Abstract. Esophageal squamous cell carcinoma (ESCC) is a prevalent aggressive malignant tumor with poor prognosis. Investigations into the molecular changes that occur as a result of the disease, as well as identification of novel biomarkers for its diagnosis and prognosis, are urgently required. Long non-coding RNAs (lncRNAs) have been reported to play a critical role in tumor progression. The present study performed data mining analyses for ESCC via an integrated study of accumulated datasets and identification of the differentially expressed lncRNAs from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases. The identified intersection of differentially expressed genes (lncRNAs, miRNAs and mRNAs) in ESCC tissues between the GEO and TCGA datasets was investigated. Based on these intersected lncRNAs, the present study constructed a competitive endogenous RNA (ceRNA) network of lncRNAs in ESCC. A total of 81 intersection lncRNAs were identified; 67 of these were included in the ceRNA network. Functional analyses revealed that these 67 key lncRNAs primarily dominated cellular biological processes. The present study then analyzed the associations between the expression levels of these 67 key lncRNAs and the clinicopathological characteristics of the ESCC patients, as well as their survival time using TCGA. The results revealed that 31 of these lncRNAs were associated with tumor grade, tumor-node-metastasis (TNM) stage and lymphatic metastasis status (P<0.05). In addition, 15 key lncRNAs were demonstrated to be associated with survival time (P<0.05). Finally, 5 key lncRNAs were selected for validation of their expression levels in 30 patients newly diagnosed with ESCC via reverse transcription-quantitative PCR (RT-qPCR). The results suggested that the fold changes in the trends of up- and down-regulation between GEO, TCGA and RT-qPCR were consistent. In addition, it was also demonstrated that a select few of these 5 key lncRNAs were significantly associated with TNM stage and lymph node metastasis (P<0.05). The results of the clinically relevant analysis and the aforementioned bioinformatics were similar, hence proving that the bioinformatics analysis used in the present study is credible. Overall, the results from the present study may provide further insight into the functional characteristics of lncRNAs in ESCC through bioinformatics integrative analysis of the GEO and TCGA datasets, and reveal potential diagnostic and prognostic biomarkers for ESCC.

Introduction

According to statistics from the World Health Organization (WHO) in 2015, esophageal cancer ranks seventh in terms of global cancer incidence (572,000 new cases per year) and sixth overall in global cancer-associated mortality (509,000 deaths) (1). Approximately 85% of all esophageal cancer cases are esophageal squamous cell carcinoma (ESCC), and the most common cause of death is metastasis (2). Tumor staging and grading systems are important in the clinical diagnosis of cancer, but are currently not adequate for prognosis and prediction of the disease (3). Frequently, when patients with ESCC are diagnosed in clinical practice, the majority have already progressed to advanced stage and/or lymphatic metastasis. Furthermore, the survival rate in patients with ESCC is very poor, and there is a lack of specific biomarkers for early diagnosis and prognosis (4). Therefore, more efficient and accurate ESCC diagnostic and prognostic biomarkers are urgently required in order to improve these areas, including screening for the early stages of ESCC, as well as new effective treatment methods.

Accumulating evidence has revealed that dysregulated long non-coding RNAs (lncRNAs) play a number
of key biological roles in the progression of various types of cancer (5,6). An increasing number of studies have suggested that aberrant expression of lncRNAs in ESCC is closely associated with histological type, tumor-node-metastasis (TNM) stage, lymph node metastasis and prognosis (7,8).

ESCC is a multistep disease, which involves multiple interactions between genetic and environmental factors. Furthermore, lncRNAs play an important regulatory role in epigenetics. Investigating the ESCC-associated alterations of lncRNAs may aid the identification of valuable biomarkers for ESCC diagnosis and prognosis. Previous studies have primarily focused on the diagnostic and prognostic performance of a small portion of lncRNAs in ESCC, but a larger number of lncRNAs remain unexplored. Therefore, elucidating the functions of dysregulated lncRNAs, particularly in ESCC, is currently an important research topic.

Currently, high-throughput RNA sequencing technologies are being widely used for the detection of lncRNA alterations in carcinogenesis and screening for potential biomarkers of numerous diseases (9). However, the small sample sizes used for microarray detection often present a bias toward the identification of ESCC-associated lncRNAs due to lack of RNA sequencing data and, therefore, often generate errors (10). By using large sample sizes that integrate multiple RNA sequencing datasets, sufficient information regarding patients with ESCC can be obtained, thereby providing more convincing results. Thus far, with the advent of high-throughput RNA sequencing technologies, The Cancer Genome Atlas (TCGA; https://portal.gdc.cancer.gov/) and Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) database platforms have allowed the collected RNA sequencing data from microarray chips to be uploaded and standardized for quality control (11). Therefore, identification of the ESCC-associated lncRNAs would be more reliable with the use of large sample sizes integrating multiple analyses from different RNA sequencing databases.

