


















Research Article

Three Survival-Related Genes of Esophageal Squamous Cell Carcinoma Identified by Weighted Gene Coexpression Network Analysis

Di Lu ¹, He Wang ², Xuanzhen Wu ¹, Jianxue Zhai ¹, Xiguang Liu ¹, Xiaoying Dong ¹, Siyang Feng ¹, Xiaoshun Shi ¹, Jianjun Jiang ¹, Zhizhi Wang ¹, Zhiming Chen ¹, Shuhua Zhao ³, Jinhua Zhong ³, Gang Xiong ¹, Hua Wu ¹, Haofei Wang ¹ and Kaican Cai ¹

¹Department of Thoracic Surgery, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China

²Department of Thoracic Surgery, Peking University Shenzhen Hospital, Shenzhen 518000, China

³Department of Biological Information Research, HaploX Biotechnology, Shenzhen 518000, China

Correspondence should be addressed to Kaican Cai; doc_cai@163.com

Received 23 March 2021; Accepted 24 July 2021; Published 9 August 2021

Academic Editor: Chao Huang

Copyright © 2021 Di Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. The aim of this study was to identify novel biomarkers associated with esophageal squamous cell carcinoma (ESCC) prognosis. **Methods.** 81 ESCC samples collected from The Cancer Genome Atlas (TCGA) were used as the training set, and 179 ESCC samples collected from the Gene Expression Omnibus database (GEO) were used as the validation set. The protein-coding genes of 25 samples from patients who completed the follow-up in TCGA were analyzed to construct a coexpression network by weighted gene coexpression network analysis (WGCNA). Gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analyses were performed for the selected genes. The least absolute shrinkage and selection operator (LASSO) Cox regression model was constructed to analyze survival-related genes, and an optimal prognostic model was developed as well as evaluated by Kaplan–Meier and ROC curves. **Results.** In this study, a module containing 43 protein-coding genes and strongly related to overall survival (OS) was identified through WGCNA. These genes were significantly enriched in retina homeostasis, antimicrobial humoral response, and epithelial cell differentiation. Besides, through the LASSO regression model, 3 genes (PDLIM2, DNASE1L3, and KRT81) significantly related to ESCC survival were screened and an optimal prognostic 3-gene risk prediction model was constructed. ESCC patients with low and high OS in both sets could be successfully discriminated by calculating a risk score with the linear combination of the expression level of each gene multiplied by the LASSO coefficient. **Conclusions.** Our study identified three novel biomarkers that have potential in the prognosis prediction of ESCC.

1. Introduction

Esophageal cancer (EC) is a highly aggressive malignancy and one of the leading causes of cancer-related death worldwide [1]. There are two histologic subtypes of EC: esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC). Among them, ESCC is the main type of EC, accounting for about 90% [2]. Currently, the main treatments for ESCC include chemotherapy, radiation, and surgery [3]. Despite significant advances in the treatment of

ESCC in recent years, the overall 5-year survival rate for ESCC patients is still less than 25%, and the prognosis is still poor, with metastasis and recurrence frequently occurring [4]. In addition, due to late diagnosis and the lack of effective targeted therapy, most patients are diagnosed at an advanced stage [5]. Therefore, it is urgent to explore new therapeutic methods and new therapeutic targets.

Although many genes and mechanisms have been shown to be closely related to the occurrence and development of ESCC, the overall genes and regulation of ESCC

are still unclear. In recent years, with the development of high-throughput technology, bioinformatics has been increasingly applied to explore diseases, molecular mechanisms, and find biomarkers for diseases [6]. The Cancer Genome Atlas (TCGA) contains genomic and clinical information on many types of cancers, which is publicly available [7]. For example, Shergalis et al. performed bioinformatics analysis using TCGA and found that 20 genes were overexpressed and associated with poor survival outcomes in GBM patients [8]. Weighted gene coexpression network analysis (WGCNA) is one of the important methods to understand gene function and gene association from whole genome expression. It can be used to detect coexpression modules of highly related genes and modules of interest related to clinical features, providing good insights for predicting the function of coexpressed genes and discovering genes that play a key role in human diseases [9]. The least absolute shrinkage and selection operator (LASSO) is a penalty regression method that can be used to analyze gene expression profiles. In addition, due to the high dimension and collinearity of the Lasso Cox regression model, it can be combined with WGCNA for biomarker identification [10]. Therefore, a better understanding and application of the above methods are more conducive to the identification of ESCC-related genes and to explore their potential clinical roles and molecular mechanisms to understand the occurrence and development of ESCC. It has been applied to gene-network constructions and survival model identification for many kinds of cancer like lung adenocarcinoma [11]. In this study, RNA-Seq data from TCGA and microarray data from the Gene Expression Omnibus (GEO) database of ESCC were downloaded. Then, we conducted a WGCNA-based analysis in order to identify key modules and hub genes associated with ESCC pathogenesis. Additionally, the potential functions of genes in this identified module were analyzed by gene ontology (GO) and pathway-enrichment analyses. More importantly, a 3-gene risk prediction model was constructed using Cox and LASSO regression models, which could help us better predict ESCC prognosis.

2. Methods

2.1. Gene Expression Data and Clinical Data. Gene expression data and clinical data for patients with ESCC were obtained from TCGA (<https://cancergenome.nih.gov/>) and the GEO data repository (<http://www.ncbi.nlm.nih.gov/geo/>) on September 20, 2018, and included 173 samples and 358 samples, respectively. Gene expression levels from TCGA were measured by RNA sequencing, denoted by fragments per kilobase of transcript per million mapped reads (FPKM): $FPKM = 10^9 \times \text{number of reads mapped to the gene} / (\text{number of reads mapped to all protein-coding genes} \times \text{length of the gene in base pairs})$. Gene expression profiles from GEO were assessed by the Agilent-038314 CBC *Homo sapiens* microarray V2.0 and were analyzed using the GeneSpring software V11.5 (Agilent). Clinical information, including pathologic TNM stage and follow-up information, was also collected. This research flow chart is shown in Figure 1.

2.2. WGCNA Network Construction and Module Detection. A weighted gene coexpression network was constructed using the WGCNA package in R. The adjacency coefficient (a_{ij}) was calculated by the absolute value of Pearson's correlation coefficient of genes i and j to the power of β . $a_{ij} = |\text{cor}(x_i, x_j)|^\beta$, where x_i is the series of expression values for gene i . The lowest power β is chosen when the scale-free topology fit index curve flattens out upon reaching a high value. In addition to considering the connection between two correlated genes, WGCNA also takes into account associated genes, and the topological overlaps (T_{ij}) are calculated from a_{ij} as follows, to compose a topological overlap matrix (TOM), as a similarity evaluation reflecting relevancy and overlap between genes:

$$T_{ij} = \begin{cases} \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}, & i \neq j, \\ 1, & i = j, \end{cases} \quad (1)$$

$$l_{ij} = \sum_{u \neq i, j} a_{iu} a_{uj},$$

$$k_i = \sum_{u \neq i} a_{iu},$$

where u represents the common genes linking genes i and j together and T_{ij} takes account of the overlap between neighboring genes of genes i and j . TOM is subtracted from one and converted into a topological overlap dissimilarity matrix referred to as the corresponding dissimilarity of TOM (dissTOM). A hierarchical clustering tree (dendrogram) of genes is then created based on dissTOM. Finally, modules of highly correlated and coexpressed genes are created via a Dynamic Tree Cut algorithm.

2.3. Relating Modules to External Clinical Traits and Identifying Hub Genes. Correlations between modules and clinical traits, including pathologic stage and survival time, were estimated by Spearman's correlation tests. Significantly correlated modules were visualized using Cytoscape 3.7.1. Genes with multiple links were defined as hub genes.

