24E-14	1.61E-12	Down
AF001548.6	−2.08138	1.45E-08	1.29E-07	Down
RP11-7 K24.3	−2.0415	9.89E-14	2.90E-12	Down
RP11-800A18.4	−2.53106	1.73E-12	4.18E-11	Down

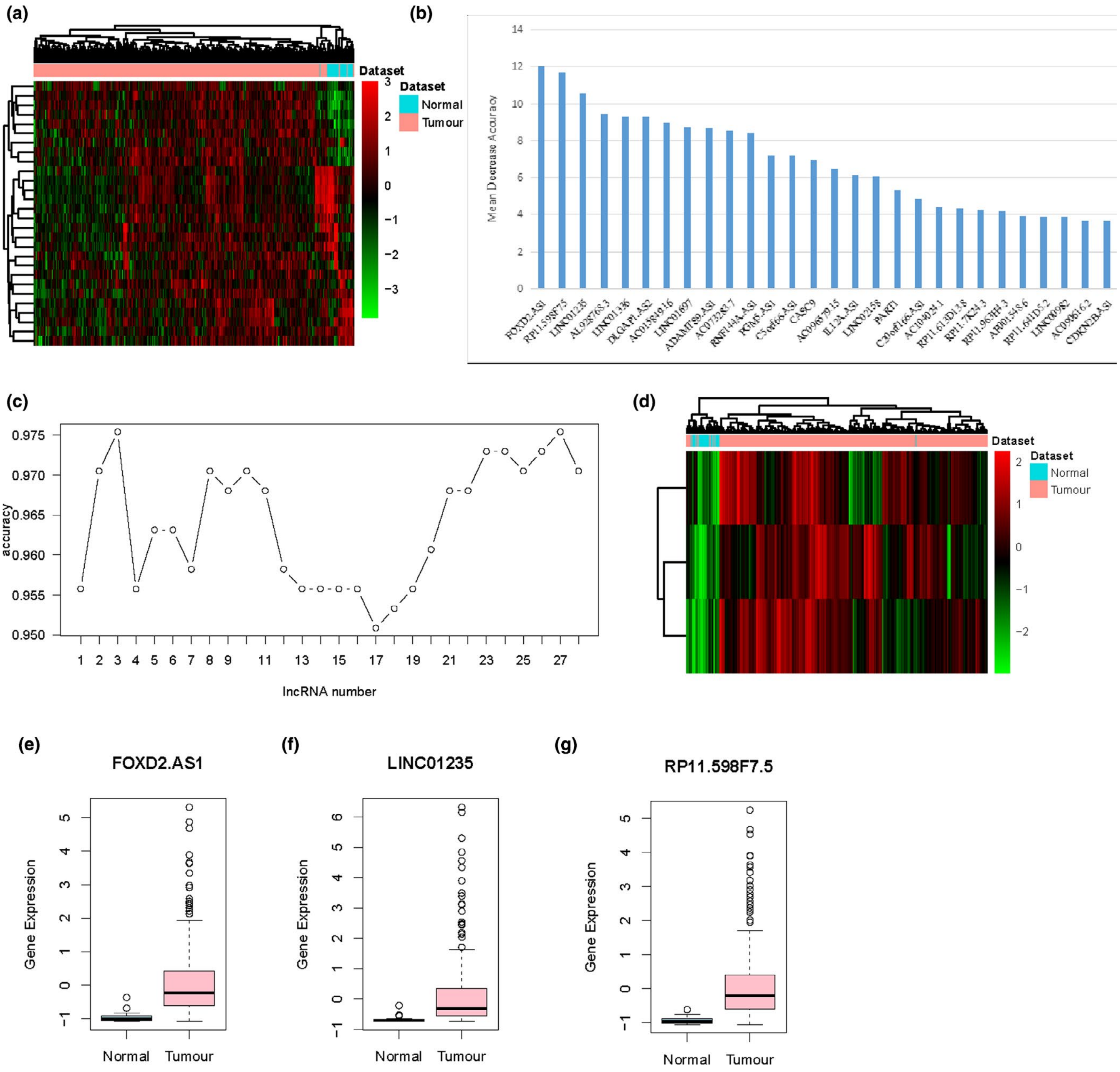

FIGURE 2 Identification of optimal lncRNA biomarkers for STAD. (a) Hierarchical clustering analysis of the 28 DElncRNAs. (b) The importance value of each DElncRNA ranked according to the mean decrease in accuracy by using the random forest analysis. (c) The variance rate of classification performance when increasing numbers of the predictive DElncRNAs. (d) Hierarchical clustering analysis of three lncRNAs biomarkers (FOXD2-AS1, LINC01235, and RP11-598F7.5). (e–g) Box-plot displayed the expression levels of three lncRNAs biomarkers between STAD and normal tissues. The *x*-axis represented normal and STAD groups. The *y*-axis represented gene expression levels