The RNA sequencing databases obtained from the tissues of patients with ESCC are expected to provide a novel analysis strategy to identify diagnostic and prognostic biomarkers for this disease. In the present study, data mining analyses for ESCC were performed by integrating the significant differences in RNA obtained from the GEO and TCGA databases. Through these efforts, co-differentially expressed lncRNAs in ESCC may be identified. Based on these lncRNAs, subsequent analyses were performed, including gene functional enrichment analyses, competing endogenous RNA network construction, assessment of the association between differentially expressed lncRNAs and the clinicopathological characteristics of patients with ESCC, and survival analyses. Finally, reverse transcription-quantitative PCR (RT-qPCR) was used to validate the bioinformatics analysis results in the tumor and adjacent non-tumor tissues of 30 patients with newly diagnosed ESCC. This new approach may contribute to an improved method for identifying potential IncRNA biomarkers for the diagnosis of ESCC, as well as its classification and prediction of prognosis.

### Materials and methods

**Microarray dataset collection.** The present study collected RNA sequencing expression datasets (lncRNAs, miRNAs and mRNAs) comprising tissue samples from 802 patients with ESCC, which included cancer tissues and adjacent normal esophageal tissues from the GEO genomics database. The RNA sequencing datasets of patients with ESCC were downloaded from the GSE23400 database (12) (208 tissue samples: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23400), GSE26886 (13) (69 tissue samples: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26886), GSE45670 (14) (38 tissue samples: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45670), GSE97049 (14 tissue samples: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97049), GSE6188 (15) (257 tissue samples: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55856) datasets. All the aforementioned GEO datasets were obtained from Affymetrix Human Genome Array platforms. A total of 312 patients with ESCC were obtained from the TCGA database (up to November 1, 2018). Annotation information of the RNA sequencing datasets was obtained using the Affymetrix Human Genome Array platforms. The present study was fully compliant with the publication guidelines provided by the GEO and TCGA databases. The details of GEO and TCGA database ESCC patient tissue RNA sequencing datasets, sample descriptions and clinicopathological characteristics are provided in Tables I and II. A flow diagram of the integrated bioinformatics analysis from the GEO and TCGA databases is provided in Fig. 1.

**Annotation of microarray probes.** In order to enhance the comparability of the GEO database RNA sequencing data, the signature values of the GSE23400, GSE26886, GSE45670, GSE97049, GSE6188 and GSE55856 RNA sequencing datasets were downloaded and converted to new comparable transcripts. Probe reannotation methods were used to obtain a standard measurement of the RNA sequencing expression profiles. Once the aforementioned ESCC RNA sequencing probe signature values and probe IDs were downloaded from the GEO database, they underwent a mapped detection analysis in a database of human genomes, and those that were mismatched were excluded. The remaining probe positions on the chromosomes were identified using the GENCODE tool (https://www.gencodegenes.org/). Probes matched to both protein-coding genes and lncRNAs were also excluded. If multiple probes matched the same gene, the median values were instead used as the expression level of that gene. In addition, prior to any further analyses, RNA sequencing data normalization was performed to ensure the data were fully compliant with the publication guidelines of TCGA.

**Integration of microarray data and differential expression analysis.** All the ESCC tissue samples from the GEO database were downloaded from three or more datasets in order to enlarge the sample number and avoid generating less
A differential analysis was then separately performed for each dataset, comparing ESCC tumor tissues to adjacent normal tissues using the limma R 3.4.4 software [false discovery rate (FDR) <0.05, fold change >2, P<0.05]. Subsequently, overlapping subclass analyses were used to identify the separate co-differentially expressed genes in each dataset, including lncRNAs, miRNAs and mRNAs, with the Venn 2.1 tool (http://bioinfogp.cnb.csic.es/tools/venny/index.html). The integrated ESCC tissue dysregulated lncRNA, miRNA and mRNA lists were saved for further analysis.

Finally, according to the fold changes of differentially expressed lncRNAs, mRNAs and miRNAs in ESCC tissues from the GEO and TCGA databases, the common genes were selected for subsequent analysis.

### Table I. Details of ESCC studies and RNA sequencing microarray datasets from the GEO database.

<table>
<thead>
<tr>
<th>GSE</th>
<th>Publication</th>
<th>RNA sequencing styles</th>
<th>Sample size for each group</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE23400</td>
<td>Clinical cancer research</td>
<td>lncRNA</td>
<td>Tumor, 104</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mRNA</td>
<td>Adjacent normal tissues, 104</td>
</tr>
<tr>
<td>GSE26886</td>
<td>BMC Cancer</td>
<td>lncRNA</td>
<td>Tumor, 35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mRNA</td>
<td>Adjacent normal tissues, 34</td>
</tr>
<tr>
<td>GSE45670</td>
<td>Annals of oncology</td>
<td>lncRNA</td>
<td>Tumor, 28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mRNA</td>
<td>Adjacent normal tissues, 10</td>
</tr>
<tr>
<td>GSE97049</td>
<td>Unrecorded</td>
<td>miRNA</td>
<td>Tumor, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adjacent normal tissues, 7</td>
</tr>
<tr>
<td>GSE6188</td>
<td>Cancer research</td>
<td>miRNA</td>
<td>Tumor, 153</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adjacent normal tissues, 104</td>
</tr>
<tr>
<td>GSE55856</td>
<td>Gut</td>
<td>miRNA</td>
<td>Tumor, 108</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adjacent normal tissues, 108</td>
</tr>
</tbody>
</table>