2.4. Gene Ontology and Pathway-Enrichment Analysis. The potential biological functions and signaling pathways of the genes in the selected modules were analyzed by enrichment GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in Metscape (<http://metascape.org>).

2.5. LASSO Cox Regression Model Construction. LASSO Cox regression models were constructed using the glmnet package in R. By utilizing several hub genes from the selected modules, the function returned a series of values of λ and models. The coefficients of most original genes were penalized to zero in line with increasing values of the tuning parameter λ . λ was chosen when the partial likelihood deviance reached its lowest. A suitable model was then chosen

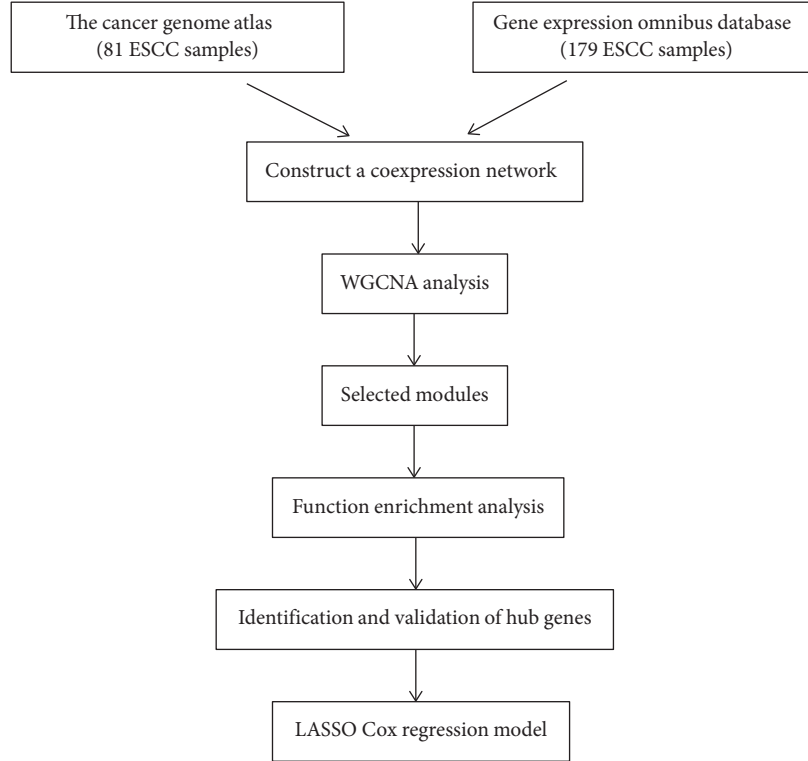


FIGURE 1: This flow chart of this study.

by 10-fold cross-validation using the function `cv.glmnet`. Using the function `lambda.min`, the remaining genes with nonzero LASSO coefficients were obtained. The risk score for each ESCC patient was calculated by a linear combination of the FPKM of each gene (G_k) multiplied by the LASSO coefficient (c_k):

$$\text{risk score} = \sum_{k=1}^n G_k \times c_k. \quad (2)$$

We followed the methods and routes of Wang et al. [11] to accomplish the gene network and survival model construction.

2.6. Statistical Analysis. Statistical analyses were conducted using SPSS (version 20.0). Receiver operating characteristic (ROC) curves were drawn, and the areas under the curves (AUC) were calculated to predict 1-year survival. The cut-off risk score was decided when the Youden index (sensitivity plus specificity minus 1) for the ROC curve was highest. The samples were then divided into high-risk and low-risk groups according to the cut-off. Survival was compared between the high- and low-risk groups using univariate and multivariable Cox regression analyses which were used to calculate the Hazard ratios (HRs). The independent significance of different factors was tested by a multivariate Cox regression analysis using backward selection to remove nonsignificant variables from the analysis. The P value threshold was 0.10 ($P > 0.10$).

3. Results

3.1. Data Preprocessing. Samples from TCGA were regarded as a training set and used for network and LASSO Cox regression model construction. A total of 173 samples from patients diagnosed between the ages of 28 and 90 years and classified as stage IA–stage IV were collected from TCGA, and 81 samples with both gene expression and clinical information were used for subsequent analysis. 25 samples from patients who completed the follow-up were subjected to sample clustering, and no outlier samples were removed before network construction (Figure 2(a)). The threshold for average gene expression value was set as 1. Protein-coding genes with average expression values less than the threshold value in all samples were excluded. A total of 25 samples including 12,439 protein-coding genes and clinical information were obtained for WGCNA.

Samples from GEO were regarded as a validation set and used for model verification. The microarray platform provided only a 60-base sequence for each probe, and the maximal exact matches (MEM) algorithm of the Burrows–Wheeler Alignment (BWA) Tool was used for sequence alignment. The parameters used were $-t 6$ and $-k 19$, and genome hg19 was used for reference. It was required that a transcript completely covered the genome position compared with the probe. 45,099 symbols were retained as protein-coding genes, and then levels of each probe were subjected to exponential transformation. One hundred and seventy-nine samples of paired normal tissues were removed, and the final validation set comprised 179 ESCC samples.

