

Six-gene signature for predicting survival in patients with head and neck squamous cell carcinoma

Juncheng Wang^{1,*}, Xun Chen^{2,*}, Yuxi Tian³, Gangcai Zhu⁴, Yuexiang Qin⁵, Xuan Chen⁶, Leiming Pi⁷, Ming Wei¹, Guancheng Liu⁸, Zhexuan Li¹, Changhan Chen¹, Yunxia Lv⁹, Gengming Cai¹⁰

¹Department of Otolaryngology, Head and Neck Surgery, Xiangya Hospital, Central South University, Changsha 410008, Hunan, People's Republic of China

²Department of Oral and Maxillofacial Surgery, First Affiliated Hospital of Quanzhou, Fujian Medical University, Quanzhou 362000, Fujian, People's Republic of China

³Department of Oncology, Xiangya Hospital, Central South University, Changsha 410008, Hunan, People's Republic of China

⁴Department of Otolaryngology, Head and Neck Surgery, The Second Xiangya Hospital, Central South University, Changsha 410011, Hunan, People's Republic of China

⁵Department of Health Management, The Third Xiangya Hospital, Central South University, Changsha 410011, Hunan, People's Republic of China

⁶Department of Stomatology, Changzheng Hospital, Second Military Medical University, Shanghai 200003, People's Republic of China

⁷Department of Otolaryngology, Head and Neck Surgery, HeYuan People's Hospital, Jinan University, He Yuan 517000, Guang Dong, People's Republic of China

⁸Department of Otolaryngology, Head and Neck Surgery, Affiliated Hospital of Guilin University, Guilin 541000, Guangxi, People's Republic of China

⁹Department of Thyroid Surgery, The Second Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi, People's Republic of China

¹⁰Department of Otolaryngology, Head and Neck Surgery, First Affiliated Hospital of Quanzhou, Fujian Medical University, Quanzhou 362000, Fujian, People's Republic of China

*Equal contribution

Correspondence to: Gengming Cai, Yunxia Lv; **email:** cgmkgx@hotmail.com, 83394045@qq.com

Keywords: bioinformatics, CNV, prognostic markers, TCGA

Received: September 16, 2019

Accepted: December 24, 2019

Published: January 12, 2020

Copyright: Wang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ABSTRACT

The prognosis of head and neck squamous cell carcinoma (HNSCC) patients remains poor. High-throughput sequencing data have laid a solid foundation for identifying genes related to cancer prognosis, but a gene marker is needed to predict clinical outcomes in HNSCC. In our study, we downloaded RNA Seq, single nucleotide polymorphism, copy number variation, and clinical follow-up data from TCGA. The samples were randomly divided into training and test. In the training set, we screened genes and used random forests for feature selection. Gene-related prognostic models were established and validated in a test set and GEO verification set. Six genes (*PEX11A*, *NLRP2*, *SERPINE1*, *UPK*, *CTTN*, *D2HGDH*) were ultimately obtained through random forest feature selection. Cox regression analysis confirmed the 6-gene signature is an independent prognostic factor in HNSCC patients. This signature effectively stratified samples in the training, test, and external verification sets ($P < 0.01$). The 5-year survival AUC in the training and verification sets was greater than 0.74. Thus, we have constructed a 6-gene signature as a new prognostic marker for predicting survival of HNSCC patients.

INTRODUCTION

Head and neck cancer is a highly heterogeneous malignant tumor, which can originate from various anatomical sites in the upper airway and digestive tract, including the mouth, larynx and pharynx. Most cases of head and neck cancer are squamous cell carcinoma (HNSCC), which accounts for about 4% of all new cancer diagnoses in the United States. Worldwide, there are about 600,000 new head and neck cancer patients each year [1], and the 5-year survival rate is only 40-50% [2]. The main reason for this high mortality is the high rate of diagnosis of advanced cancers, as the survival rate among advanced cancer patients is only 34.9% [3]. There is thus an urgent need for markers to help clinicians make accurate early HNSCC diagnoses, predict clinical outcomes, and provide reference for individualized medicine.

Many studies have been carried out in an effort to find predictive biomarkers to establish guidelines for the long-term prognosis of HNSCC patients. These biomarkers can be divided into two categories: 1) single molecules such as squamous cell carcinoma antigen (SCC-A), human papilloma virus (HPV), or any of the other new markers currently being studied; and 2) gene markers identified by analyzing high-throughput gene expression profiles and constructed using from several to dozens of prognostic genes. Several systems biology methods are currently being used to identify gene biomarkers related to the HNSCC prognosis and to characterize those genes [4–6]. For example, Tian et al. used weighted gene correlation network analysis and least absolute shrinkage and selection operator Cox regression to identify a 6-lncRNA signature within a gene expression profile. De et al. used gene expression meta-analysis to identify a 172-gene signature. And Zhao et al. used protein-protein interaction network analysis and Cox regression analysis to identify a 3-gene signature. All three authors tested their genetic signature using independent external data sets. However, the AUC for Zhao et al's genetic signature is not high (AUC=0.6), which means that identifying robust lncRNA signatures is still a challenge, and more investigation will be required to verify signatures. In other words, there is still an important need to identify new gene signals related to the prognosis of HNSCC through bioinformatics analysis of their biological functions.

To effectively identify a reliable gene signature associated with prognosis in HNSCC, we proposed a systematic pipeline to identify HNSCC-related gene markers. This approach enabled us to identify a 6-gene signature that can be used to effectively predict prognostic risk in HNSCC patients and provide a basis

for better understanding of the molecular mechanisms underlying the prognoses of HNSCC patients.

RESULTS

Identification of gene sets associated with total patient survival

For The Cancer Genome Atlas (TCGA) training set samples, we used single factor regression analysis to establish the relationship between overall survival (OS) and gene expression. We identified 425 single factor Cox regression logrank P values less than 0.01. Of those, 141 genes were associated a hazard ratio (HR) > 1, and 284 with a HR < 1. The 20 genes with the highest HRs are listed in Table 1.

Recognition of gene sets for genome variation

Using copy number variation data in TCGA, we used GISTIC 2.0 to identify genes exhibiting significant amplification or deletion. Significantly amplified fragments within the genomes, including *EGFR* at 7p11.2 (q value = 2.28E-43), *FGF1* at 8p11.23 (q value = 3.94E-14), and *ERBB2* at 17q12 (q value = 0.0050541) (Figure 1A). In total, 247 genes were amplified (Supplementary Table 1). Genes that were significantly missing from the genome included *CDKN2A* at 9p21.3 (q value = 5.28E-149), *CDK5* at 7q36.1 (q value = 1.87E-06), and *PTEN* at 10q23.31 (q = 0.0032849) (Figure 1B). A total of 901 genes were identified as missing from the genome.