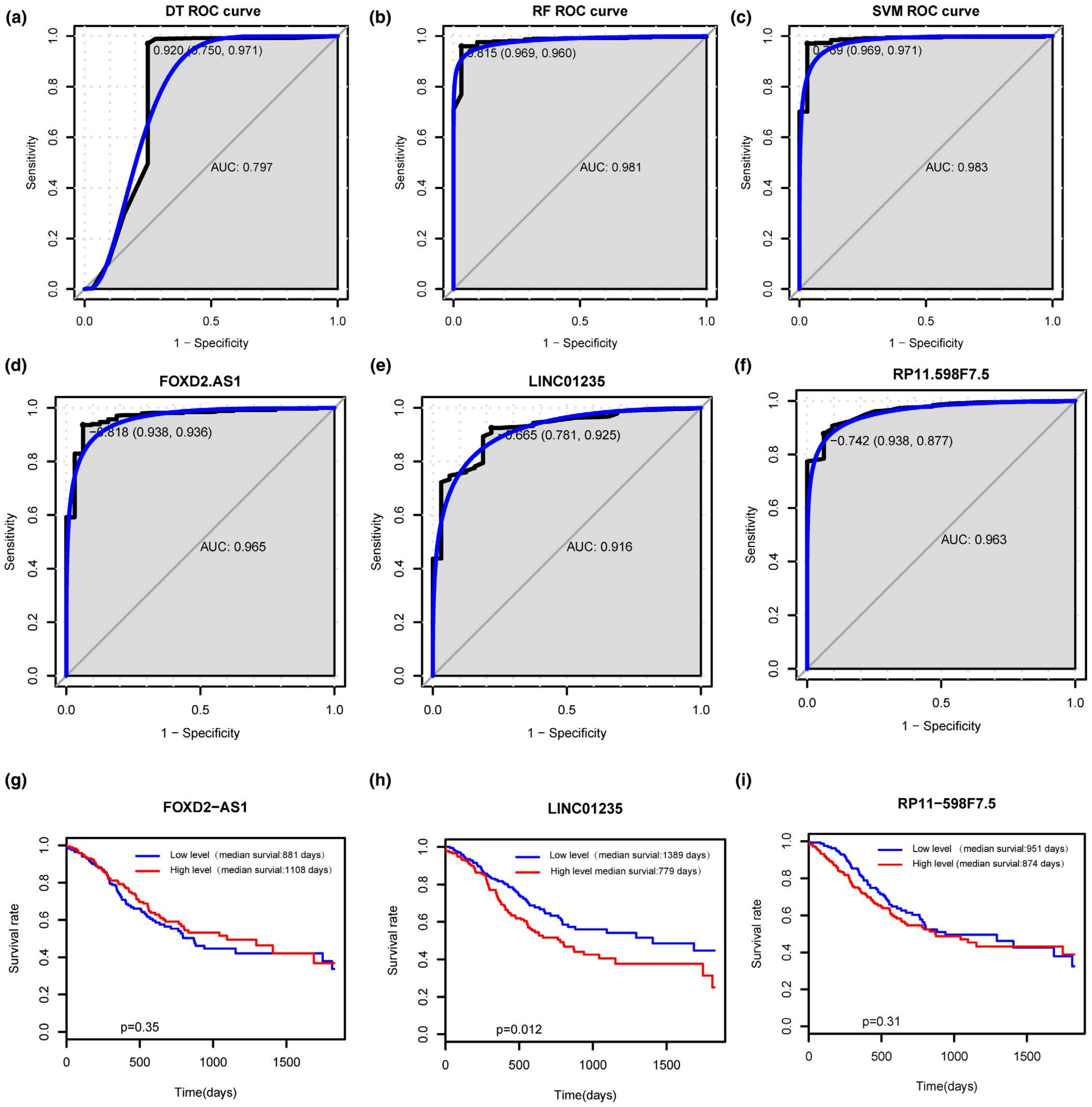

FIGURE 3 ROC analysis of three STAD-specific lncRNAs biomarkers. The ROC results of these three diagnostic lncRNAs biomarkers (FOXD2-AS1, LINC01235, and RP11-598F7.5) their combination based on decision tree model (Figure 3a), support random forest (Figure 3b) and SVM model (Figure 3c) and individual FOXD2-AS1 (Figure 3d), LINC01235 (Figure 3e), and RP11-598F7.5 (Figure 3f). The *x*-axis shows 1-specificity and y-axis shows sensitivity. Figure 3g–i Survival analysis of FOXD2-AS1, LINC01235, and RP11-598F7.5

3.6 | Confirmation by qRT-PCR

We performed the confirmation of three DEmRNAs (*ESM1*, *WNT2*, and *COL10A1*) and three optimal diagnostic lncRNAs biomarkers (FOXD2-AS1, RP11-598F7.5, and LINC01235) by qRT-PCR. Based on TCGA, *ESM1*, *WNT2*, *COL10A1*, FOXD2-AS1, RP11-598F7.5, and LINC01235 were upregulated in STAD compared to adjacent tissues. According to the qRT-PCR results, *ESM1*, *WNT2*, *COL10A1*, FOXD2-AS1, RP11-598F7.5, and LINC01235 were upregulated which was consistent with the results of TCGA (Figure 6).

3.6.1 | Validation in GEO dataset

The expression pattern of selected DEmRNAs (ESM1, WNT2, and COL10A1) and DElncRNAs (FOXD2-AS1 and LINC01235) was verified using GSE27342 dataset. The