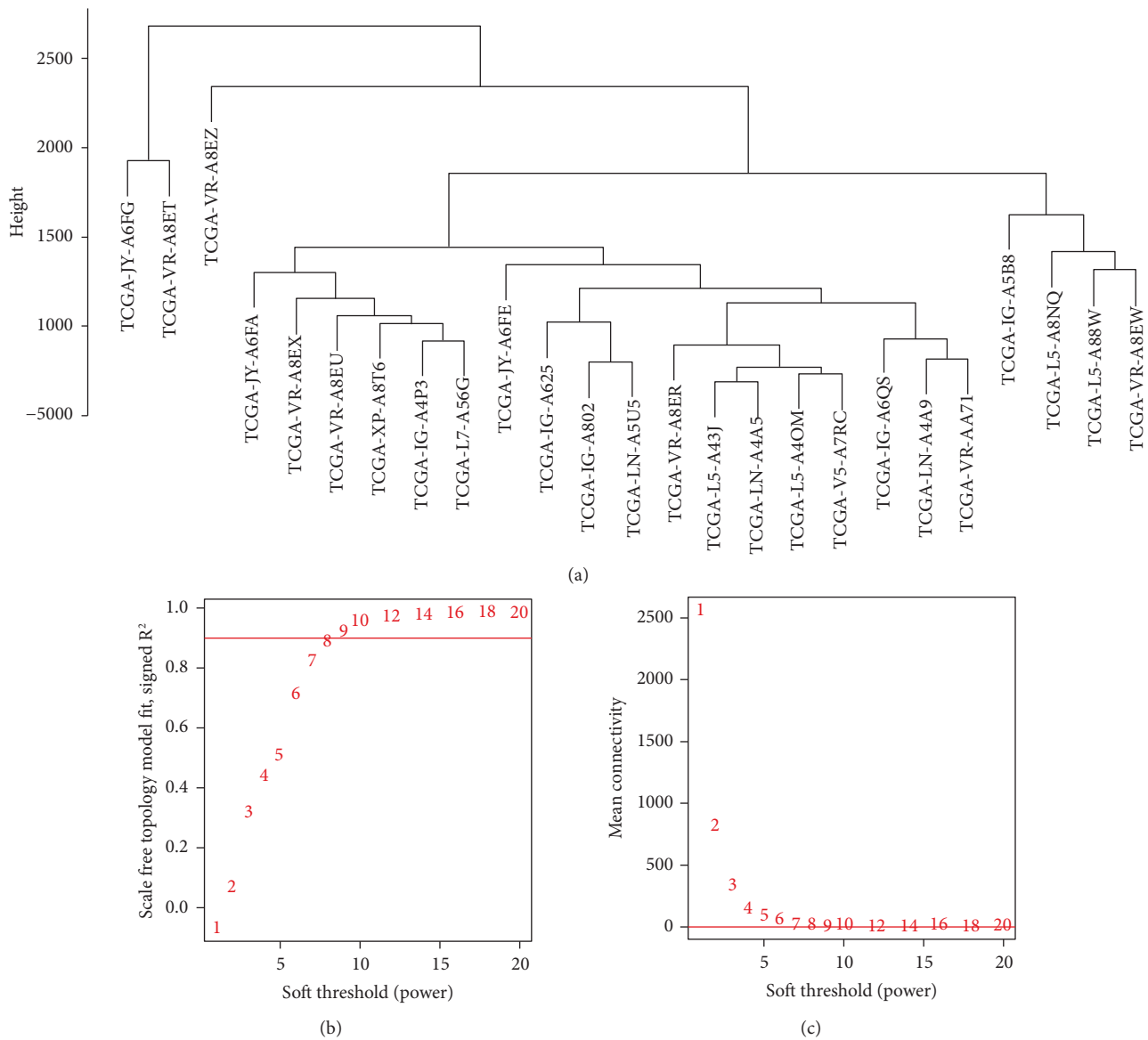


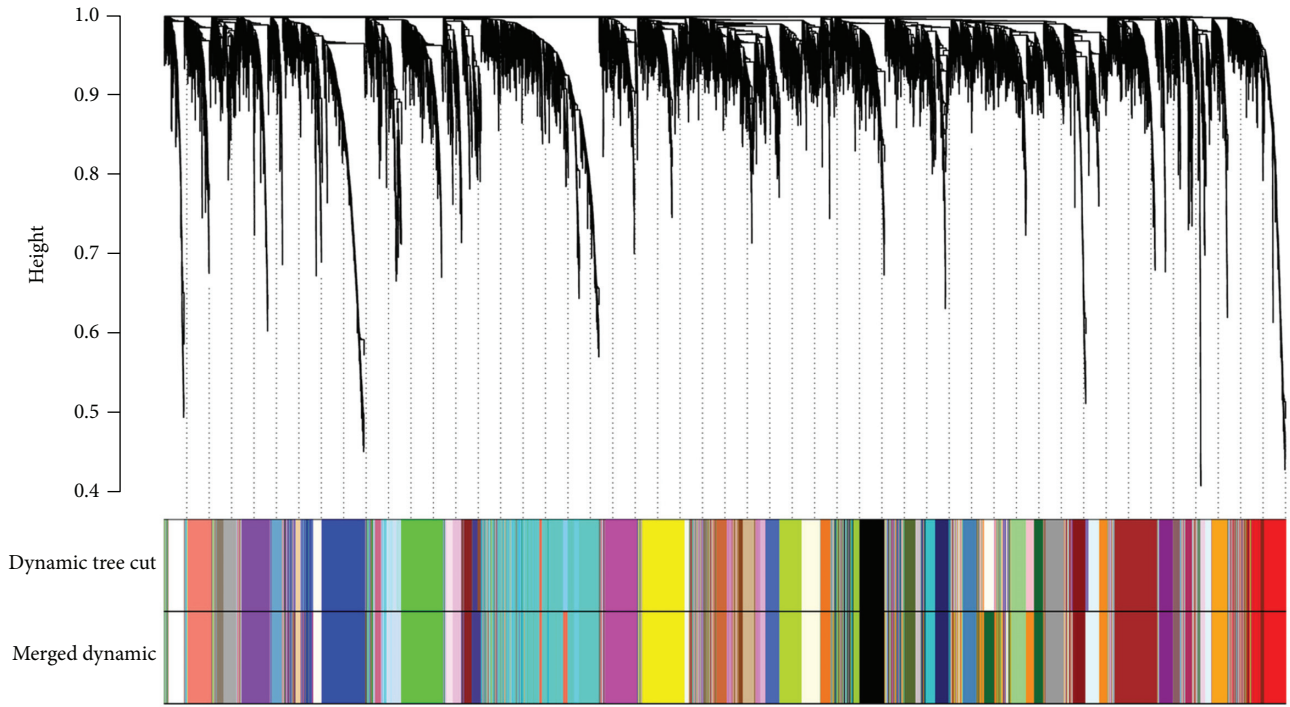
FIGURE 2: Sample clustering and determination of soft-threshold power in the WGCNA. (a) Sample clustering. Clustering dendrogram of samples based on their Euclidean distances. No outlier samples were removed. (b) β decision. Scale-free topology fit R^2 and series of soft thresholds. The red line indicates an R_2 value of 0.90. (c) Mean connectivity and series of soft thresholds. The red line indicates a mean connectivity value of 0.

3.2. Weighted Gene Coexpression Network of ESCC. As presented in Figure 2(b), the soft threshold is higher with the elevated R^2 , indicating that the network closely approaches to scale-free distribution. In this study, the soft thresholding power β was set as 8 and the scale-free topology fit index curve flattened out at 0.90 (Figure 2(c)). Then, cluster dendrogram was constructed and dynamic tree cut was performed (Figure 3(a)). Specifically, the constructed weighted gene coexpression network included 55 modules, including 36–875 genes. 36 genes that were not successfully integrated into any other modules were integrated into the gray module and were omitted in downstream analysis.

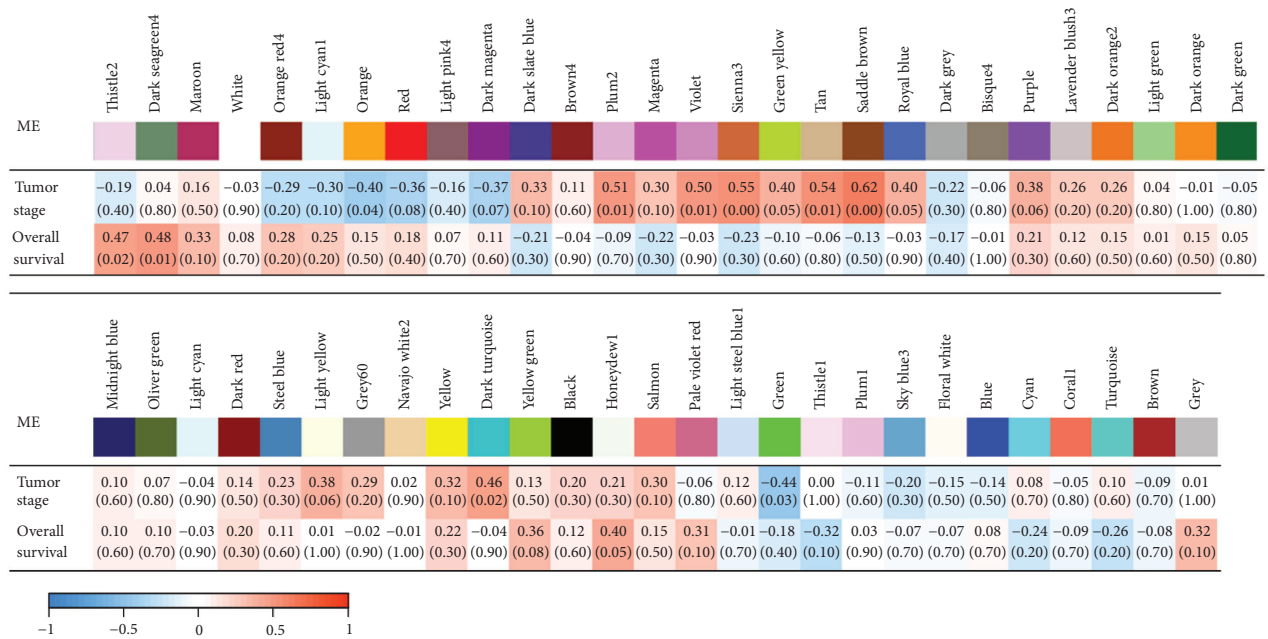
3.3. Identifying Modules with Clinical Significance. The correlations between each module and clinical traits including pathologic TNM stage and survival were calculated.