Using TCGA mutation annotation data with Mutsig2, we identified genes with significant mutations. A total of 302 genes with significant mutation frequencies were detected (Supplementary Table 2). The most significant types mutations in the Top 50 genes were synonymous mutations, missense mutations, frame insertions or deletions, frame movement, nonsense mutations, distribution of shear sites, and other non-synonymous mutations (Figure 1C). It was clear that there were differences in the mutations to different genes, including *B2M*, *CDKN2A*, *PTEN*, *TP53* and *FBXW7*. A common feature, however, is that mutation of these genes is reported to be closely related to the occurrence and development of tumors [7–10].

Functional analysis of genome variant genes

To analyze the function of genomic variant genes, we integrated 1321 amplified or deleted genes and significantly mutated genes identified based on copy number variation. Gene Ontology (GO) biological process and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analyses were

Table 1. List of the most relevant 20 genes.

ENSG ID	SYMBOL	HR	coefficient	z score	P
ENSG00000127084	<i>TLL2</i>	0.630	-0.462	-4.562	5.07E-06
ENSG00000254656	<i>FAM69A</i>	1.360	0.308	4.202	2.64E-05
ENSG00000041515	<i>HIST2H3PS2</i>	1.337	0.290	4.200	2.67E-05
ENSG00000126353	<i>TMCO1</i>	0.647	-0.435	-3.969	7.22E-05
ENSG00000115085	<i>MAP2K7</i>	0.649	-0.432	-3.916	9.01E-05
ENSG00000170482	<i>SPINK1</i>	0.538	-0.620	-3.896	9.80E-05
ENSG00000162545	<i>GLYCTK</i>	1.503	0.407	3.880	0.0001
ENSG00000125910	<i>ERRFI1</i>	0.638	-0.449	-3.838	0.0001
ENSG00000174652	<i>MSANTD3</i>	0.674	-0.395	-3.795	0.0001
ENSG00000197540	<i>BTLA</i>	0.655	-0.423	-3.722	0.0001
ENSG00000184903	<i>ATP6V0E1</i>	1.392	0.331	3.719	0.0001
ENSG00000132465	<i>CALML5</i>	0.692	-0.369	-3.698	0.0002
ENSG00000089199	<i>MORF4L2</i>	1.382	0.324	3.683	0.0002
ENSG00000189319	<i>ZZEF1</i>	0.696	-0.362	-3.674	0.0002
ENSG00000198198	<i>TYK2</i>	0.692	-0.369	-3.609	0.0003
ENSG00000153531	<i>EOMES</i>	1.318	0.276	3.553	0.0003
ENSG00000159176	<i>PTGR1</i>	1.359	0.307	3.552	0.0003
ENSG00000127241	<i>GPR171</i>	0.598	-0.515	-3.537	0.0004
ENSG00000197992	<i>LIMD2</i>	0.617	-0.483	-3.537	0.0004
ENSG00000182866	<i>DCP2</i>	0.700	-0.357	-3.494	0.0004

then performed. The results of the KEGG enrichment analysis revealed that Pathways in cancer, HPV infection, PI3K-Akt signaling pathway, Human T-cell leukemia virus 1 infection, Human cytomegalovirus infection, and numerous other related KEGG biological pathways are important to the development of cancer (Figure 2A). GO terms such as developmental process, positive regulation of cellular process, cell differentiation, and regulation of localization were mainly enriched in the “biological process” category (Figure 2B). These terms were also closely related to the occurrence and development of cancer; that is, these genes exhibiting genomic variation are closely related to cancer.

Identification of a 6-gene signature for head and neck cancer survival

To identify a gene signature, we integrated genomic variant and prognosis-related genes, and then selected the intersection of the groups as candidate genes, which yielded 36 genes. We then used random forests for feature selection. The relationship between error rate and number of taxonomic trees was used to reveal genes with relative importance greater than 0.4 as the final signature (Figure 3A). Ultimately, we identified 6 genes

(Table 2). The important order of the out-of-bag scores for the 6 genes is displayed in Figure 3B. A 6-gene signature was established by multivariate COX regression analysis. The Equation 1 is as follows:

$$Risk_a = -0.3929517 * D2HGDH + 0.3627132 * PEX11A + 0.3038125 * NLRP2 + 0.275704 * SERPINE1 + 0.188539 * UPK2 + 0.1112888 * CTTN \quad (1)$$

The risk score of each sample was calculated, and the samples were grouped according to the median risk score (cutoff = -0.0236503). The prognoses of the high-risk and low-risk groups significantly differed (Figure 3C). The average 1-, 3-, and 5-year AUC for the 6-gene signature was 0.75 (Figure 3D). High expression of *PEX11A*, *NLRP2*, *SERPINE1*, *UPK2* and *CTTN* was associated with high risk, while high expression of *D2HGDH* was associated with low risk and was a protective factor.

Verification of the robustness of the 6-gene signature model

To verify the robustness of the 6-gene signature model, we first calculated a risk score for each sample in the