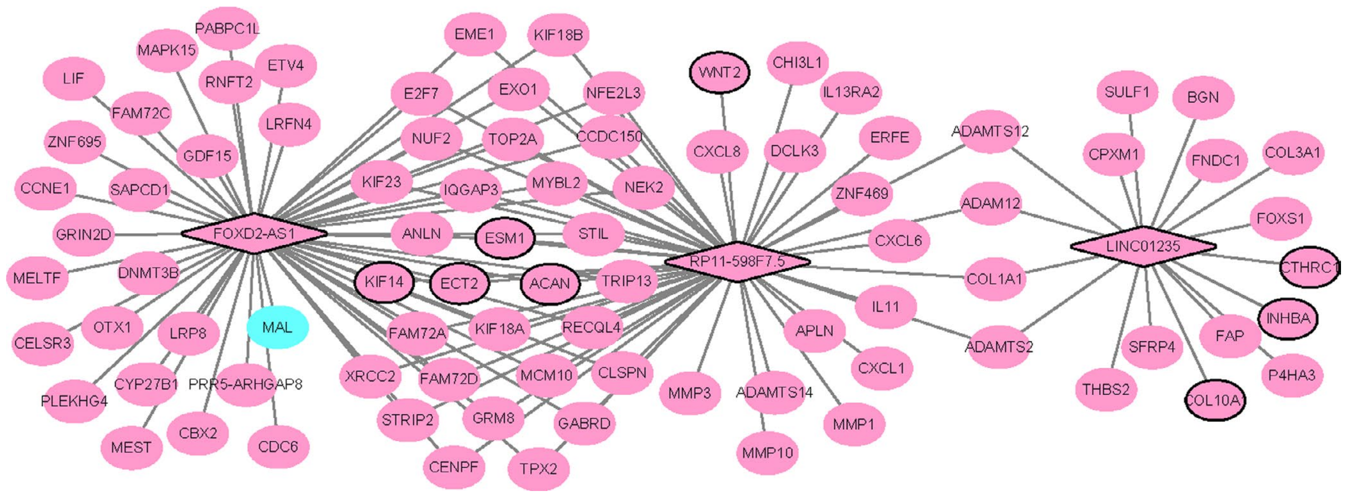

FIGURE 4 lncRNAs-mRNAs co-expression network. The ellipses and rhombuses were represented the mRNAs and lncRNAs, respectively. Red and green color represented up and downregulation, respectively. The black border indicates top 10 DEmRNAs and DElncRNAs

raw data of box-plots are displayed in Table S4. As shown in Figure 7, FOXD2-AS1 was downregulated, which was inconsistent with our integration results. ESM1, WNT2, COL10A1, and LINC01235 were upregulated in STAD, which was consistent with our integration results, suggesting that the results were convincing.

3.7 | Gene set enrichment analysis

Gene set enrichment analysis was performed in the present study using the GSE27342 dataset. The enrichment results showed that DNA replication, Cell cycle, ECM-receptor interaction, and P53 signaling pathway were four significantly enriched pathways (Figure 8).

4 | DISCUSSION

STAD is one of the leading causes of cancer death, accounting for about 10% of newly diagnosed cancer (Liu et al., 2018). Therefore, searching for novel diagnosis and prognosis biomarkers of STAD is needed. In this study, the expression profiles of lncRNAs and mRNA in STAD was obtained from TCGA dataset. We identified 814 DEmRNAs (550 downregulated and 264 upregulated mRNAs) and 106 DElncRNAs (55 downregulated and 51 upregulated lncRNAs) of STAD. A total of three optimal diagnostic lncRNA biomarkers, including FOXD2-AS1, LINC01235, and RP11-598F7.5, for STAD were identified by machine learning.

KEGG pathway enrichment analysis showed that DEmRNAs was were enriched TNF signaling pathway, JAK-STAT signaling pathway, Transcriptional misregulation in cancer, and ECM-receptor interaction, indicating that DElncRNAs and DEmRNAs might play crucial roles by participating in these pathways in STAD. Gu et al. performed the lncRNA and mRNA expression profile of three STAD tissues and three matched adjacent non-tumor tissues via RNA-sequencing, and found that the DEmRNAs co-expressed with DElncRNAs were significantly enriched in JAK-STAT signaling pathway, which was considered as a signaling pathway associated with STAD (Gu et al., 2017). Li et al. also carried out the RNA-sequencing in 15 pairs of STAD tissues and the adjacent normal tissues, and found that DEGs of most significantly enriched in ECM-receptor interaction signaling pathway (Li et al., 2019). The results indicated that our KEGG pathway enrichment analysis results were convincing. To our knowledge, except of FOXD2-AS1 and RP11-598F7.5, the present study was the first to identify the LINC01235 in STAD. At present, FOXD2-AS1 have been proven to be abnormally regulated in various human cancers. For instance, Su et al. have demonstrated that FOXD2-AS1 promotes the progression of bladder cancer by regulation AKT and E2F1 (Su et al., 2018). Chen et al. have found FOXD2-AS1 acts as a tumor promoter in nasopharyngeal carcinoma by modulating miR-363-5p/S100a1 signaling pathway (Chen et al., 2018). Ni et al. have found that FOXD2-AS1 promotes tumorigenesis and progression of glioma via miR-185-5p/HMGA2 axis (Ni et al., 2019). Xu et al. have reported that FOXD2-AS1 expression was upregulated in stomach tumor tissues, and FOXD2-AS1 promotes carcinogenesis in stomach cancer through *EZH2* and *LSD1* mediated *EphB3* downregulation (Xu et al., 2018). In current study, we found that FOXD2-AS1 was upregulated in both TCGA integration analysis and qRT-PCR validation. *LIF* co-expressed with FOXD2-AS1 was enriched in TNF signaling pathway and JAK-STAT signaling pathway. Therefore, we hypothesized that FOXD2-AS1 might play pivotal roles in STAD by regulating the TNF signaling pathway and JAK-STAT signaling pathway.