Among them, 6 modules were positively correlated pathologic TNM stage, whereas 2 modules are negatively correlated with pathological TNM staging. Besides, the dark seagreen4 module and thistle2 module are positively correlated with survival time, suggesting that these two modules may be involved in ESCC tumorigenesis (Figure 3(b)).

3.4. Functional Characterization of Genes in the Selected Module. To explore the biological function of the identified genes in the dark seagreen4 and thistle2 modules, GO term and KEGG pathway-enrichment analyses were performed on 43 and 106 genes in the two modules, respectively. Among the genes in the dark seagreen4 module, retina homeostasis, antimicrobial humoral response, and epithelial cell differentiation were the most significantly enriched



(a)



(b)

FIGURE 3: WGCN and their module-trait associations identified by the WGCNA. (a) Weighted gene coexpression network of ESCC identified 55 modules. A dendrogram was produced based on the WGCNA package in R by average linkage hierarchical clustering of 12,439 protein-coding genes. (b) Module-trait associations. Each column represents a module eigengene, and each row represents a clinical trait. Each cell contains the correlation coefficient (first line) and P value (in parentheses). A P value < 0.05 using Spearman's correlation test was considered statistically significant. ESCC: esophageal squamous cell carcinoma; ME: module eigengene; TNM: tumor-node-metastasis; WGCN: weighted gene coexpression network; WGCNA: weighted gene coexpression network analysis.

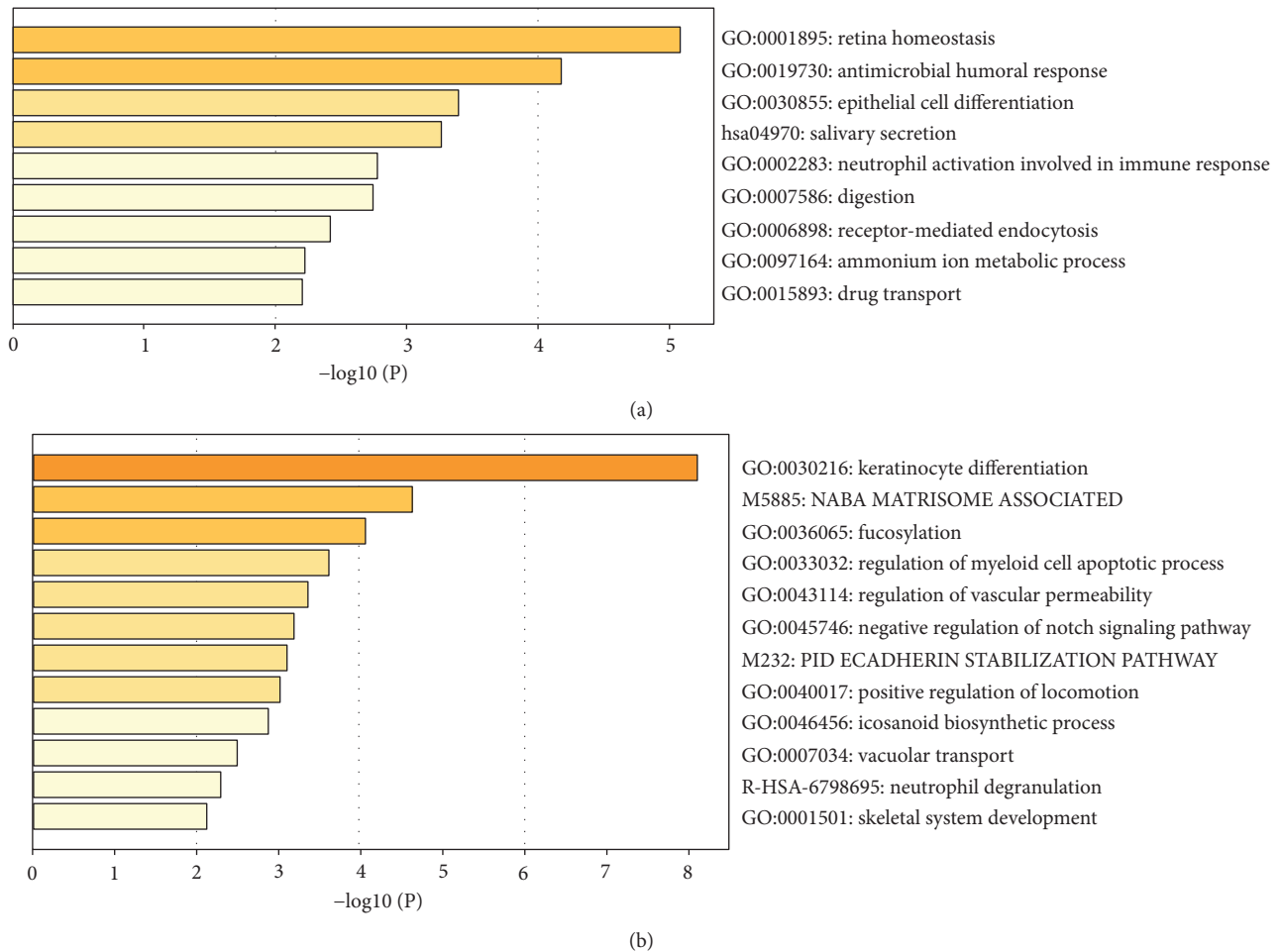


FIGURE 4: Top clusters in the dark seagreen4 module and the thistle2 module: (a) heatmap of enrichment clusters in the dark seagreen4 module; (b) heatmap of enrichment clusters in the thistle2 module.

(Figure 4(a)). While keratinocyte differentiation, NABA MATRISOME ASSOCIATED, and fucosylation were the most significantly enriched among the genes in the thistle2 module (Figure 4(b)).

Further, we analyzed the dark seagreen4 module with the strongest correlation with survival time by Cytoscape 3.7 to screen for hub genes. As shown in Figure 5, the top 20 genes with the most links were defined as hub genes in the dark seagreen4 module, including *AZGP1*, *CRISP3*, *KRT13*, *PDLIM2*, *PIGR*, *BPIFB1*, *BPIFB2*, *CLIC3*, *DNASE1L3*, *ENDOU*, *KRT81*, *MUC5B*, *PRR4*, *SCGB3A1*, *TFF3*, *SPINK7*, *ASPG*, *KRT6C*, *AQP5*, and *PCP4*.