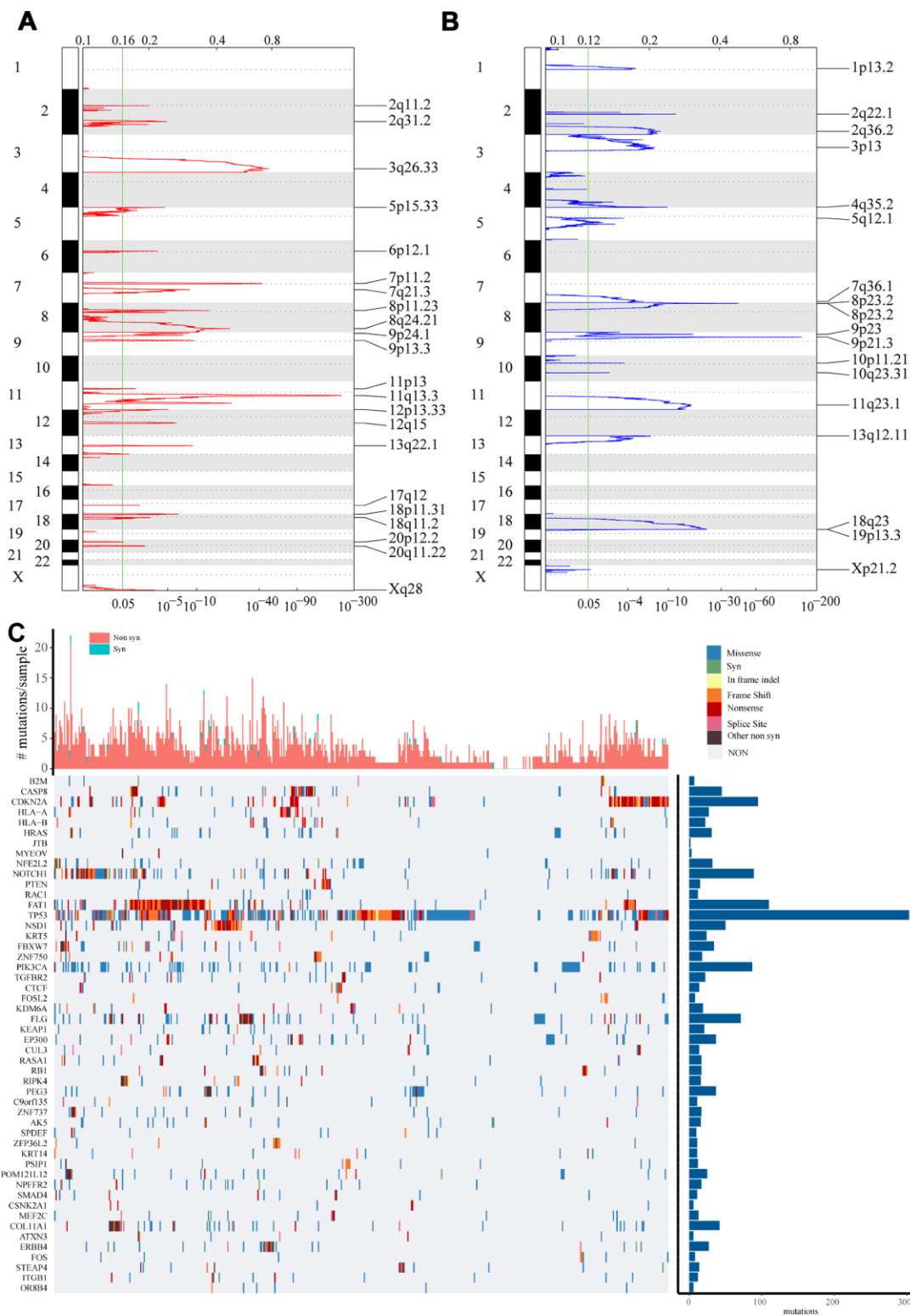


Figure 1. Identification of genes with significant amplification or deletion. (A and B) The mRNA located in the focal CNA peaks are HNSCC-related. False-discovery rates (q values) and scores from GISTIC 2.0 for alterations (x-axis) are plotted against genome positions (y-axis). Dotted lines indicate the centromeres. Amplifications (A) are shown in red, deletion (B) in blue. The green line represents 0.25 q value cut-off point that determines significance. (C) Top 50 genes with the most significant mutations. The bar chart above shows the total number of synonymous and non-synonymous mutations in each patient's top 50 genes. The bar chart on the right shows the number of samples in which the 50 genes were mutated in all samples. The different colors in the thermogram indicate the type of mutation; gray indicates no mutation.

test set. Based on the threshold for the training set, the samples were divided into two groups with significantly different prognoses (Figure 4A). The relationship between the expression of the 6 genes and the risk score was also consistent with the training set (Figure 4C). We similarly applied the model to all TCGA tumor samples and found that the low risk group fared significantly better than the high-risk group (Figure 4B). The relationship between the expression of these 6 genes and the risk score was also consistent with the training set (Figure 4D). Overall, the model effectively provided prognostic classifications with the TCGA datasets.

To verify the classification performance of the 6-gene signature model with different data platforms, we used GEO platform data as external datasets, used the model to calculate a risk score for each sample, and used the cutoff for the training set to divide the samples into high-risk and low-risk groups. The prognosis of the low-risk group was significantly better than that of the high-risk group (Figure 5A). ROC analysis showed that the 5-year AUC was up to 0.74, compared with the training set. As in Figure 5B, the data in Figure 5C show that the relationship between the expression of the 6 genes and risk score is also consistent with the training set. Thus, the 6-gene signature model we

selected was prognostic with both internal and external datasets.

Clinical independence of the 6-gene signature model

To assess the independence of the 6-gene signature model in clinical application, we used single-factor and multi-factor COX regression to analyze HRs, 95% CIs, and P values from TCGA training set, TCGA test set, and the GSE65858 data. We systematically analyzed the clinical information from the patients as recorded in TCGA and GSE65858 datasets, including their age, sex, disease stage, pathological TNM stage, and tumor stage, as well as our 6-gene signature (Table 3). In TCGA test set, single factor COX regression analysis revealed that in the high-risk group, pathologic T3, T4, M1, and N2 all significantly correlated with survival. However, the corresponding multifactor COX regression analysis found that only high-risk group (HR = 2.17, 95% CI = 1.04-4.53, P = 0.037979), pathologic T4 (HR = 7.58, 95% CI = 1.90-30.22, P = 0.004054) and pathologic N2 (HR = 5.01, 95% CI = 1.97-12.75, P = 0.000703) had clinical independence. In TCGA test set, single factor COX regression analysis revealed that in the high-risk group, pathologic T3 and N2 correlated significantly with survival, but the corresponding multivariate COX regression analysis found that no factor had clinical