ESC, esophageal squamous cell carcinoma; GEO, Gene Expression Omnibus.

Table II. Clinical information and samples size for TCGA ESCC datasets.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Total cases, n=312 (%)</th>
<th>Alive, n=138 (%)</th>
<th>Deceased, n=174 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>211 (67.63)</td>
<td>83 (60.14)</td>
<td>128 (73.56)</td>
</tr>
<tr>
<td>Female</td>
<td>101 (32.37)</td>
<td>55 (39.86)</td>
<td>46 (26.44)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>162 (51.92)</td>
<td>5 (39.86)</td>
<td>107 (61.49)</td>
</tr>
<tr>
<td>Asia</td>
<td>127 (40.71)</td>
<td>68 (49.28)</td>
<td>59 (33.91)</td>
</tr>
<tr>
<td>Black</td>
<td>23 (7.37)</td>
<td>15 (10.87)</td>
<td>8 (4.60)</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤50</td>
<td>86 (27.56)</td>
<td>53 (38.41)</td>
<td>33 (18.97)</td>
</tr>
<tr>
<td>&gt;50</td>
<td>226 (72.44)</td>
<td>85 (61.59)</td>
<td>141 (81.03)</td>
</tr>
<tr>
<td>Tumor grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GI</td>
<td>49 (15.71)</td>
<td>31 (22.46)</td>
<td>18 (10.34)</td>
</tr>
<tr>
<td>GII</td>
<td>146 (46.79)</td>
<td>48 (34.78)</td>
<td>98 (56.32)</td>
</tr>
<tr>
<td>GIII-IV</td>
<td>117 (37.50)</td>
<td>59 (42.75)</td>
<td>58 (33.33)</td>
</tr>
<tr>
<td>TNM stage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I/II</td>
<td>99 (31.73)</td>
<td>30 (21.74)</td>
<td>69 (39.66)</td>
</tr>
<tr>
<td>III/IV</td>
<td>213 (68.27)</td>
<td>108 (78.26)</td>
<td>105 (60.34)</td>
</tr>
<tr>
<td>Lymph node status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No metastasis</td>
<td>35 (11.22)</td>
<td>32 (23.19)</td>
<td>3 (1.72)</td>
</tr>
<tr>
<td>Metastasis</td>
<td>277 (88.78)</td>
<td>106 (76.81)</td>
<td>171 (98.28)</td>
</tr>
</tbody>
</table>


Table II. Clinical information and samples size for TCGA ESCC datasets.

Table II. Clinical information and samples size for TCGA ESCC datasets.

### Construction of the competing endogenous (ce) RNA network.

In the present study, an lncRNA, miRNA and mRNA ceRNA network was built based on the theory that lncRNAs can regulate miRNA abundance by sequestration binding, acting as ‘miRNA sponges’ through miRNA binding to the mRNAs and negatively regulating gene expression. Common ESCC tissues with significantly differentially expressed lncRNAs, miRNAs and mRNAs were selected to build the ceRNA network, in which the fold changes of genes were root in the TCGA database, in order to investigate whether these intersection lncRNAs, miRNAs and mRNAs were involved in ceRNA regulation. MiRcode (https://omictools.com/mircode-tool), miRanda (http://www.microrna.org/microrna/home.do) and
Targetscan (http://www.targetscan.org/) were used to predict the miRNA target lncRNAs and miRNA-mRNA interactions in the different databases. Finally, the predicted miRNA target genes and the significantly differentially expressed intersection genes in the GEO and TCGA databases were used to build the ceRNA network. A flow chart for the lncRNA, miRNA and mRNA ceRNA network construction is presented in Fig. 2.

During this process, the lncRNA, miRNA and mRNA ceRNA network was constructed using the fold change data of the significantly expressed RNA of patients with ESCC from TCGA database using the R package. Pearson’s correlation matrix models were used to identify the potential relevance of all pair-wise genes. Finally, based on the weighted adjacency matrix, the network connectivity of genes with other genes was investigated, and the ceRNA network was built using Cytoscape software (version 3.0).