3.5. Prognostic Signature Construction via LASSO Cox Regression Model Using the Training Set. The LASSO Cox regression model was constructed using the glmnet package in R by utilizing several hub genes in the dark-seagreen4 modules. Three genes (*PDLIM2*, *DNASE1L3*, and *KRT81*) with nonzero coefficients at the selected λ were obtained (Figures 6(a) and 6(b)). Based on the genes with nonzero coefficients, the risk score of every patient was calculated according to the linear combination of the

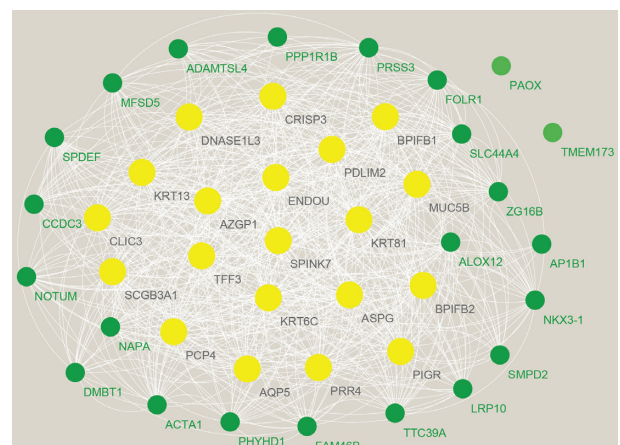
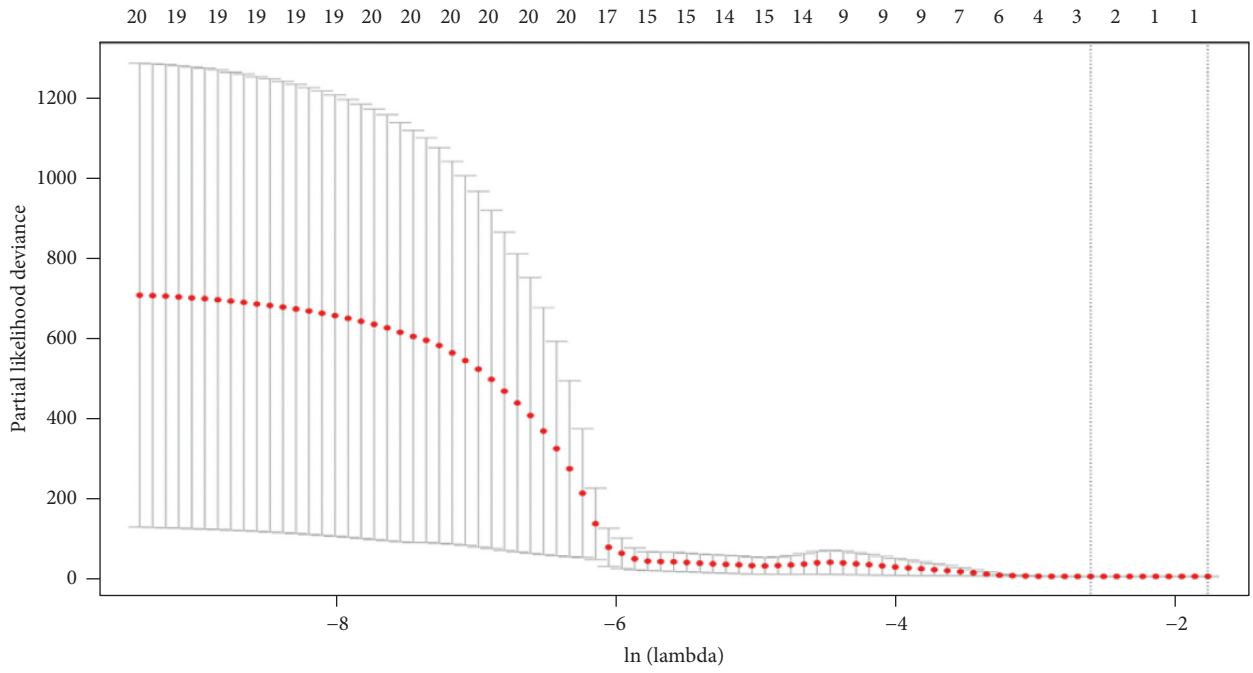
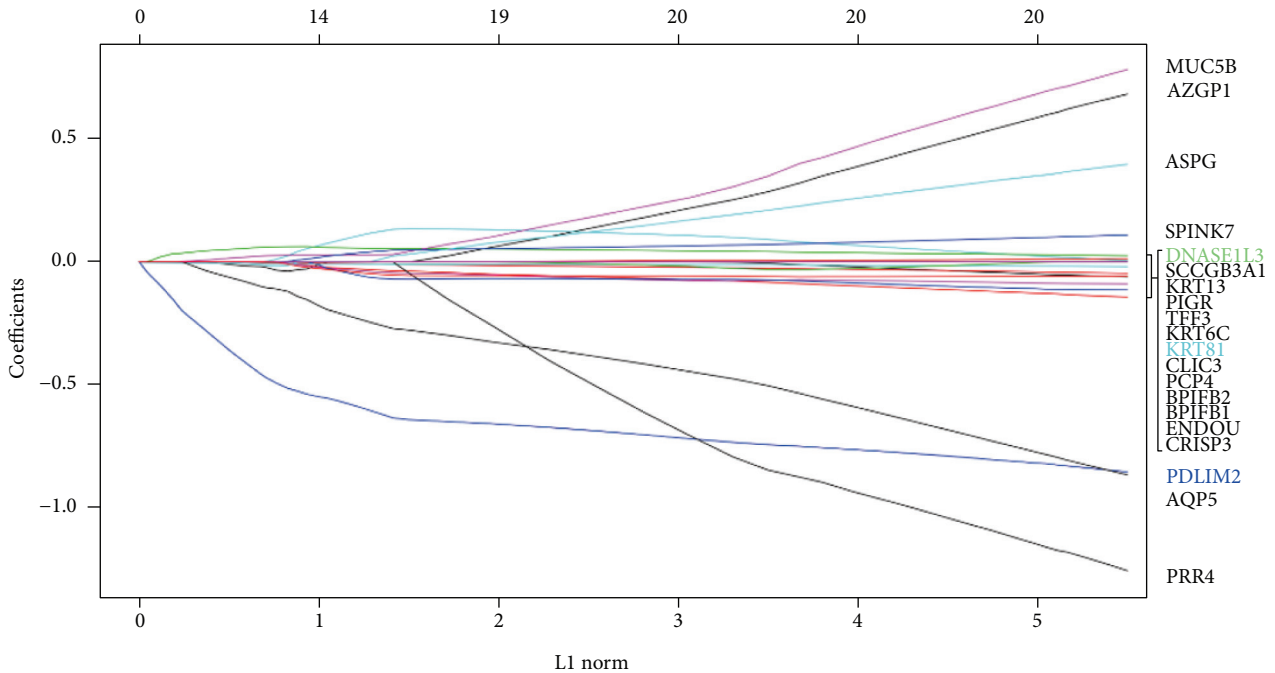


FIGURE 5: Coexpression network of the dark seagreen4 module. There are 43 connected genes in the dark seagreen4 module. Nodes are genes, and lines represent their connections. Twenty yellow nodes are the hub genes of the network.

expression of each gene multiplied by the LASSO coefficients: $(-0.0529 \times PDLIM2) + (0.0045 \times DNASE1L3) + (-0.0021 \times KRT81)$.



(a)



(b)

FIGURE 6: LASSO Cox regression model construction. (a) λ selection by 10-fold cross-validation. Continuous upright lines are partial likelihood deviance \pm standard error (SE); imaginary lines are depicted at the optimal values by minimum criteria (λ_{min} , left vertical dotted line) and 1-SE criteria ($\lambda_{1\text{se}}$, right vertical dotted line). The partial likelihood deviance with changing of $\log(\lambda)$ is plotted. The value 0.073 is chosen for λ by 10-fold cross-validation with the minimum criteria. (b) Process of LASSO Cox model fitting. Each curve represents a gene. The trend of each coefficient against the L1-norm is plotted when λ changes. L1-norm is the total absolute of nonzero coefficients. LASSO: least absolute shrinkage and selection operator; L1-norm: L1 regularization, the total absolute of nonzero coefficients; SE: standard error.