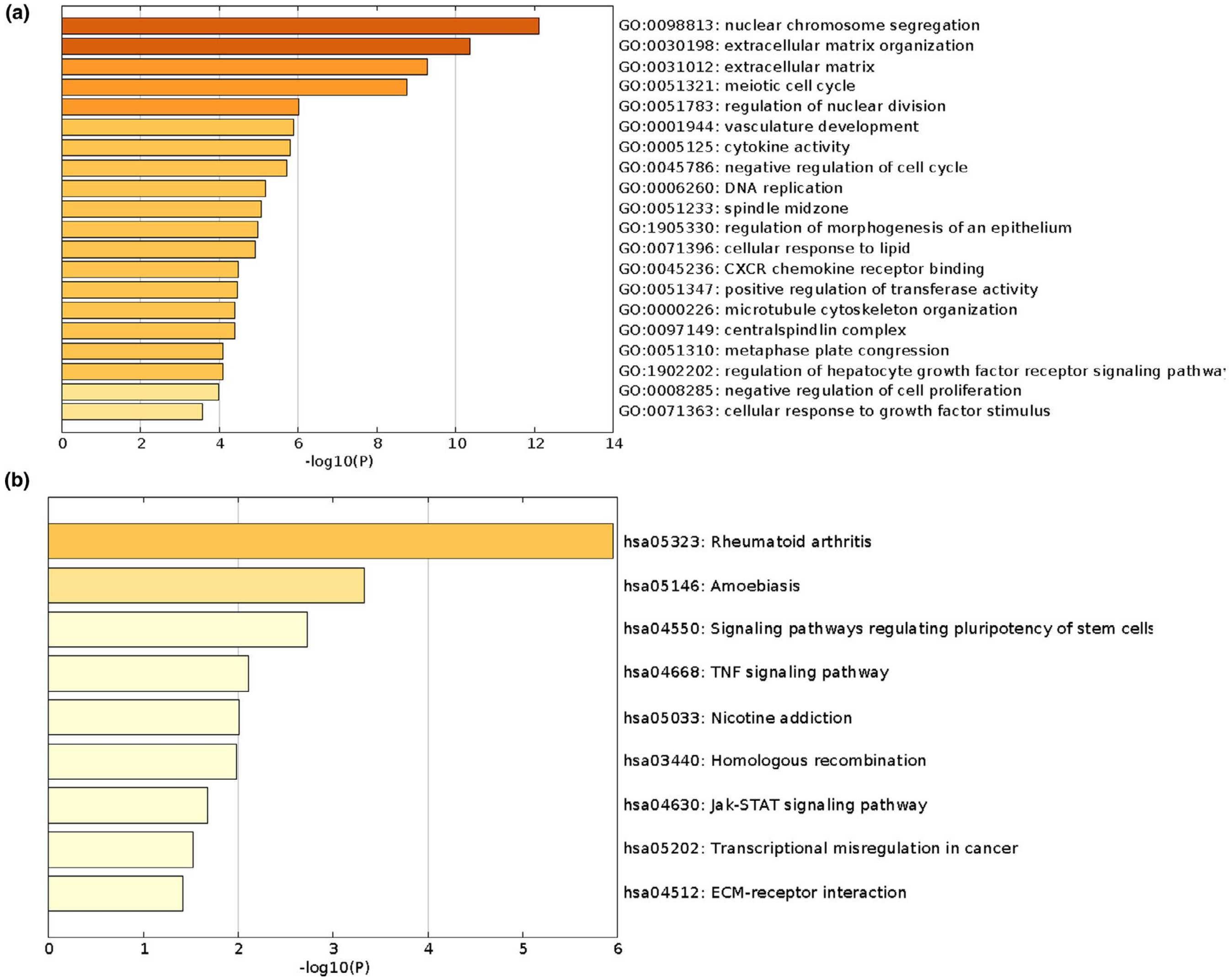

FIGURE 5 The function annotation of DEmRNAs co-expressed with the identified optimal diagnostic lncRNAs. The *x*-axis shows -log P and *y*-axis shows GO terms or KEGG pathways. (a) GO terms. (b) KEGG pathways

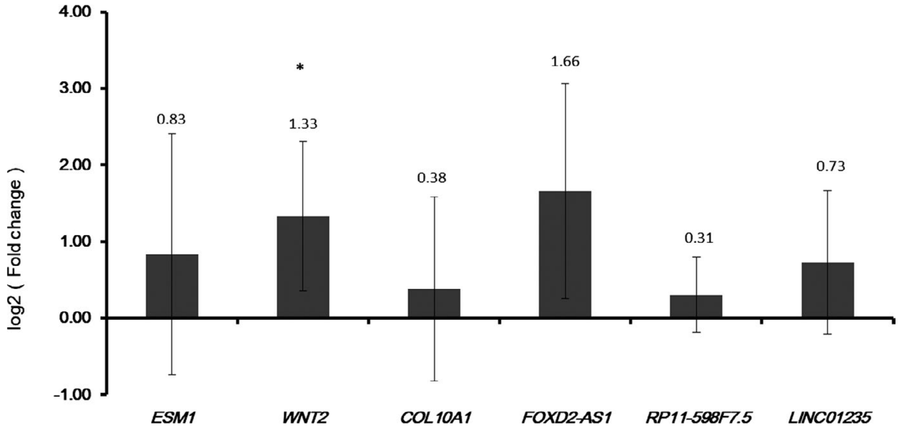

FIGURE 6 Validation optimal lncRNA biomarkers in STAD tissue by qRT-PCR. * indicated *p*-value < 0.05

The survival analysis results showed that LINC01235 was significantly associated with the prognosis of patients with STAD, which provide evidence emphasize its prognostic value for STAD. *COL10A1* (Type X collagen gene), belongs to the collagen family, has been found in various human cancers (Sole et al., 2014). Huang et al. reported that the high