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. In order to investigate the potential gene functional enrichment and signaling pathway regulation of the consensus mRNAs that were involved in the ceRNA network, GO (http://www.geneontology.org) and KEGG pathway analyses were performed. The mRNAs that were involved in the ceRNA network were uploaded to the GO database to investigate the enriched molecular functions of these mRNAs. Upregulated and downregulated mRNAs from the ceRNA network were analyzed. Furthermore, the KEGG tool (http://www.kegg.jp/) was used to identify the potential regulated signaling pathways

Figure 1. Flowchart for integrated bioinformatics analysis of ESCC publicly available RNA sequencing datasets from the GEO and TCGA databases. ESCC, esophageal squamous cell carcinoma; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas.
of these genes. R software was used to visualize the GO and KEGG results.

**Association between ceRNA network key lncRNAs and ESCC clinical status from TCGA.** Based on the co-expression of the lncRNAs, miRNAs and mRNAs in the ceRNA network, the key lncRNAs that were involved in the network were selected as target lncRNAs potentially associated with ESCC progression. Subsequently, the potential association between ceRNA network key lncRNAs and TCGA ESCC patients’ clinicopathological characteristics were investigated, which included sex, TNM stage, tumor grade, lymphatic metastasis status and pathological stage using multiple linear regression analysis.

**Kaplan-Meier survival analysis.** In order to investigate whether the key lncRNAs that were involved in the ceRNA network were associated with the TCGA database status, a Kaplan-Meier survival analysis of the ESCC patients from TCGA was performed. Based on the TCGA ESCC patient datasets, the fold changes of the aforementioned selected key lncRNAs in the cancer tissues and overall survival rates of patients with ESCC were determined using the GEPIA tool (http://gepia.cancer-pku.cn/). Kaplan-Meier survival analysis parameters were calculated using the publicly available TCGA ESCC patient datasets and GEPIA tools. The survival distributions of different TCGA with ESCC, and the expression changes in key lncRNAs, were examined using Kaplan-Meier analysis, log-rank test and hazard ratio (HR).

**Preparation of human ESCC samples and RT-qPCR validation of bioinformatics analysis results.** Next, 30 tumor and paired non-tumor esophageal tissues from patients with ESCC were collected from the Gansu Wuwei Tumor Hospital (Wuwei, China); the patients were aged 40-75 years. The samples were collected and stored at -80°C. All patients were diagnosed with ESCC according to their pathological examination results. Clinical basic information and informed consent forms were obtained from all ESCC patients (Table III). The collection of the tumor samples from patients with ESCC was approved by the Ethics Committee of the Gansu Wuwei Tumor Hospital.

Total RNA was isolated from tissue samples using TRizol® reagent (Invitrogen; Thermo Fisher Scientific, Inc.). Reverse Transcription Kit (Promega Corporation) and GoTaq® qPCR Master Mix of Power SYBR® Green (Promega Corporation) were used to synthesize cDNA and perform RT-qPCR analysis. The reaction was performed at 95°C for 2 min, followed by 40 cycles at 95°C for 15 sec, 60°C for 30 sec and 72°C for 30 sec. A dissociation curve was analyzed from 60-95°C. Finally, RT-qPCR was performed using the Step One Plus™ PCR System (Applied Biosystems; Thermo Fisher Scientific, Inc.). RT-qPCR relative fold change results were calculated using the 2^{-ΔΔCq} method (17).

**Statistical analysis.** All ESCC tissue RNA sequencing datasets from the GEO database were obtained from at least three independent datasets. R software was used to normalize the RNA sequencing data and compare significantly differentially expressed genes. Cytoscape 3.0 and GEPIA software tools were used to construct the ceRNA network and perform the survival analysis. Two-tailed Student's t-test was used to assess the differences between subgroups. P<0.05 was considered to indicate a statistically significant difference.