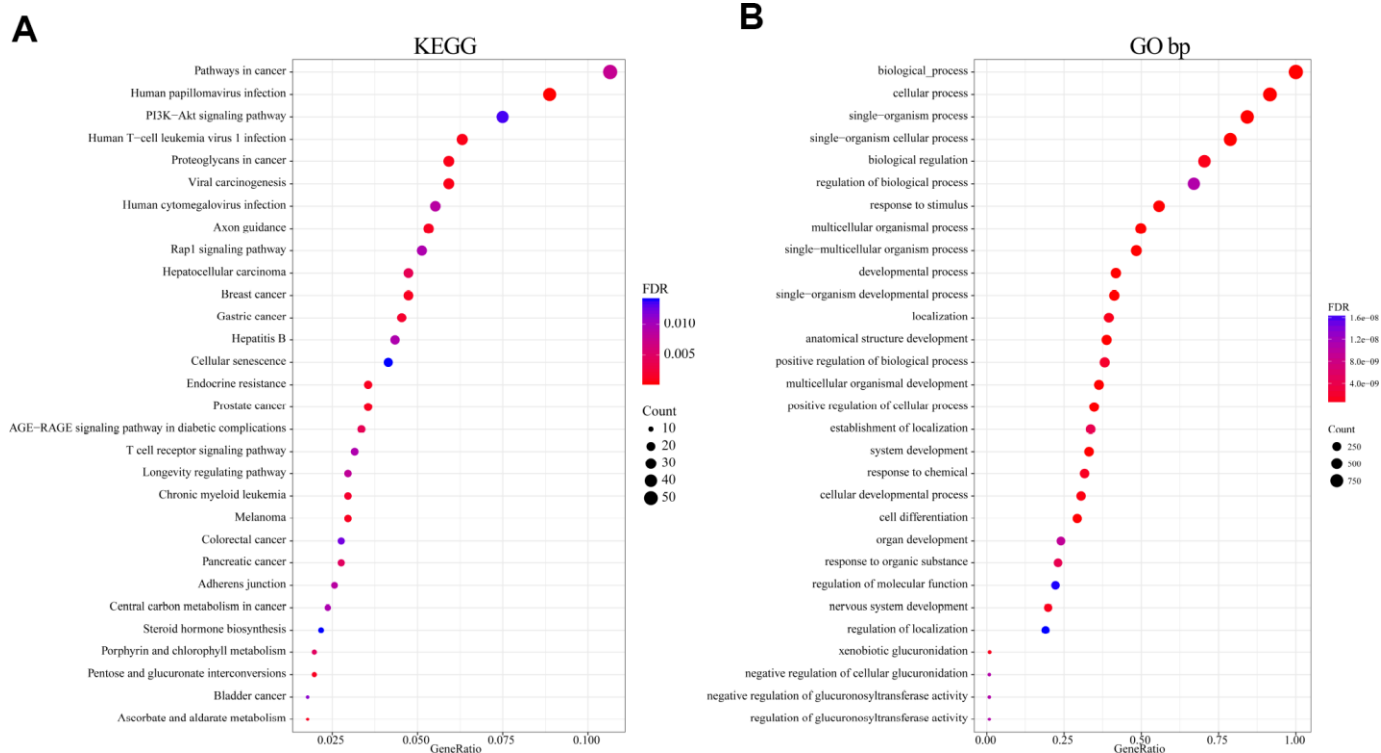


Figure 2. Functional enrichment analysis of 1321 genome variant genes. (A) Enriched KEGG biological pathways. **(B)** Enriched GO terms in the “biological process” category. Different colors indicate different significances, while different sizes indicate the number of genes.

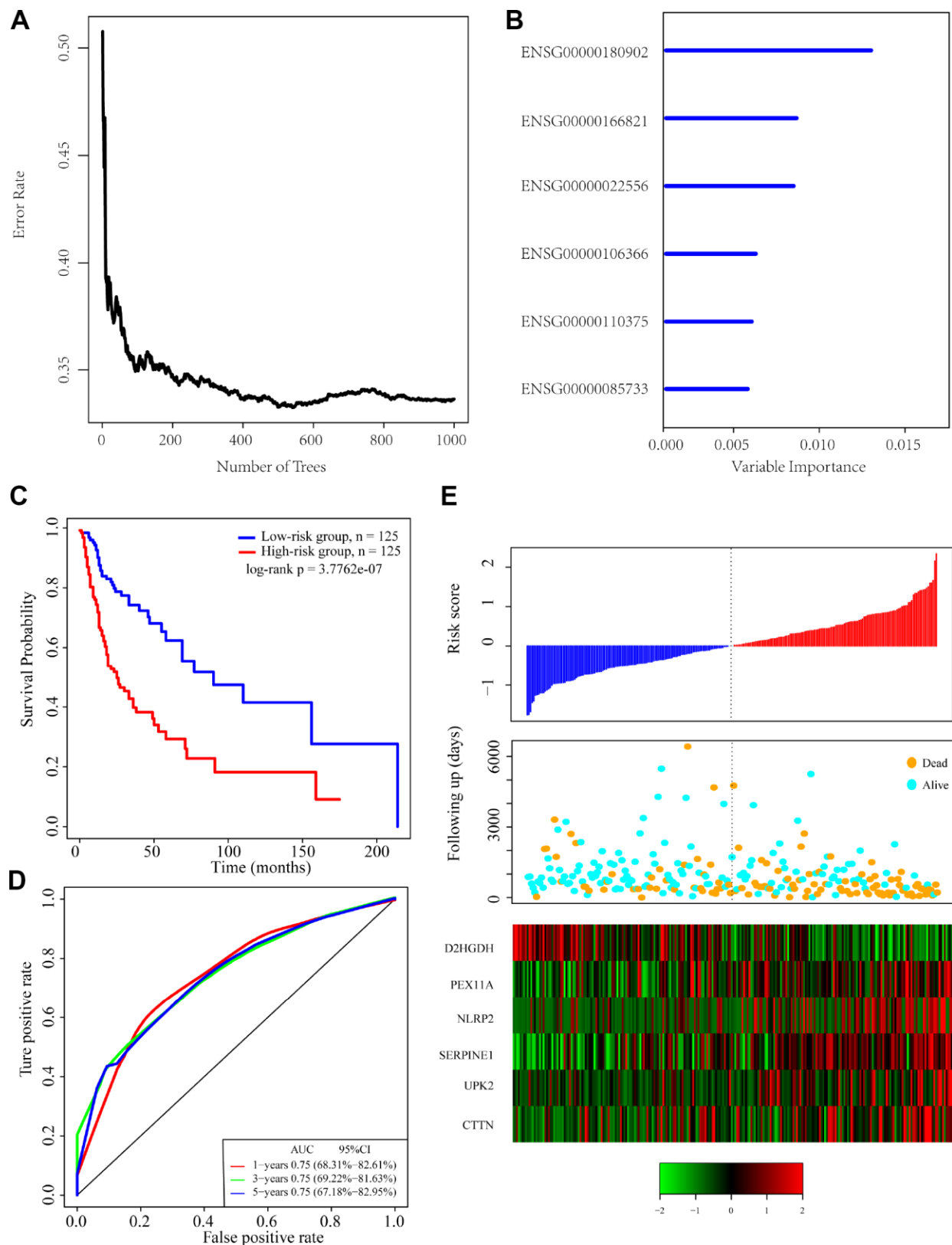


Figure 3. Identification of genomic variant genes and prognosis-related genes in head and neck cancer. (A) Relationship between the error rate and number of classification trees. (B) Importance the sequencing of 6 out-of-bag genes. (C) Distribution of the 6-gene signature in Kaplan-Meier survival curve for the TCGA training set. (D) ROC curve and AUC for the 6-gene signature classification. (E) TCGA training focused on risk score, survival time, survival status, and expression of the 6-gene signature.