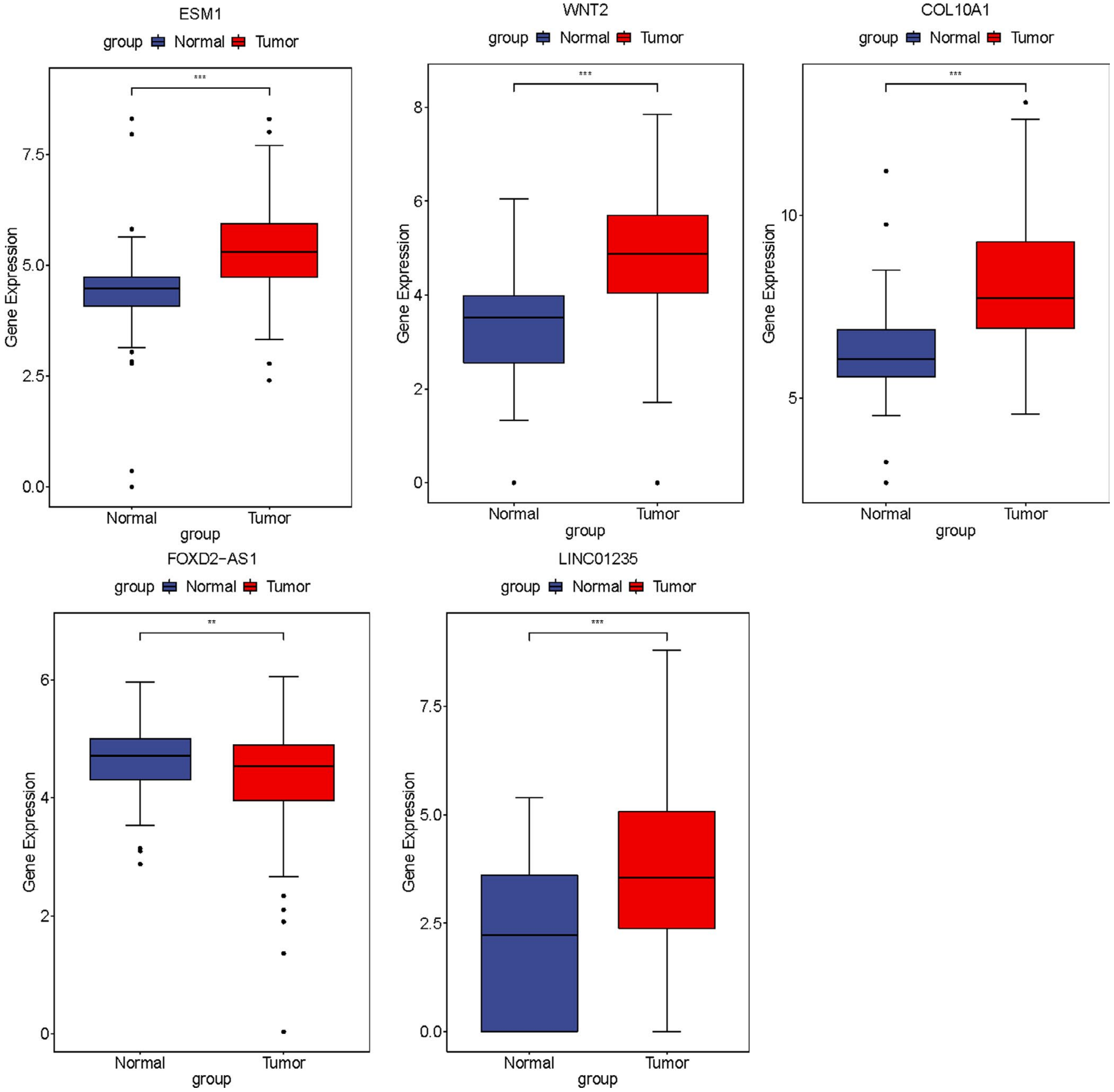

FIGURE 7 Validation in the GEO dataset. The *x*-axis shows healthy normal control (blue color) and STAD (red color) groups and *y*-axis shows a log2 transformation to the intensities. **indicated *p*-value < 0.01, ***indicated *p*-value < 0.001

expression of *COL10A1* is an independent biomarkers of prognosis and survival in colon cancer patients (Huang et al., 2018). *COL10A1* might promote STAD tumor aggressiveness by regulating of the TGF-β1-SOX9 axis (Li et al., 2018). In this study, *COL10A1* co-expressed with LINC01235 was enriched in ECM-receptor interaction pathway. Therefore, we presumed that LINC01235 might be involved in the occurrence of STAD by regulating ECM-receptor interaction pathway.

WNT2, a member of the WNT protein family, is frequently overexpressed in and colorectal cancer and gastric cancer (Katoh, 2001). Zhang et al. found that *WNT2* is upregulated in gastric cancer, and *WNT2* contributes to promoting the gastric cancer cells migration and invasion abilities (Zhang et al., 2018). In this study, we found that *WNT2* was upregulated in both TCGA integration analysis and qRT-PCR validation. The DElncRNA-DEmRNA co-expression network results showed that *WNT2* co-expressed with RP11-598F7.5. Hence, we hypothesized that RP11-598F7.5 might play important roles in STAD by regulating *WNT2*.