**Results**

**Convergence of gene expression signatures across the different datasets of patients with ESCC from GEO.** The ESCC RNA sequencing data and other information were obtained from the GEO database. In order to increase the veracity and reliability of the signal values, and also decrease the possibility of false-positives, three independent datasets were downloaded, including ESCC tissue lncRNA, miRNA and mRNA.
from the GEO database (Fig. 3; Table IV). Among those, 45 miRNAs with significantly changed levels in ESCC tissues expressed genes were analyzed, and it was demonstrated each dataset are presented in Table IV. The detailed information of these six datasets and GSE55856 dataset contained 198 differentially expressed miRNAs. Finally, the GSE6188 dataset contained 178 differentially expressed miRNAs. The GSE97049 dataset contained 7,407 mRNAs. The GSE45670 dataset contained 6,336 differentially expressed genes, including 1,340 differentially expressed lncRNAs and 45,518 mRNAs. The GSE23400 dataset contained 5,310 differentially expressed genes, including 1,340 differentially expressed lncRNAs and 7,407 mRNAs. The GSE45670 dataset contained 6,336 differentially expressed genes, including 914 differentially expressed lncRNAs and 5,422 mRNAs. The GSE97049 dataset contained 178 differentially expressed miRNAs. The GSE6188 dataset contained 198 differentially expressed miRNAs. Finally, the GSE55856 dataset contained 198 differentially expressed miRNAs. The detailed information of these six datasets and the number of differentially expressed genes identified from each dataset (Fig. 3). The GSE23400 dataset contained 5,310 differentially expressed genes, including 792 differentially expressed lncRNAs and 4,518 mRNAs. The GSE26886 dataset contained 8,747 differentially expressed genes, including 1,340 differentially expressed lncRNAs and 7,407 mRNAs. The GSE45670 dataset contained 6,336 differentially expressed genes, including 914 differentially expressed lncRNAs and 5,422 mRNAs. The GSE97049 dataset contained 178 differentially expressed miRNAs. The GSE6188 dataset contained 198 differentially expressed miRNAs. Finally, the GSE55856 dataset contained 198 differentially expressed miRNAs. The detailed information of these common significantly differentially expressed genes in ESCC tissues from GEO database can be found in Tables SI-SII.

ESCC tissues with co-differentially expressed genes identified in the GEO and TCGA databases. The ESCC patient tissues RNA sequencing results and their clinicopathological information were collected from TCGA database. Compared with the tumor and non-tumor tissue RNA sequencing results of patients with ESCC from TCGA, we identified a total of 889 differentially expressed lncRNAs (597 upregulated and 292 downregulated, 4,796 differentially expressed mRNAs (3,170 upregulated and 1,626 downregulated), and 438 differentially expressed mRNAs (299 upregulated and 139 downregulated) (Table IV). The Venn diagram in Fig. 4 presents the intersections of genes between the GEO and TCGA databases. A total of 81 co-differentially expressed lncRNAs, 39 miRNAs and 357 mRNAs were identified (Fig. 4 and Table V).

Construction of ceRNA network. In order to determine whether the intersection of co-differentially expressed lncRNAs, miRNAs and mRNAs in GEO and TCGA (Table V) exist in competing endogenous regulating relationships, the intersected 81 lncRNAs, 39 miRNAs and 357 mRNAs were applied to construct the ceRNA network. The aforementioned 39 miRNAs target lncRNAs and mRNAs were then predicted based on MiRcode, miRanda and Targetscan. Subsequently, the intersection genes of the abovementioned miRNAs target predicted the lncRNAs and mRNAs to build the ceRNA network (Table V). The ceRNA network was visualized using Cytoscape software (version 3.0). The results revealed that there were 67 lncRNAs, 37 miRNAs and 80 mRNAs involved in the ceRNA network (Fig. 5). The detailed information of ceRNA network is presented in Table III.

Functional analysis of mRNAs in the ceRNA network. The present study identified 67 lncRNAs, 37 miRNAs and 80 mRNAs involved in the ceRNA network. Therefore, it may be suggested that the co-differentially expressed genes in the ESCC tissues that were included in the ceRNA network may play key roles in ESCC progression. Based on these suggestions, the present study investigated the potential biological regulatory functions of these 80 mRNAs that were involved in the ceRNA network via GO enrichment of functions and KEGG pathway analyses. The upregulated and downregulated intersected mRNAs were further analyzed. The results suggested that the most enriched GO function by upregulated mRNAs was ‘Regulation of biological process’ (GO:0005789). The most enriched GO function by downregulated mRNAs was ‘Single-organism cellular process’ (GO:0044763) (Fig. 6). The KEGG pathway analysis indicated 25 signaling pathways involved in regulating upregulated mRNAs, and the most enriched pathway was ‘MicroRNAs in cancer’ (hsa05206). In addition, 95 signaling pathways were involved in regulating downregulated mRNAs, and the most enriched pathway was ‘Insulin secretion’ (hsa04911). Some of these signaling pathways, such as the ‘Wnt signaling pathway’, were involved in the progression of ESCC (18), the ‘MAPK signaling pathway’ was revealed as a key pathway affecting esophageal carcinoma cell proliferation and apoptosis (19), and the ‘mTOR signaling pathway’ was involved in the progression of ESCC (20). Furthermore, ‘Bladder cancer, Non-small-cell lung cancer, Pathways in cancer and PI3K-AKT signaling pathway’ were also reported as cancer-associated signaling pathways (21-24) (Fig. 7).

Association between lncRNA signature and the clinical characteristics of patients with ESCC. The aforementioned 67 lncRNAs that were involved in the ceRNA network were selected and further investigated in order to identify the association between these key lncRNAs and TCGA database with the clinicopathological characteristics of the 312 patients.
Figure 3. (A) Volcano plots of differentially expressed genes in GEO datasets; (B) Venn diagram demonstrates the differentially expressed intersection genes (lncRNAs, miRNAs and mRNAs) of GEO datasets. GEO, Gene Expression Omnibus; lncRNA, long non-coding RNA.