3.6. *Survival Analysis.* ROC curve was plotted to assess the 1-year survival, and results shown in Figure 7 reveal that the cut-off risk score was decided as -0.136 . As shown in the

Kaplan–Meier curve in Figure 8(a), patients in the high-risk group had worse overall survival (OS) than those in the low-risk group. The mean OS was 26.3 months (95% CI,

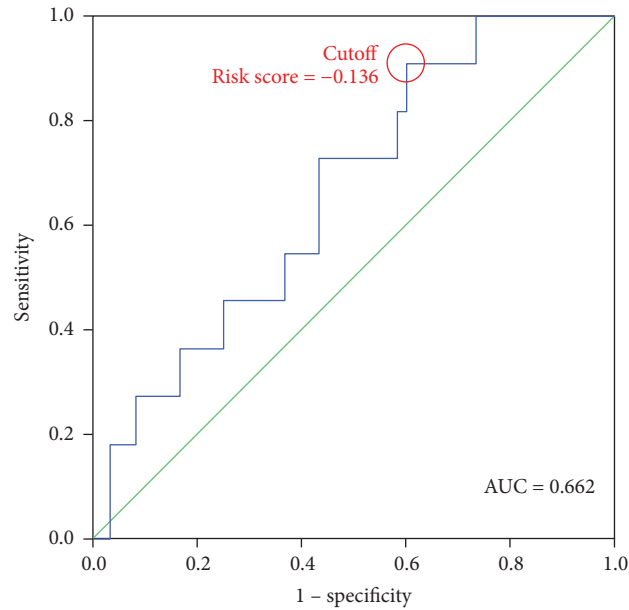


FIGURE 7: Cut-off risk score decision. The ROC curve is drawn, and AUC is calculated. The AUC is 0.662 (95% CI, 0.504–0.821) to predict 1-year survival using the training set. The cut-off risk score is decided as -0.136 . AUC: area under the curve; ROC: receiver operating characteristic.

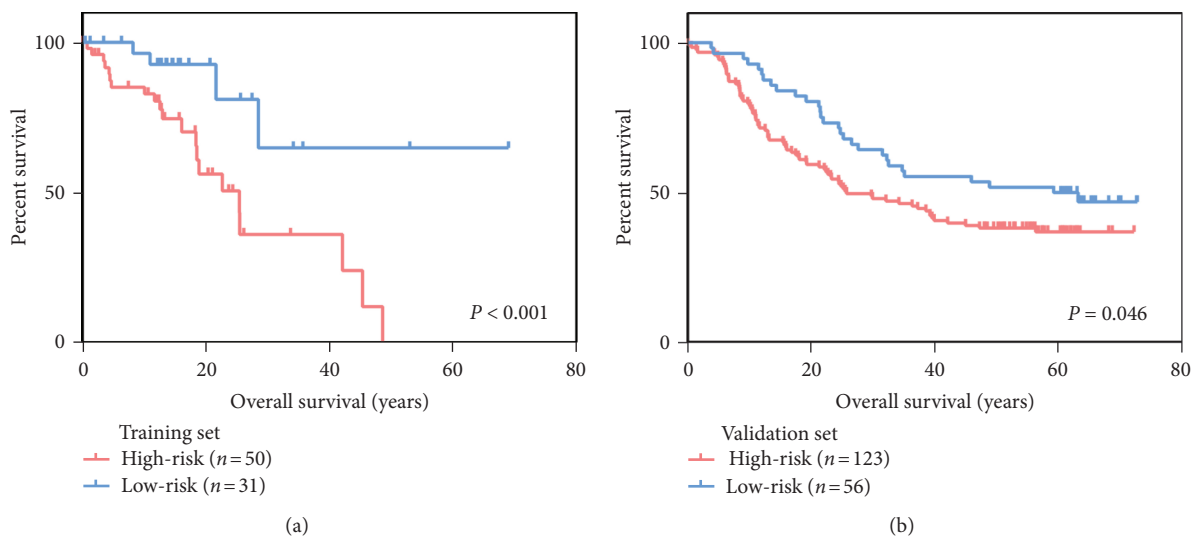


FIGURE 8: LASSO Cox regression model validation. (a) Survival comparison between the high- and low-risk groups of the training set using Kaplan–Meier analysis and log-rank tests. (b) Survival comparison between the high- and low-risk groups of the validation set using Kaplan–Meier analysis and log-rank tests. LASSO: least absolute shrinkage and selection operator.

20.1–32.5) in the high-risk group and 54.0 months (95% CI, 38.8–69.2) in the low-risk group ($P < 0.001$). Similarly, in the validation set, the mean OS was 38.0 months (95% CI, 33.1–43.0) in the high-risk group and 47.8 months (95% CI, 41.0–54.6) in the low-risk group ($P = 0.046$, Figure 8(b)).

3.7. Prognostic Factors Analysis. The risk score was then verified as an independent prognostic factor for OS, and univariate and multivariate analyses of potential prognostic factors in both training and validation set for OS were performed. In both sets, the risk score was correlated with

OS in univariable analyses. After multivariable analyses adjustment by clinicopathological variables, the risk score and TNM stage remained prognostic factors for OS in the training set (Table 1) and the risk score remained the exclusive independent predictive factor for OS in the validation set (Table 2). ESCC patients with a high-risk score had poorer OS (training set: HR 5.319, 95% CI: 1.576–17.951, and $P = 0.007$; validation set: HR 1.767, 95% CI: 1.112–2.807, and $P = 0.016$). Clinical and pathological TNM staging plays an important role in predicting the prognosis of patients with esophageal cancer. Compared with the TNM stage, except for stage IV patients in the training set (HR 10.372,

TABLE 1: Univariate and multivariable Cox regression analysis of prognosis factors in the training set for OS.

Variable	Univariate analysis			Multivariable analysis		
	HR	95% CI	P value	HR	95% CI	P value
OS						
Risk score (high vs. low)	5.319**	1.576–17.951	0.007	6.815***	1.840–25.248	0.004
TNM stage						
(Stage II vs. I)	1.035	0.221–4.843	0.965	1.758	0.326–9.485	0.512
(Stage III vs. I)	1.927	0.404–9.188	0.410	3.254	0.581–18.216	0.179
(Stage IV vs. I)	3.929	0.626–24.640	0.144	10.372*	1.345–79.977	0.025
Age (≥ 60 vs. < 60 years)	1.321	0.585–2.983	0.503			

Abbreviation. CI: confidence interval; HR: hazard ratio; OS: overall survival; TNM: tumor-node-metastasis; * $P < 0.05$; ** $P < 0.01$; and *** $P < 0.01$.

TABLE 2: Univariate and multivariable Cox regression analysis of prognosis factors in the validation set for OS.

Variable	Univariate analysis			Multivariable analysis		
	HR	95% CI	P value	HR	95% CI	P value
OS						
Risk score (high vs. low)	1.543*	1.004–2.372	0.048	1.767*	1.112–2.807	0.016
TNM stage						
(Stage II vs. I)	1.260	0.500–3.174	0.625	1.362	0.540–3.437	0.513
(Stage III vs. I)	1.327	0.531–3.317	0.545	1.762	0.690–4.500	0.236
Age (≥ 60 vs. < 60 years)	0.911	0.623–1.335	0.634			

Abbreviation. CI: confidence interval; HR: hazard ratio; OS: overall survival; TNM: tumor-node-metastasis; and * $P < 0.05$.

95% CI: 1.345–79.977 $P = 0.025$), stage II patients and stage III patients did not have poorer OS than stage I patients. Unexpectedly, age had little effect on ESCC.