Table 2. Six genes significantly associated with OS in the training set patients.

Ensemble Gene ID	Symbol	HR	Z-score	P	Importance	Relative importance
ENSG00000180902	<i>D2HGDH</i>	0.77	-2.61	8.88E-03	0.012	1
ENSG00000166821	<i>PEX11A</i>	1.33	2.98	2.87E-03	0.0076	0.6359
ENSG00000022556	<i>NLRP2</i>	1.32	2.96	3.03E-03	0.0074	0.6214
ENSG00000106366	<i>SERPINE1</i>	1.43	3.33	8.47E-04	0.0052	0.4369
ENSG00000110375	<i>UPK2</i>	1.29	2.98	2.84E-03	0.005	0.4175
ENSG00000085733	<i>CTTN</i>	1.27	2.69	7.08E-03	0.0048	0.4

independence; the HR in the high-risk group was 1.45, 95% CI = 0.67-3.12, P = 0.3389. In the GSE65858 dataset, univariate COX regression analysis revealed that the in the high-risk group, age, pathologic T3, N3, and M1 correlated significantly with survival. The corresponding multivariate COX regression analysis found that the high-risk group (HR = 1.83, 95% CI = 1.16-2.87, P = 0.0083) and age (HR = 1.03, 95% CI = 1.01-1.05, P = 0.00313) had clinical independence. These results show that our model 6-gene signature is a prognostic index independent of other clinical factors and exhibits independent predictive performance upon clinical application.

GSEA analysis of pathway differences enriched in the high-risk and low-risk groups

GSEA was used in TCGA training to analyze the pathways significantly enriched in the high-risk and low-risk groups. Twenty enriched pathways were detected (Supplementary Table 3), including focal adhesion, TGF-β signaling pathway, WNT signaling pathway, and ERBB signaling pathway, all of which are closely related to tumor occurrence, development and metastasis. Notably, these pathways are significantly enriched in the high-risk samples (Figure 6).

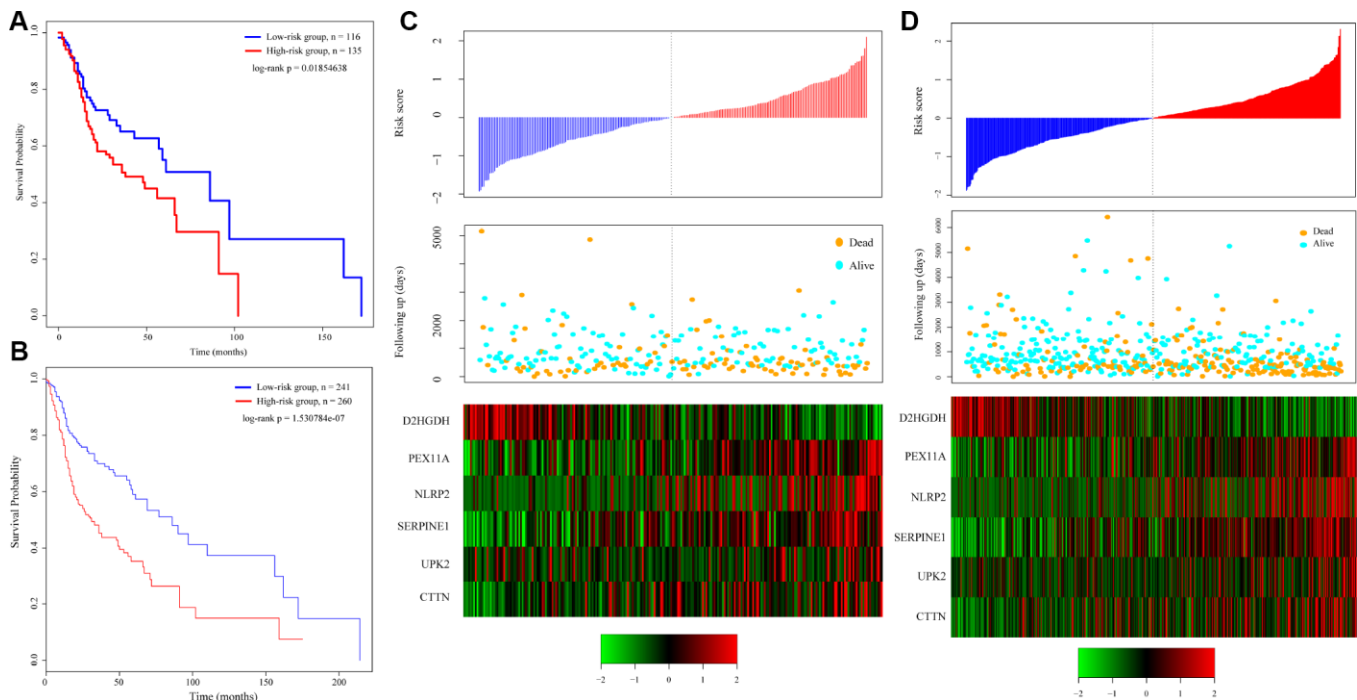


Figure 4. Relation between the 6-gene signature and cancer risk. (A) Kaplan-Meier curve for the test set sample. **(B)** Kaplan-Meier curve in all TCGA tumor samples. **(C)** Relationship between expression of the 6-gene signature and risk scores in test set samples. **(D)** Relationship between expression of the 6-gene signature and the risk score in all TCGA samples.