Table IV. Details information of differentially expressed genes in GEO database.

<table>
<thead>
<tr>
<th>GSE</th>
<th>Differentially expressed genes</th>
<th>Upregulated</th>
<th>Downregulated</th>
<th>Intersection differently expressed genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE23400</td>
<td>lncRNA</td>
<td>792</td>
<td>378</td>
<td>414</td>
</tr>
<tr>
<td>GSE26886</td>
<td>lncRNA</td>
<td>1340</td>
<td>599</td>
<td>741</td>
</tr>
<tr>
<td>GSE45670</td>
<td>lncRNA</td>
<td>914</td>
<td>508</td>
<td>406</td>
</tr>
<tr>
<td>GSE23400</td>
<td>mRNA</td>
<td>4518</td>
<td>2219</td>
<td>2299</td>
</tr>
<tr>
<td>GSE26886</td>
<td>mRNA</td>
<td>7407</td>
<td>3933</td>
<td>3474</td>
</tr>
<tr>
<td>GSE45670</td>
<td>mRNA</td>
<td>5422</td>
<td>2599</td>
<td>2823</td>
</tr>
<tr>
<td>GSE97049</td>
<td>miRNA</td>
<td>178</td>
<td>68</td>
<td>110</td>
</tr>
<tr>
<td>GSE6188</td>
<td>miRNA</td>
<td>198</td>
<td>48</td>
<td>150</td>
</tr>
<tr>
<td>GSE55856</td>
<td>miRNA</td>
<td>219</td>
<td>109</td>
<td>110</td>
</tr>
</tbody>
</table>

GEO, Gene Expression Omnibus.

Figure 4. Venn diagram demonstrating the intersections of genes between GEO and TCGA data. (A) The intersection of lncRNAs; (B) The intersection of miRNAs; (C) The intersection of mRNAs. GEO, Gene Expression Omnibus; lncRNA, long non-coding RNA.
with ESCC. The ESCC patients' clinicopathological characteristics included age, sex, race, tumor grade, TNM stage and lymphatic metastasis, and were obtained from TCGA database. The relevance analysis results suggested that 31 lncRNAs were significantly differentially expressed in ESCC patients with different clinicopathological characteristics (P<0.05).
It was revealed that PTGES2-AS1, CMAHP, LINC00472, LINC01105, BTN2A3P, LOC728743, LINC00346, PSMG3-AS1, SBF1P1, LINC01588, DIRC3, RPLPOP2, SMIM10L2A, ZNF300P1 and TP73-AS1 were associated with tumor grade; LOC148696, HCP5, LOC148696, LINC00472, RPLPOP2, DDX12P, SMIM10L2A, PVT1, ZNF300P1, SMIM10L2B and HAND2-AS1 were associated with TNM stage; and PTGES2-AS1, FOXD2-AS1, LOC148696, TCAM1P, LINC00982, LINC00176, DIRC3, ALOX12P2 and SBF1P1 were associated with lymphatic metastasis in patients with ESCC. Furthermore, it was indicated that PART1 and LINC00341 may be associated with sex and race, respectively (Table VI).

Prognostic analysis of lncRNA expression levels and survival of patients with ESCC. Based on the RNA sequencing data and clinical information of the patients with ESCC from TCGA, a Kaplan-Meier survival analysis was used to identify the associations between the ceRNA network 67 key lncRNAs and overall survival time of the ESCC patients. The expression levels of these 67 key lncRNAs and ESCC prognosis data in TCGA database were synthetically calculated using a Cox proportional hazard regression model and it was revealed that there were 15 lncRNAs statistically associated with the overall survival rate (log-rank P<0.05). Among these 15 lncRNAs, 7 (ZNF300P1, TP73-AS1, UG0898H09, CMAHP, DIRC3, LOC148696 and SMIM10L2A) were negatively associated with the prognosis of patients with ESCC (P<0.05), and 8 (RPLPOP2, HCP5, PVT1, LINC01588, PTGES2-AS1, DDX12P, CYP2D7 and LOC399815) were positively associated with the prognosis of patients with ESCC (P<0.05) (Fig. 8).