ESM1 (endothelial cell-specific molecule-1) was correlated to tumorigenesis and tumor progression and was

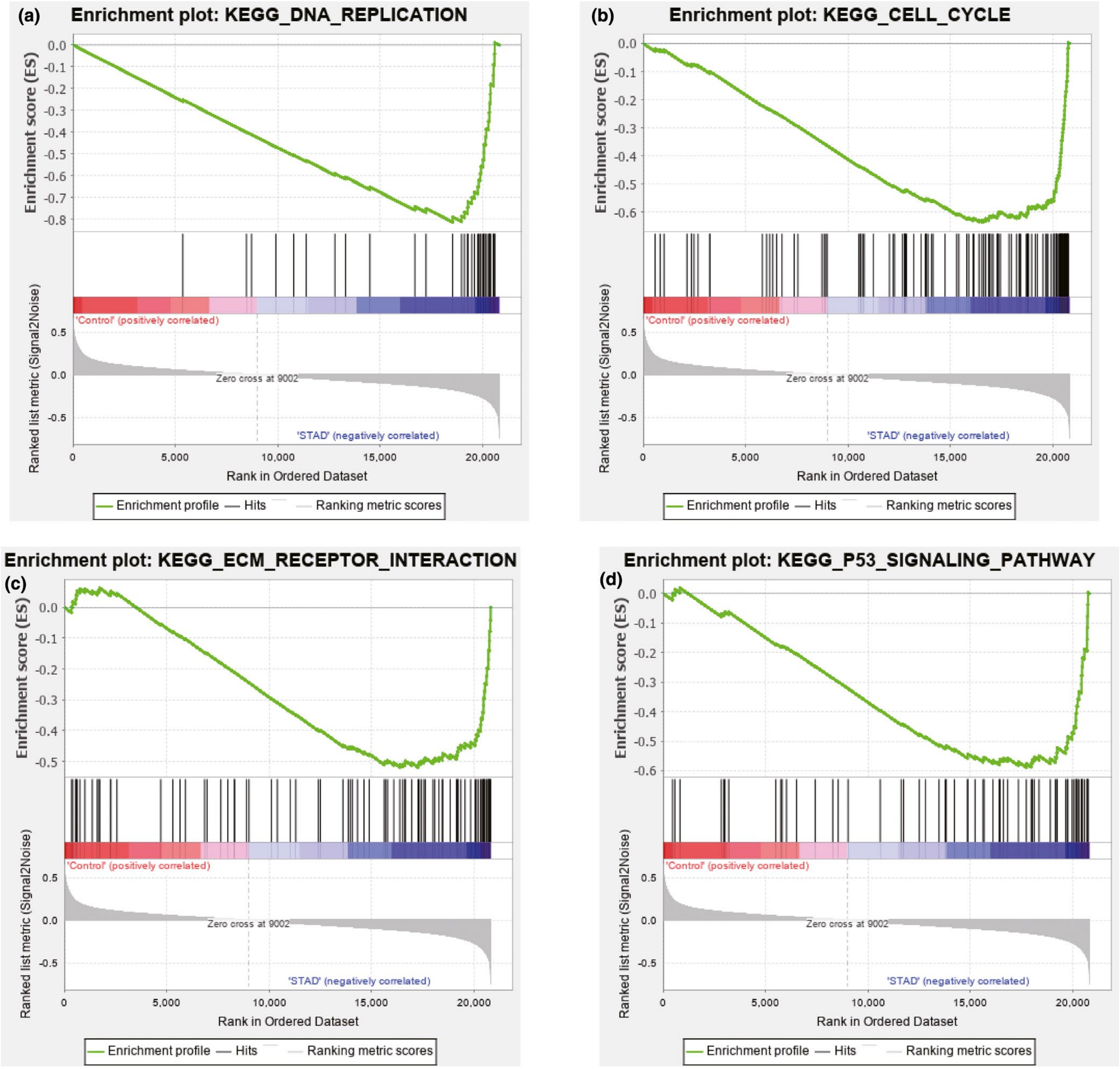

FIGURE 8 Enrichment plots from gene set enrichment analysis. (a) DNA replication. (b) Cell cycle. (c) ECM-receptor interaction. (d) P53 signaling pathway

regarded as a marker of angiogenesis in various cancer (Miao et al., 2016; Ozaki et al., 2014). Lv et al. demonstrated that *ESM1* level is upregulated in gastric cancer patients, and *ESM1* can be used as a potential biomarkers for early detection and prognosis of gastric cancer (Lv et al., 2014). *ESM1* expressing microvessel density correlates with the expression of vascular endothelial growth factor and is a prognostic factor for survival in gastric cancer (Chang et al., 2016). *ESM1* promotes gastric cancer cell proliferation, and *ESM1* level is associated with the pathological tumor stage (Zhao et al., 2014). In this study, our results displayed that *ESM1* was upregulated in both TCGA integration analysis and qRT-PCR validation. The results also showed that *ESM1* co-expressed with FOXD2-AS1. Therefore, we hypothesized that ESM1 might be involved in initiation and progression of STAD. Recently, Li et al. performed a more comprehensive study, and their study found that MAGI2-AS3 was overexpressed in STAD and associated with poor prognosis, and MAGI2-AS3 promotes tumor progression through sponging miR-141/200a to maintain overexpression of *ZEB1* in STAD and *BRD4* is a transcriptional regulator of MAGI2-AS3 in STAD (Li et al., 2020). We are collecting STAD samples to validate the expression of the identified optimal diagnostic lncRNAs in our following research with larger sample size. Then, the biological significances of optimal diagnostic lncRNAs will be investigated in in vivo and in vitro experiments.