DISCUSSION

In terms of prognosis, head and neck cancer is a highly heterogeneous disease in that survival times vary substantially among patients with similar TNM stages. With the diagnosis and treatment of head and neck cancers at earlier stages, traditional clinicopathological indicators such as tumor size, vascular invasion, portal vein thrombus and TNM stage have proven inadequate for predicting individual outcomes, especially risk stratification, as no one-size-fits-all treatment strategy appears to be effective [11, 12]. Consequently, screening prognostic molecular markers that adequately reflect the biological characteristics of tumors would be of great significance for individualized prevention and treatment of head and neck cancer patients. In the

present study, we analyzed the expression profiles of 771 head and neck cancer samples from TCGA and the GEO and identified 6 genes robustly associated with OS. This signature is independent of other clinical factors.

Gene signatures are currently being used in clinical practice. Two examples are Oncotype DX [13–15], which provides a breast cancer recurrence score based on expression of 21 genes, and ColoPrint, which provides a colon cancer recurrence score based on expression of 18 genes [16–18]. Results obtained with these assays have shown that screening new prognostic cancer markers based on gene expression profiles is a promising high-throughput molecular identification method. In that regard, Tian et al. [6] identified a 6-gene

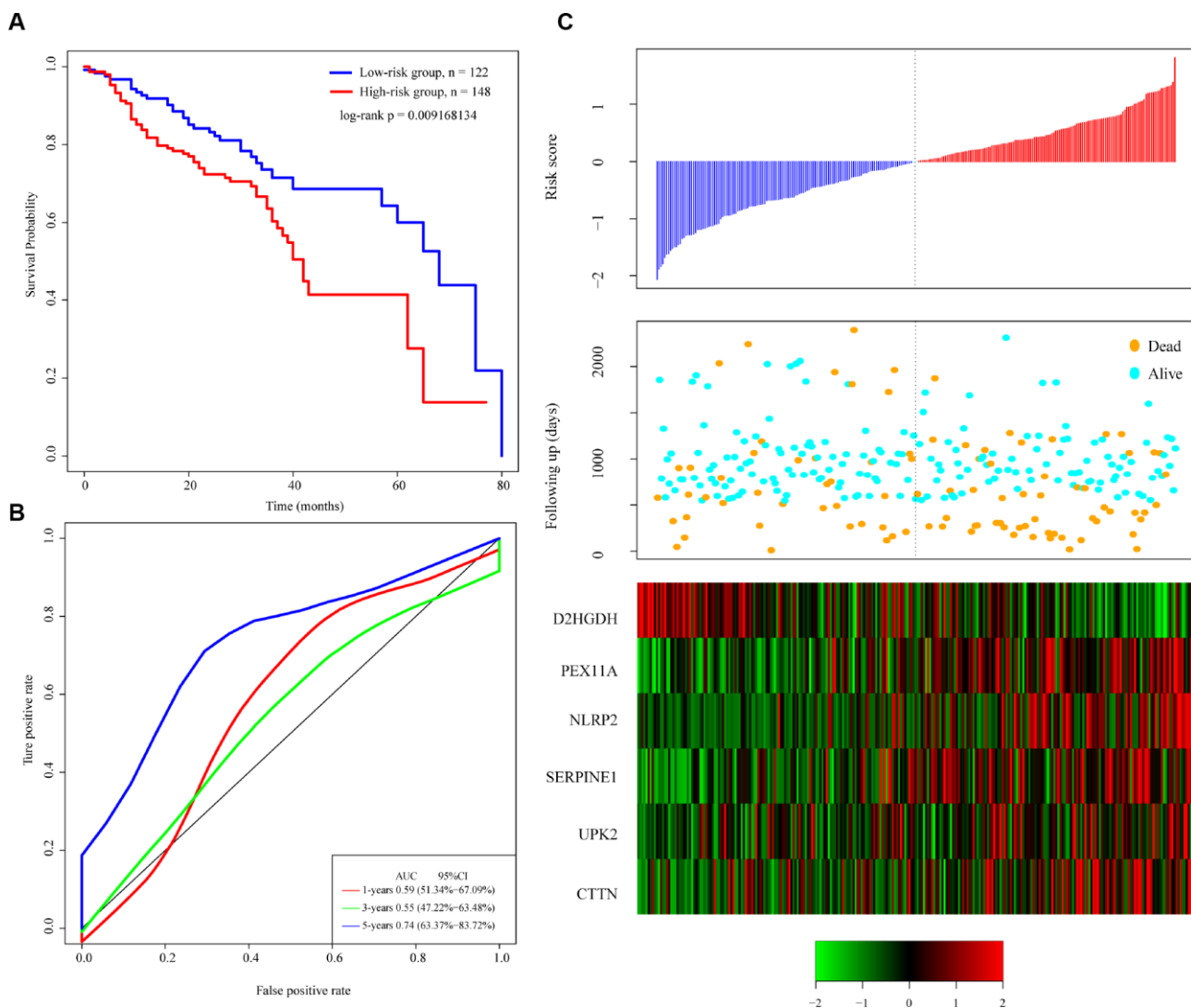


Figure 5. Performance of the 6-gene signature model with GEO data. (A) Kaplan-Meier survival curve distribution of 6-gene signature for the GSE65858 dataset. (B) ROC curve and AUC for the 6-gene signature classification. (C) Risk score, survival time, survival status, and expression of the 6-gene signature in the GSE65858 dataset.

Table 3. Univariate and multivariate COX regression analyses of clinical factors and independence associated with prognosis.