RT-qPCR validation. Through comprehensive analysis of the aforementioned results, it was inferred that LINC00982, TP73-AS1, SMIM10L2A, PVT1 and FOXD2-AS1 may play important roles in ESCC progression. Therefore, these five lncRNAs were selected and their actual expression levels were detected in 30 patients with newly diagnosed ESCC and paired non-tumor esophageal epithelial tissue samples via RT-qPCR to assess the reliability and validity of the bioinformatics analysis results (Tables III and VII). The results revealed that LINC00982, TP73-AS1 and SMIM10L2A were downregulated, whereas PVT1 and FOXD2-AS1 were upregulated in ESCC tumor tissues (Fig. 9). The results suggested that the RT-qPCR validation in 30 newly diagnosed patients with ESCC and the aforementioned bioinformatics analysis results (Table VII) exhibited the same trends (Fig. 5). Subsequently, the association between the real expression levels of the aforementioned 5 lncRNAs and the clinical characteristics of the 30 newly diagnosed patients with ESCC was assessed. The results revealed that TP73-AS1, SMIM10L2A and PVT1 were significantly associated with TNM stage (P<0.05), whereas LINC00982, SMIM10L2A and FOXD2-AS1 were significantly associated with lymph node
metastasis status (P<0.05). In addition, it was also revealed that TP73-AS1, SMIM10L2A and FOXD2-AS1 were correlated with the age and sex of the patients with ESCC (P<0.05; Fig. 10). The expression levels of the 5 lncRNAs were detected,
and the results from the clinically relevant analysis and the aforementioned bioinformatics analysis (Tables V and VII) were similar, suggesting that the bioinformatics analysis used in the present study was credible.

**Discussion**

ESCC is a common malignant tumor originating from the mucosa of the esophagus at vulnerable sites, including the
pharyngoesophageal junction, the part of the esophagus crossing with the posterior surface of the left bronchus, and the part of the esophagus passing through the diaphragmatic esophageal hiatus (25). Gastrointestinal endoscopy detection cannot identify all precancerous conditions or the early stages of ESCC. The incidence and mortality rate of ESCC in the
general population are high, and a large number of patients are usually diagnosed late and have a poor prognosis (26). In addition, ESCC lymph node metastasis may be regional, bidirectional, continuous or jumping, and effective complete eradication methods are currently lacking (27). In order to improve survival rate and the detection and treatment of early-stage ESCC, more effective therapeutic tools are required, including the identification of novel diagnostic and prognostic biomarkers. Recent studies that have focused on the abnormal expression of lncRNAs have revealed a new investigative approach to the pathological changes observed in ESCC, and have also indicated that the search for novel biomarkers may hold promise for ESCC diagnosis and prognosis (28-30).

The main concern is that small amounts of RNA sequencing samples may not accurately reflect the abnormal changes in lncRNAs when used as biomarkers that are associated with the diagnosis and prognosis of ESCC (31,32). Therefore, the ESCC-associated dysregulated lncRNA expression profiles should be identified based on large sample sizes in order to improve their accuracy and reliability as diagnostic and prognostic markers.

The development of high throughput RNA sequencing technologies has allowed thousands of dysregulated RNAs to be observed in a number of diseases (33-36). The present study identified abnormally expressed lncRNAs implicated in the pathogenesis of ESCC by synthetically analyzing RNA sequencing datasets obtained from the GEO and TCGA databases. A number of studies have proved that integration of multiple RNA sequencing datasets is a better method for enhancing the accuracy and reliability of results compared with using individual small samples (37). RNA sequencing allows reannotation, identification of differentially expressed RNAs in the GEO and TCGA databases, ceRNA network construction based on gene discovery, association between key lncRNAs and clinical characteristics of ESCC patients and survival analysis, in order to investigate the potential biomarkers for the diagnosis and prognosis of ESCC.

A multiple subset analysis and integrated bioinformatics approach was applied in order to identify dysregulated RNAs in ESCC in the present study. In order to increase the gene information accuracy, the RNA sequencing signals from the GEO database were matched to their chromosomal position using the GENCODE tool. All dataset samples were used in order to increase the sample numbers so as to avoid errors in the results and narrow the gene objects range. In addition, to avoid errors occurring from different ESCC individual tissues, only paired RNA sequencing sample datasets were included in the GEO database.

According to the integrated bioinformatics approach, it was revealed that there were 81 lncRNAs, 39 miRNAs and 357 mRNAs co-differentially expressed ESCC-associated intersection genes between the GEO and TCGA database. Among these 81 dysregulated lncRNAs, some have been indicated to exhibit significantly different expression levels in ESCC, including downregulated lncRNA HAND2-AS1 which may act as an anti-oncogene that can inhibit ESCC cell proliferation, migration and invasion (38). Dong et al (39) and Huang et al (40) have reported that downregulated MEG3 in ESCC tissues can inhibit ESCC cell growth and induce apoptosis. In addition, upregulated lncRNAs PVT1, TP73-AS1, UCA1 and ZFAS1 were also reported to be significantly differentially expressed in esophageal cancer, and also to be involved in regulating the progression of this disease (41-44).

To the best of our knowledge, the functions of the remaining dysregulated lncRNAs in the development and progression of ESCC have not yet been reported.