In summary, our study found three DElncRNAs (FOXD2-AS1, LINC01235, and RP11-598F7.5) with diagnostic value for STAD. Among them, LINC01235 was not only an optimal diagnostic lncRNA biomarkers, but also related to survival time. Our results warrant further studies on these DEmRNAs and DElncRNAs to improve our comprehending of the STAD progression mechanisms. However, there are limitations to our study. First, the sample size for qRT-PCR confirmation was small and large numbers of STAD samples are needed for further research. Second, optimal diagnostic lncRNAs of STAD were identified and biological functions were not studied. Therefore, in vivo and in vitro experiments were necessary to uncover the biological functions of optimal diagnostic lncRNAs of STAD in the future work.

ACKNOWLEDGMENT

We thank Beijing Medintell Bioinformatic Technology Co., LTD for assistance in data analysis.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

Qun Li and Xiaofeng Liu contributed to the conception of the study. Jia Gu and Jinming Zhu contributed the materials and performed the experiment. Zhi Wei and Hua Huang performed the data analyses. Qun Li and Xiaofeng Liu contributed significantly in writing the manuscript. All authors read and approved the final manuscript.

ORCID

Xiaofeng Liu https://orcid.org/0000-0001-5596-7102

REFERENCES

Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202–209. https://doi.org/10.1038/nature13480

Cervantes, A., Roda, D., Tarazona, N., Rosello, S., & Perez-Fidalgo, J. A. (2013). Current questions for the treatment of advanced gastric cancer. *Cancer Treatment Reviews*, *39*(1), 60–67. https://doi.org/10.1016/j.ctrv.2012.09.007

Chang, Y., Niu, W., Lian, P. L., Wang, X. Q., Meng, Z. X., Liu, Y., & Zhao, R. (2016). Endocan-expressing microvessel density as a prognostic factor for survival in human gastric cancer. *World Journal of Gastroenterology*, *22*(23), 5422–5429. https://doi.org/10.3748/wjg.v22.i23.5422

Chen, G., Sun, W., Hua, X., Zeng, W., & Yang, L. (2018). Long non-coding RNA FOXD2-AS1 aggravates nasopharyngeal carcinoma carcinogenesis by modulating miR-363-5p/S100A1 pathway. *Gene*, *645*, 76–84. https://doi.org/10.1016/j.gene.2017.12.026

Cutler, D. R., Edwards, T. C. Jr, Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792.

Dey, B. K., Mueller, A. C., & Dutta, A. (2014). Long non-coding RNAs as emerging regulators of differentiation, development, and disease. *Transcription*, *5*(4), e944014. https://doi.org/10.4161/21541272.2014.944014

Ding, B., Gao, X., Li, H., Liu, L., & Hao, X. (2017). A novel microRNA signature predicts survival in stomach adenocarcinoma. *Oncotarget*, *8*(17), 28144–28153. https://doi.org/10.18632/oncotarget.15961

Gu, J., Li, Y., Fan, L., Zhao, Q., Tan, B., Hua, K., & Wu, G. (2017). Identification of aberrantly expressed long non-coding RNAs in stomach adenocarcinoma. *Oncotarget*, *8*(30), 49201–49216. https://doi.org/10.18632/oncotarget.17329

Huang, H., Li, T., Ye, G., Zhao, L., Zhang, Z., Mo, D., & Liu, H. (2018). High expression of COL10A1 is associated with poor prognosis in colorectal cancer. *OncoTargets and Therapy*, *11*, 1571–1581. https://doi.org/10.2147/ott.s160196

Katoh, M. (2001). Frequent up-regulation of WNT2 in primary gastric cancer and colorectal cancer. *International Journal of Oncology*, *19*(5), 1003–1007. https://doi.org/10.3892/ijo.19.5.1003

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500. https://doi.org/10.1093/bioinformatics/btp692

Li, D., Wang, J., Zhang, M., Hu, X., She, J., Qiu, X., Zhang, X., Xu, L. I., Liu, Y., & Qin, S. (2020). LncRNA MAGI2-AS3 is regulated by BRD4 and promotes gastric cancer progression via maintaining ZEB1 overexpression by sponging miR-141/200a. *Molecular Therapy—Nucleic Acids*, *19*, 109–123. https://doi.org/10.1016/j.omtn.2019.11.003

Li, L. I., Zhu, Z., Zhao, Y., Zhang, Q. I., Wu, X., Miao, B., Cao, J., & Fei, S. (2019). FN1, SPARC, and SERPINE1 are highly expressed and significantly related to a poor prognosis of gastric adenocarcinoma revealed by microarray and bioinformatics. *Scientific Reports*, *9*(1), 7827. https://doi.org/10.1038/s41598-019-43924-x

Li, T., Huang, H., Shi, G., Zhao, L., Zhang, Z., Liu, R., & Li, G. (2018). TGF-beta1-SOX9 axis-inducible COL10A1 promotes invasion and metastasis in gastric cancer via epithelial-to-mesenchymal transition. *Cell Death & Disease*, *9*(9), 849. https://doi.org/10.1038/s41419-018-0877-2

Liu, J., Liu, F., Shi, Y., Tan, H., & Zhou, L. (2018). Identification of key miRNAs and genes associated with stomach adenocarcinoma from The Cancer Genome Atlas database. *FEBS Open Bio*, *8*(2), 279–294. https://doi.org/10.1002/2211-5463.12365

Lv, Z., Fan, Y., Chen, H., & Zhao, D. (2014). Endothelial cell-specific molecule-1: A potential serum marker for gastric cancer. *Tumour Biology*, *35*(10), 10497–10502. https://doi.org/10.1007/s13277-014-2319-9

Miao, Y., Zong, M., Jiang, T., Yuan, X., Guan, S., Wang, Y., & Zhou, D. (2016). A comparative analysis of ESM-1 and vascular endothelial cell marker (CD34/CD105) expression on pituitary adenoma invasion. *Pituitary*, *19*(2), 194–201. https://doi.org/10.1007/s11102-015-0698-6