Variables	Univariate analysis			Multivariable analysis		
	HR	95%CI of HR	P	HR	95%CI of HR	P
TCGA training datasets						
6-gene risk score						
Low risk group	1(reference)			1(reference)		
High risk group	2.66	1.79-3.92	1.060E-06	2.18	1.04-4.53	0.038
Age	1.03	1.01-1.04	8.550E-04	0.99	0.95-1.02	0.604231
Gender female	1(reference)			1(reference)		
Gender male	0.73	0.49- 1.07	0.11	0.50	0.23-1.09	0.081554
Grade 1	1(reference)			1(reference)		
Grade 2	1.78	0.95-3.32	0.07	0.91	0.25-3.18	0.88025
Grade 3 / 4	1.47	0.75-2.87	0.26	1.94	0.53-7.04	0.315185
Pathologic T 1/ T 2	1(reference)			1(reference)		
Pathologic T 3	2.26	1.28-3.95	4.53E-03	2.07	0.51-8.37	0.309
Pathologic T 4	2.46	1.50-4.00	3.19E-04	7.59	1.90-30.22	0.004
Pathologic N 0	1(reference)			1(reference)		
Pathologic N 1	0.89	0.37-2.08	0.782	0.76	0.19-3.01	0.694
Pathologic N 2	2.29	1.40-3.72	0.001	5.02	1.97-12.75	0.001
Pathologic M 0	1(reference)			1(reference)		
Pathologic M 1/ M X	1.31	0.66-2.55	0.433	1.16	0.43-3.08	0.76
Tumor stage I	1(reference)			1(reference)		
Tumor stage II	1.08	0.23-4.90	0.918	0.55	0.03-9.85	0.686335
Tumor stage III	1.40	0.30-6.42	0.667	0.77	0.05-10.60	0.845959
Tumor stage IV	2.75	0.67-11.22	0.158	0.27	0.02-3.63	0.32
Validation cohort, TCGA test datasets and GSE65858						
TCGA test datasets						
6-gene risk score						
Low risk group	1(reference)			1(reference)		
High risk group	1.62	1.08- 2.42	0.020	1.45	0.67-3.12	0.339
Age	1.01	0.98-1.02	0.444	1.00	0.96-1.03	0.993
Gender female	1(reference)			1(reference)		
Gender male	0.84	0.55-1.28	0.420	0.45	0.19-1.01	0.054
Grade 1	1(reference)			1(reference)		
Grade 2	1.86	0.92-3.77	0.08	0.90	0.28-2.88	0.865
Grade 3	1.57	0.73-3.36	0.24	1.54	0.42-5.52	0.508
Pathologic T 1/ T 2	1(reference)			1(reference)		
Pathologic T 3	1.84	1.09-3.11	0.022	1.76	0.47-6.54	0.399
Pathologic T 4	1.36	0.84-2.21	0.213	1.20	0.35-4.07	0.770
Pathologic N 0	1(reference)			1(reference)		
Pathologic N 1	1.14	0.57-2.25	0.706	1.74	0.50-6	0.379
Pathologic N 2	2.49	1.50-4.12	0.000	1.87	0.65-5.32	0.239
Pathologic N 3	2.90	0.87-9.6	0.082	1.86	0.26-12.82	0.529
Pathologic M 0	1(reference)			1(reference)		

Pathologic M 1	1.05	0.49-2.21	0.905	1.34	0.54-3.28	0.524
Tumor stage I	1(reference)					
Tumor stage II	2.94	0.34-0.66	0.157	0.55	0.03-9.20	0.677
Tumor stage III	3.04	0.32-0.70	0.137	1.00	0.08-11.42	0.997
Tumor stage IV	4.03	0.24-0.98	0.053	1.46	0.12-17.69	0.765
GSE65858						
6-gene risk score						
Low risk group	1(reference)			1(reference)		
High risk group	1.75	1.13-2.67	0.011	1.83	1.16-2.87	0.008
Age	1.03	1.00-1.04	0.01	1.03	1.01-1.05	0.00
Gender female	1(reference)			1(reference)		
Gender male	1.05	0.61-1.77	0.87	1.02	0.59-1.75	0.94
Pathologic T 1	1(reference)			1(reference)		
Pathologic T 2	0.49	0.21-1.15	0.10	0.53	0.15-1.78	0.31
Pathologic T 3	1.73	0.81-3.67	0.15	1.61	0.55-4.69	0.38
Pathologic T 4	1.89	0.91-3.87	0.08	1.22	0.42-3.53	0.71
Pathologic N 0	1(reference)			1(reference)		
Pathologic N 1	0.36	0.12-1.02	0.06	0.34	0.10-1.08	0.07
Pathologic N 2	1.60	0.99-2.56	0.05	1.03	0.51-2.05	0.94
Pathologic N 3	2.93	1.34-6.42	0.01	1.27	0.41-3.84	0.67
Pathologic M 0	1(reference)			1(reference)		
Pathologic M 1/ M X	3.71	1.63-8.4	0.00	2.18	0.70-6.79	0.18
Tumor stage I	1(reference)			1(reference)		
Tumor stage II	0.39	0.11-1.33	0.13	0.60	0.10-3.40	0.56
Tumor stage III	0.46	0.14-1.45	0.19	0.66	0.13-3.31	0.61
Tumor stage IV	1.49	0.60-3.70	0.39	1.14	0.24-5.20	0.87

signature, but the verification set AUC was only about 0.65, and Zhao et al. [4] identified a 3-gene signature, but the AUC was only about 0.6. In addition, De et al. [5] identified a 172-gene signature through meta-analysis of gene expression. Although the AUC is high, the large number of genes that needs to be detected makes this analysis impractical for clinical use. By contrast, our 6-gene signature has a high AUC using only 6 genes, which makes it conducive to clinical application.

The six genes in our signature include *PEX11A*, *NLRP2*, *SERPINE1*, *UPK*, and *CTTN* as risk factors, and *D2HGDH* as a protective factor. It has been reported that *NLRP2* can be used as a marker of leukemia [19] and has an important impact on the prognosis after stem cell transplantation [20]. *SERPINE1* is closely related to prognosis in ovarian cancer, gastric cancer, thyroid cancer and other tumors [21–25]. *CTTN* is closely related to prognosis in head and neck cancer, esophageal cancer, thyroid cancer, and glioma, among others [26–29]. *D2HGDH* is a marker of colorectal cancer, glioma, and prostate cancer [30–33]. *PEX11A*

and *UPK* have not been previously reported to be related to cancer. Ours is the first study to suggest that they can be used as new prognostic markers of head and neck cancer. At the same time, our GSEA results show that the 6-gene signature enrichment significantly correlates with pathways and biological processes associated with the occurrence and development of HNSCC. This indicates that our model has potential clinical application value and could provide a potential target for diagnosis and for development of new targeted therapies that include, for example, novel alkylating agents [34, 35].

Although we have identified potential candidate genes affecting tumor prognosis using bioinformatics technology with large samples, our study has limitations. First, the sample lacks some clinical follow-up information, so we did not consider factors such as the presence of other health conditions to differentiate prognostic biomarkers. Second, the results obtained using bioinformatics analysis alone are insufficient and need to be confirmed through experimental verification.

Therefore, further genetic and experimental studies with larger samples and experimental validation are needed.

CONCLUSIONS

In our study, we developed a 6-gene signature prognostic stratification system, which has good AUC

values in both the training set and validation set, and is independent of other clinical features. Compared to clinical features, gene classifiers can improve survival risk prediction. We therefore recommend using this classifier as a molecular diagnostic test to assess prognostic risk in patients with head and neck cancer.