Based on the theory that lncRNAs can regulate miRNA abundance by sequestration binding, acting as ‘miRNA sponges’, and the fact that miRNAs can bind to mRNAs and negatively regulate gene expression (45), the present study built an lncRNA-miRNA-mRNA ceRNA network according to their...
negative regulatory associations. Among 81 ESCC-associated intersection lncRNAs, a total of 67 key lncRNAs were included in the lncRNA-miRNA-mRNA ceRNA network. The ceRNA network construction can reveal the potential regulatory associations between lncRNAs, miRNAs and mRNAs in ESCC. According to the ceRNA network created in the present study, some of these 67 lncRNAs, such as FIRRE, FOXD2-AS1, IPW and LINC00472, were also reported as potential diagnostic and prognostic biomarkers in numerous human diseases, including ESCC (46‑49). In order to investigate the potential regulatory functions of these 67 key lncRNAs, the present study further analyzed the 80 negatively regulated mRNAs in the ceRNA network, and revealed that some of those also play important roles in the progression of ESCC, including CYBRD1, EDA, ERBB4, FGF2 and IKBKE (50‑52). Subsequently, the present study analyzed the 67 key lncRNAs in the ceRNA network and identified enrichment of the functions of 80 mRNAs that were indirectly involved in signaling pathways. The GO analysis results revealed that the enriched GOs primarily targeted the metabolism and cell biological process. The KEGG analysis revealed that certain pathways were also associated with cancer, such as the Wnt signaling pathway, the MAPK signaling pathway, non-small cell lung cancer, pathways in cancer and the PI3K-AKT signaling pathway (53). Therefore, the present bioinformatics analysis results revealed that these 67 key lncRNAs in the ceRNA network may be implicated in the progression of ESCC.

The associations between the expression levels of the 67 aforementioned key lncRNAs and the clinicopathological characteristics from TCGA database were analyzed, and it was revealed that 32 lncRNAs were associated with the characteristics of the 312 patients with ESCC. These lncRNAs were primarily associated with tumor grade, TNM stage and lymphatic metastasis status. Among these 31 lncRNAs, PVT1, FOXD2-AS1 and PART1 have been reported to be the key genes involved in lymphatic metastasis and invasion and as diagnostic biomarkers for ESCC (54‑56). However, the functions of other lncRNAs have not yet been reported.
in association with ESCC progression. Therefore, the present study also analyzed the association between the 67 key lncRNAs in the ceRNA network and overall survival of patients with ESCC in TCGA database. It was revealed that 15 lncRNAs were associated with overall survival time. Among these 15 lncRNAs, only TP73-AS1 and PVT1 have been reported to be associated with survival in ESCC (42,57), whereas other lncRNAs have not been reported to date. The bioinformatics analysis results in the present study revealed potential lncRNAs biomarkers for the diagnosis, classification and prognosis of ESCC.

Subsequently, RT-qPCR validation of 5 randomly selected lncRNAs in 30 ESCC tissue samples was performed and the accuracy and credibility of the bioinformatics results were assessed. The results of the RT-qPCR validation were almost the same as the expression data in the GEO and TCGA databases. The present study then analyzed the association between these 5 randomly selected lncRNAs and the collected clinicopathological data of the patients with ESCC. The results also suggested that the results with these 5 lncRNAs were similar to the aforementioned bioinformatics analysis results. Therefore, the synthetic bioinformatics analysis results of the present study are reliable.

In summary, the present study has successfully identified specific ESCC-associated lncRNAs from large-scale samples through integrated analysis of RNA expression profile datasets of patients with ESCC from the GEO and TCGA databases. Differentially expressed lncRNAs and their potential functions in ESCC were investigated in the present study, along with specific ESCC-associated lncRNAs by different clinicopathological characteristics and overall survival time of patients with ESCC. Overall, these 67 lncRNAs are worth investigating further in terms of their applicability as biomarkers in the diagnosis, clinicopathological classification and prognosis of patients with ESCC.

Acknowledgements

Not applicable.

Funding

The present study was supported by the Fundamental Research Funds for the Central Universities (grant no. Izujbyk-2018-13), the National Natural Science Foundation of China (grant no. 81803188), and the Gansu Province Science and Technology Project (grant no. 1606RJYA270).

Availability of data and materials

The authors declare that all relevant raw data will be made freely available to any researchers who wish to use them for non-commercial purposes, whilst preserving any necessary confidentiality and anonymity. All data generated or analyzed during this study are included in this published article.

Authors' contributions

CYL conceived and designed the study. CYL and XHW performed the experiments. WWZ, JLW and JL analyzed and interpreted the results. JLX and XHW performed the patient tissue sample collection and quality control. CYL performed analysis and quality control, and was a major contributor to writing the manuscript. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

The present study was approved by the Ethics Committee of the Gansu Wuwei Tumor Hospital. All patients provided written informed consent to participate in the present study.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References


33. Li et al: FUNCTIONS of IncRNA in ESOPHAGEAL SQUAMOUS CELL CARCINOMA.


This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.