4. Discussion

With the development of microarray and RNA sequencing technology, a new era of biological big data is coming [12]. Data mining strategies can be used to explore key biological phenotypes related to high-dimensional data sets and can also realize the characterization of human cancer, the identification and definition of important genes in the tumorigenesis process, and the diagnosis of prognostic characteristics [13–15]. In this study, 10 gene coexpression modules were identified by WGCNA analysis of coexpression gene networks in TCGA and GSE53625 datasets. Of these modules, dark seagreen4 and thistle2 module are closely related to the survival time of ESCC, especially dark seagreen4. Therefore, it is speculated that the genes contained in module 4 are the key regulatory factors of ESCC, so we conducted further analysis. GO and KEGG enrichment analyses were used to evaluate the potential functional role of gene modules in ESCC [16]. It was observed that the genes in dark seagreen 4 were mainly involved in antimicrobial humoral response, epithelial cell differentiation, and neutrophil activation involved in immune response, all of which were involved in the cancer development. More importantly, based on the LASSO Cox regression model analysis of the TCGA data set, we screened out 3 genes that can predict overall survival, namely, PDLIM2, DNASE1L3, and KRT81 and constructed the best prognostic 3-gene risk prediction model.

PDZ and LIM domain containing protein 2 (PDLIM2), encoded by the *PDLIM2* gene, is a member of the actin-associated LIM family of proteins, which plays an important

role in cytoskeletal organization and cell differentiation and is related to tumorigenesis [17]. Previous studies have shown that PDLIM2 plays a biological role as a tumor suppressor or a tumor promoter in different malignancies. For example, in human castration-resistant prostate cancer (CRPC)-like cells, PDLIM2 was significantly upregulated, while the inhibition of PDLIM2 can reduce the malignant phenotype in CRPC-like cells [18]. In colon cancer, PDLIM2 was significantly downregulated and involved in cancer regulation by inhibiting NF- κ B activation [19]. These studies confirmed the tumor specificity of PDLIM2.

DNASE1L3 encodes an enzyme called deoxyribonuclease gamma. Monogenic lupus can be caused by nucleic acid degradation and repair defects because DNA that has not been properly cleared can serve as an immunogen library to drive the response of T cells and B cells [20]. Interestingly, the mutation in *DNASE1L3* abolished the functional activity of the nuclease and caused defective DNA degradation, resulting in complete permeability of systemic lupus erythematosus [21]. DNase1L3, a genetically engineered human recombinant DNase, was described as attaining safeguarding stem cell-based regenerative therapy against iatrogenic cancerogenesis [22]. Malecki et al. found that the targeted expression of recombinant DNASE1, DNASE1L3, DNASE2, and DFFB can completely eradicate ovarian cancer cells in vitro, while healthy cells are not affected [23].

KRT81 encodes Hb-1, a type of keratin, which is expressed in all epithelial cell types. According to reports, it plays an important role in maintaining cell integrity, protein synthesis, and intracellular signal transduction [24,25]. Xie et al. confirmed that *KRT81* rs3660GG type is an independent prognostic marker in non-Hodgkin's lymphoma [26]. Campayo et al. revealed that SNP-*KRT81* plays an important role in the recurrence of nonsmall cell lung cancer

[27]. In addition, KRT81 also proved to be a novel and promising marker for squamous cell lung cancer [27].

Recent studies constructed prognostic models for ESCC based on gene expression data, with the aim of improving its early diagnosis and personalized treatment. Song et al. integrated the analysis of ESCC microarray data from GDS3838 and TCGA-ESCC and confirmed the prognostic value of PDLIM2 expression, which was independently associated with longer OS in ESCC patients [28]. In the current study, the risk score tends to be larger with lower expression of PDLIM2, and ESCC patients with a high-risk score had significantly poorer OS. As two independent studies, these similar findings suggested the accuracy and reliability of the analysis process. Alaei et al. constructed a large coexpression network of coding and noncoding genes using the WGCNA method and found important functional modules. Kaplan–Meier estimators and log-rank test statistics were then used to identify the majority of selected protein-coding and noncoding genes associated with poor prognosis in ESCC [29]. In this study, LASSO Cox regression model analysis was applied to construct a prognostic 3-gene risk prediction model. After the optimal cut-off point was determined by ROC analysis, the patients were divided into high-risk group and low-risk group. The results showed that the high-risk group had a lower percentage of survival, while the low-risk group had a higher percentage of survival. In addition, the verification result of risk score as an independent prognostic factor of OS showed that whether it is in the training set or the validation set, the risk score is its prognostic factor.

Zhan et al. developed a 3-gene signature by an ESCC-specific protein-protein interaction (PPI) network involving LOXL2 and actin-related proteins [30]. Sun et al. identified a three-gene prognostic signature based on the expression of GASC1 and 6 GASC1-targeted genes [31]. The signature was verified as an independent prognostic factor of ESCC. Guo et al. developed an immunogenomic risk score for predicting survival outcomes among esophageal cancer (EC) patients by utilizing immune-related genes [32]. Fei et al. constructed two prognostic risk score models of two EC sub-types, respectively [33]. The risk scores were verified as prognostic factors for EC OS and a classifier for the evaluation of groups with different risks. Since the underlying biological mechanisms of most ESCC biomarkers remains unclear, screening through candidate or pathway-based strategies rather than systematic screening results in limited prediction effect [34].

However, our prognostic model had some limitations. Firstly, we did not obtain other clinical information that might have influenced OS, such as primary health problems and follow-up treatment. Secondly, gene expression profiles obtained from GSE53625 were assessed by Agilent-038314 CBC *Homo sapiens* microarray V2.0 and have already been analyzed, so we did not start with the raw data. Thirdly, further validation in an independent set such as with RT-qPCR validation is required to confirm the diagnostic value of our model. Fourthly, we are unsure whether the risk score is feasible for metastatic ESCC because these samples were from primary tumors, the initial site of the cancer. The risk

score was obtained from the dark seagreen4 module, which showed the strongest correlation with survival time. Further explorations are needed to detect markers from other modules.

5. Conclusion

In this study, a survival-related risk score for ESCC was identified using the WGCNA and a LASSO Cox regression model to explore a new molecular characterization of ESCC associated with prognosis, and we produced a risk model to aid its diagnosis and management. Our results suggested that three novel markers could effectively be of diagnostic and therapeutic value for the management of ESCC.

Data Availability

The data underlying this article are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Di Lu, He Wang, and Xuanzhen Wu contributed to this study equally.

Acknowledgments

The authors would like to thank the 27th European Conference on General Thoracic Surgery for the presentation of the abstract of this work. This work was supported by the Science and Technology Planning Project of Guangdong Province (No. 2017B020226005) and Start-Up Foundation for Scientific Research of Southern Medical University (Nos. PY2018N030 and PY2018N028).

References

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] J. R. Siewert and K. Ott, "Are squamous and adenocarcinomas of the esophagus the same disease?" *Seminars in Radiation Oncology*, vol. 17, no. 1, pp. 38–44, 2007.
- [3] A. Pennathur, M. K. Gibson, B. A. Jobe, and J. D. Luketich, "Oesophageal carcinoma," *The Lancet*, vol. 381, no. 9864, pp. 400–412, 2013.
- [4] J. Hu, R. Li, H. Miao, and Z. Wen, "Identification of key genes for esophageal squamous cell carcinoma via integrated bioinformatics analysis and experimental confirmation," *Journal of Thoracic Disease*, vol. 12, no. 6, pp. 3188–3199, 2020.
- [5] H. Zeng, R. Zheng, S. Zhang et al., "Esophageal cancer statistics in China, 2011: estimates based on 177 cancer registries," *Thoracic Cancer*, vol. 7, no. 2, pp. 232–237, 2016.
- [6] T. Can, "Introduction to bioinformatics," *miRNomics: MicroRNA Biology and Computational Analysis*, vol. 1107, pp. 51–71, 2014.