Ni, W., Xia, Y., Bi, Y., Wen, F., Hu, D., & Luo, L. (2019). FoxD2-AS1 promotes glioma progression by regulating miR-185-5P/HMGA2 axis and PI3K/AKT signaling pathway. *Aging (Albany NY)*, *11*(5), 1427–1439. https://doi.org/10.18632/aging.101843

Ozaki, K., Toshikuni, N., George, J., Minato, T., Matsue, Y., Arisawa, T., & Tsutsumi, M. (2014). Serum endocan as a novel prognostic biomarker in patients with hepatocellular carcinoma. *Journal of Cancer*, *5*(3), 221–230. https://doi.org/10.7150/jca.7691

Pan, Y., Liu, G., Wang, D., & Li, Y. (2019). Analysis of lncRNA-mediated ceRNA crosstalk and identification of prognostic signature in head and neck squamous cell carcinoma. *Frontiers in Pharmacology*, *10*, 150. https://doi.org/10.3389/fphar.2019.00150

Siegel, R., Ma, J., Zou, Z., & Jemal, A. (2014). Cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, *64*(1), 9–29. https://doi.org/10.3322/caac.21208

Solé, X., Crous-Bou, M., Cordero, D., Olivares, D., Guinó, E., Sanz-Pamplona, R., Rodriguez-Moranta, F., Sanjuan, X., de Oca, J., Salazar, R., & Moreno, V. (2014). Discovery and validation of new potential biomarkers for early detection of colon cancer. *PLoS One*, *9*(9), e106748. https://doi.org/10.1371/journal.pone.0106748

Su, F., He, W., Chen, C., Liu, M. O., Liu, H., Xue, F., Bi, J., Xu, D., Zhao, Y., Huang, J., Lin, T., & Jiang, C. (2018). The long non-coding RNA FOXD2-AS1 promotes bladder cancer progression and recurrence through a positive feedback loop with Akt and E2F1. *Cell Death & Disease*, *9*(2), 233. https://doi.org/10.1038/s41419-018-0275-9

Tsai, M.-M., Wang, C.-S., Tsai, C.-Y., Huang, H.-W., Chi, H.-C., Lin, Y.-H., Lu, P.-H., & Lin, K.-H. (2016). Potential Diagnostic, Prognostic and Therapeutic Targets of MicroRNAs in Human Gastric Cancer. *International Journal of Molecular Sciences*, *17*(6), 945. https://doi.org/10.3390/ijms17060945

Wang, Y., Chen, L., Ju, L., Qian, K., Liu, X., Wang, X., & Xiao, Y. (2019). Novel biomarkers associated with progression and prognosis of bladder cancer identified by co-expression analysis. *Frontiers in Oncology*, *9*, 1030. https://doi.org/10.3389/fonc.2019.01030

Wang, Y., Wu, N., Liu, J., Wu, Z., & Dong, D. (2015). FusionCancer: A database of cancer fusion genes derived from RNA-seq data. *Diagnostic Pathology*, *10*, 131. https://doi.org/10.1186/s13000-015-0310-4

Wei, C., Liang, Q., Li, X., Li, H., Liu, Y. I., Huang, X., Chen, X., Guo, Y., & Li, J. (2019). Bioinformatics profiling utilized a nine immune-related long noncoding RNA signature as a prognostic target for pancreatic cancer. *Journal of Cellular Biochemistry*, *120*(9), 14916–14927. https://doi.org/10.1002/jcb.28754

Xu, T.-P., Wang, W.-Y., Ma, P., Shuai, Y., Zhao, K., Wang, Y.-F., Li, W., Xia, R., Chen, W.-M., Zhang, E.-B., & Shu, Y.-Q. (2018). Upregulation of the long noncoding RNA FOXD2-AS1 promotes carcinogenesis by epigenetically silencing EphB3 through EZH2 and LSD1, and predicts poor prognosis in gastric cancer. *Oncogene*, *37*(36), 5020–5036. https://doi.org/10.1038/s41388-018-0308-y

Yang, H., & Li, H. (2019). CD36 identified by weighted gene co-expression network analysis as a hub candidate gene in lupus nephritis. *PeerJ*, *7*, e7722. https://doi.org/10.7717/peerj.7722

Zeng, J.-H., Liang, L., He, R.-Q., Tang, R.-X., Cai, X.-Y., Chen, J.-Q., Luo, D.-Z., & Chen, G. (2017). Comprehensive investigation of a novel differentially expressed lncRNA expression profile signature to assess the survival of patients with colorectal adenocarcinoma. *Oncotarget*, *8*(10), 16811–16828. https://doi.org/10.18632/oncotarget.15161

Zhang, Z., Wang, J., & Dong, X. (2018). Wnt2 contributes to the progression of gastric cancer by promoting cell migration and invasion. *Oncology Letters*, *16*(3), 2857–2864. https://doi.org/10.3892/ol.2018.9050

Zhao, W., Sun, M., Li, S., Wang, Y., & Liu, J. (2014). Biological and clinical implications of endocan in gastric cancer. *Tumour Biology*, *35*(10), 10043–10049. https://doi.org/10.1007/s13277-014-2287-0

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Li Q, Liu X, Gu J, Zhu J, Wei Z, Huang H. Screening lncRNAs with diagnostic and prognostic value for human stomach adenocarcinoma based on machine learning and mRNA-lncRNA co-expression network analysis. *Mol. Genet Genomic Med.* 2020;8:e1512. https://doi.org/10.1002/mgg3.1512