- [7] J. N. Weinstein, E. A. Collisson, E. A. Collisson et al., “The cancer genome Atlas pan-cancer analysis project,” *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.
- [8] A. Shergalis, A. Bankhead 3rd, U. Luesakul, N. Muangsinsin, and N. Neamati, “Current challenges and opportunities in treating glioblastoma,” *Pharmacological Reviews*, vol. 70, no. 3, pp. 412–445, 2018.
- [9] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, “Comparing statistical methods for constructing large scale gene networks,” *PLoS One*, vol. 7, no. 1, Article ID e29348, 2012.
- [10] R. Tibshirani, “The lasso method for variable selection in the Cox model,” *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.
- [11] H. Wang, D. Lu, X. Liu et al., “Survival-related risk score of lung adenocarcinoma identified by weight gene co-expression network analysis,” *Oncology Letters*, vol. 18, pp. 4441–4448, 2019.
- [12] H. Zhao, S. Zhang, S. Shao, and H. Fang, “Identification of a prognostic 3-gene risk prediction model for thyroid cancer,” *Frontiers in Endocrinology*, vol. 11, p. 510, 2020.
- [13] A. Hébrant, G. Dom, M. Dewaele et al., “mRNA expression in papillary and anaplastic thyroid carcinoma: molecular anatomy of a killing switch,” *PLoS One*, vol. 7, no. 10, Article ID e37807, 2012.
- [14] H. Zhang, X. Zhao, M. Wang, and W. Ji, “Key modules and hub genes identified by coexpression network analysis for revealing novel biomarkers for larynx squamous cell carcinoma,” *Journal of Cellular Biochemistry*, vol. 120, no. 12, pp. 19832–19840, 2019.
- [15] Y. Zou and L. Jing, “Identification of key modules and prognostic markers in adrenocortical carcinoma by weighted gene co-expression network analysis,” *Oncology Letters*, vol. 18, pp. 3673–3681, 2019.
- [16] N. Chen, G. Zhang, J. Fu, and Q. Wu, “Identification of key modules and hub genes involved in esophageal squamous cell carcinoma tumorigenesis,” *Using WCGNA*, vol. 27, Article ID 1073274820978817, 2020.
- [17] Z. Qu, J. Fu, P. Yan, J. Hu, S.-Y. Cheng, and G. Xiao, “Epigenetic repression of PDZ-LIM domain-containing protein 2,” *Journal of Biological Chemistry*, vol. 285, no. 16, pp. 11786–11792, 2010.
- [18] M. Kang, K.-H. Lee, H. S. Lee et al., “PDLIM2 suppression efficiently reduces tumor growth and invasiveness of human castration-resistant prostate cancer-like cells,” *The Prostate*, vol. 76, no. 3, pp. 273–285, 2016.
- [19] B. Y. Oh, J. Cho, H. K. Hong et al., “Exome and transcriptome sequencing identifies loss of PDLIM2 in metastatic colorectal cancers,” *Cancer Management and Research*, vol. 9, pp. 581–589, 2017.
- [20] P. Costa-Reis and K. E. Sullivan, “Monogenic lupus: it’s all new!,” *Current Opinion in Immunology*, vol. 49, pp. 87–95, 2017.
- [21] A. Bodaño, J. Amarello, A. González, J. J. Gómez-Reino, and C. Conde, “Novel DNASE1 mutations related to systemic lupus erythematosus,” *Arthritis and Rheumatism*, vol. 50, no. 12, pp. 4070–4071, 2004.
- [22] M. Malecki, C. LaVanne, D. Alhambra, C. Dodivenaka, S. Nagel, and R. Malecki, “Safeguarding stem cell-based regenerative therapy against iatrogenic cancerogenesis: transgenic expression of DNASE1, DNASE1L3, DNASE2, DFFB controlled by POLA1 promoter in proliferating and directed differentiation resisting human autologous pluripotent induced stem cells leads to their death,” *Journal of Stem Cell Research and Therapy*, vol. Suppl 9, 2013.
- [23] M. Malecki, J. Dahlke, M. Haig, L. Wohlwend, and R. Malecki, “Eradication of human ovarian cancer cells by transgenic expression of recombinant DNASE1, DNASE1L3, DNASE2, and DFFB controlled by EGFR promoter: novel strategy for targeted therapy of cancer,” *Journal of Genetic Syndromes & Gene Therapy*, vol. 4, p. 152, 2013.
- [24] P. A. Coulombe and M. B. Omary, ““Hard” and “soft” principles defining the structure, function and regulation of keratin intermediate filaments,” *Current Opinion in Cell Biology*, vol. 14, no. 1, pp. 110–122, 2002.
- [25] V. Karantzis, “Keratins in health and cancer: more than mere epithelial cell markers,” *Oncogene*, vol. 30, no. 2, pp. 127–138, 2011.
- [26] Y. Xie, L. Diao, L. Zhang, C. Liu, Z. Xu, and S. Liu, “A miR-SNP of the KRT81 gene is associated with the prognosis of non-Hodgkin’s lymphoma,” *Gene*, vol. 539, no. 2, pp. 198–202, 2014.
- [27] M. Campayo, A. Navarro, N. Viñolas et al., “A dual role for KRT81: a miR-SNP associated with recurrence in non-small-cell lung cancer and a novel marker of squamous cell lung carcinoma,” *PLoS One*, vol. 6, no. 7, Article ID e22509, 2011.
- [28] G. Song, J. Xu, L. He et al., “Systematic profiling identifies PDLIM2 as a novel prognostic predictor for oesophageal squamous cell carcinoma (ESCC),” *Journal of Cellular and Molecular Medicine*, vol. 23, no. 8, pp. 5751–5761, 2019.
- [29] S. Alaei, B. Sadeghi, A. Najafi, and A. Masoudi-Nejad, “LncRNA and mRNA integration network reconstruction reveals novel key regulators in esophageal squamous-cell carcinoma,” *Genomics*, vol. 111, no. 1, pp. 76–89, 2019.
- [30] X.-H. Zhan, J.-W. Jiao, H.-F. Zhang et al., “A three-gene signature from protein-protein interaction network of LOXL2 - and actin-related proteins for esophageal squamous cell carcinoma prognosis,” *Cancer Medicine*, vol. 6, no. 7, pp. 1707–1719, 2017.
- [31] L. L. Sun, J. Y. Wu, Z. Y. Wu et al., “A three-gene signature and clinical outcome in esophageal squamous cell carcinoma,” *International Journal of Cancer*, vol. 136, pp. E569–E577, 2015.
- [32] X. Guo, Y. Wang, H. Zhang et al., “Identification of the prognostic value of immune-related genes in esophageal cancer,” *Frontiers in Genetics*, vol. 11, p. 989, 2020.
- [33] Z. Fei, R. Xie, Z. Chen et al., “Establishment of a novel risk score system of immune genes associated with prognosis in esophageal carcinoma,” *Frontiers in Oncology*, vol. 11, p. 625271, 2021.
- [34] Y. Yu, Z. Li, C. Huang et al., “Integrated analysis of genomic and transcriptomic profiles identified a prognostic immunohistochemistry panel for esophageal squamous cell cancer,” *Cancer Medicine*, vol. 9, no. 2, pp. 575–585, 2020.