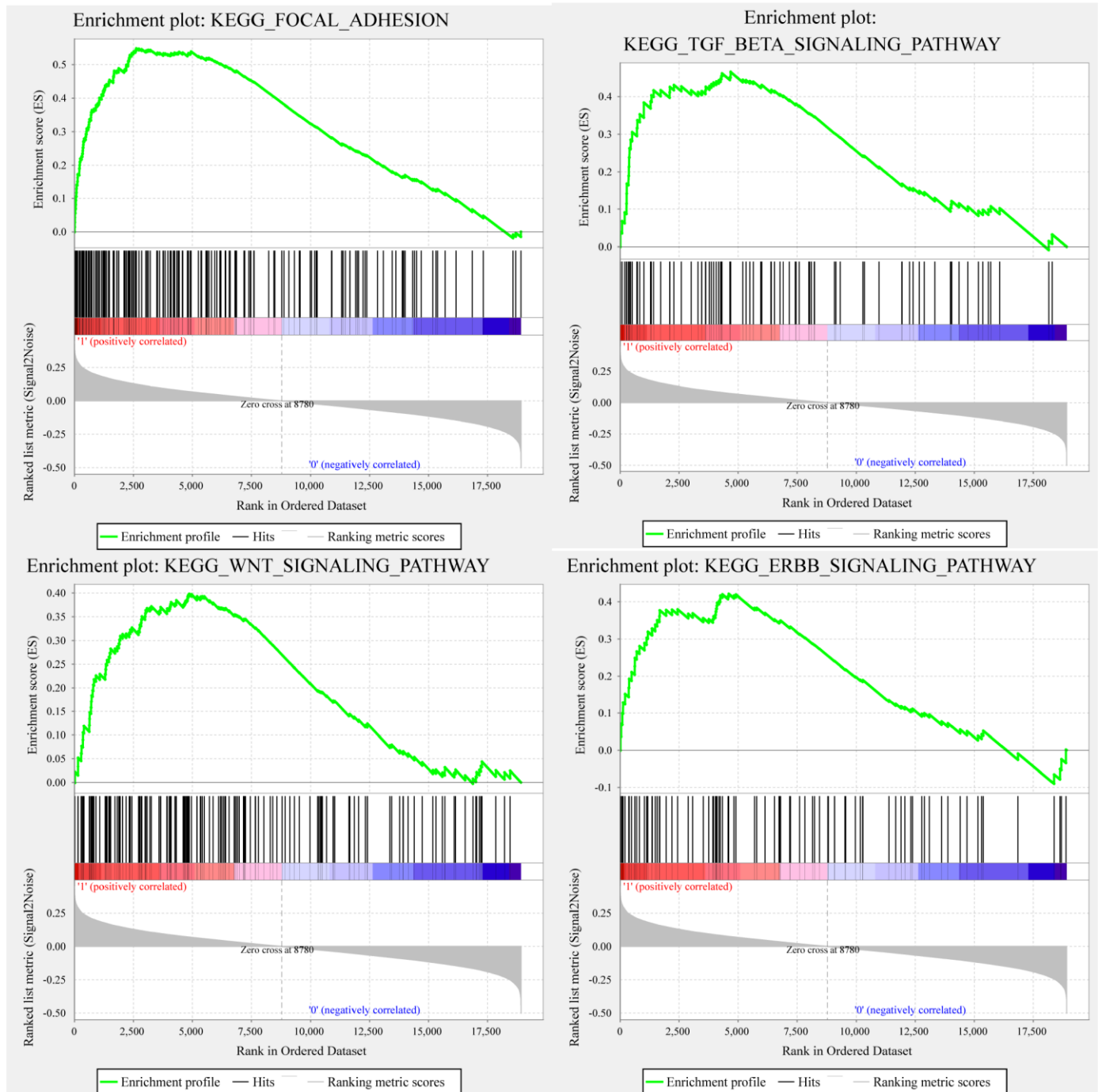


Figure 6. GSEA showing four pathways enriched in the high-risk group. GSEA enrichment results for focal adhesion, TGF- β signaling pathway, WNT signaling pathway, and ERBB signaling pathway.

Table 4. Clinical information statistics for three datasets.

Characteristic		TCGA training datasets (n=250)	TCGA test datasets (n=251)	GSE65858 (n=270)
Age(years)	<=50	38	50	41
	>50	212	201	229
Survival Status	Living	136	146	176
	Dead	141	105	94
Gender	female	68	66	47
	male	182	185	223
Grade	G 1	34	28	
	G 2	136	163	
	G 3	68	51	
	G 4	2	0	
Pathologic_T	T 1	17	28	35
	T 2	67	65	80
	T 3	48	48	58
	T 4	89	63	97
Pathologic_N	N 0	80	90	94
	N 1	26	39	32
	N 2	94	72	132
Pathologic_M	N 3	2	5	12
	M 0	94	93	263
	M 1/ M X	32	30	7
	Stage I	9	16	18
Tumor Stage	Stage II	37	32	37
	Stage III	30	48	37
	Stage IV	141	120	178

MATERIALS AND METHODS

Data acquisition and processing

The FPKM data of TCGA RNA-Seq downloaded from the UCSC Cancer Browser (<https://xenabrowser.net/datapages/>) contained 546 samples. The clinical follow-up information contained 612 samples, and the copy number variation data on the SNP 6.0 chip contained 519 samples. The mutation annotation information (MAF) downloaded from the GDC client contained 504 samples. The standardized tables of the GSE658587 [36] data set were downloaded from the GEO. For TCGA RNAseq data, 501 tumor samples with follow-up information were chosen and randomly divided into two groups: a training set (N = 250) (Supplementary Tables 4 and 5) and a test set (N = 251)

(Supplementary Tables 6 and 7). The GSE65858 data set was used as an external verification set for each group of sample information (Table 4).

Univariate Cox proportional risk regression analysis

As described previously by Jin-Cheng et al. [37], univariate Cox proportional hazard regression analysis was performed with each immune gene to screen out genes significantly associated with OS in the training data set. $P < 0.01$ was chosen as the threshold.

Analysis of copy number variation data

GISTIC software is widely used to detect both broad and focal (potentially overlapping) recurring events. We used GISTIC 2.0 [38] to identify genes exhibiting significant

amplification or deletion. The parameter threshold was that the length of the amplification or deletion was more than 0.1, and the fragment was $P < 0.05$.

Gene mutation analysis

We used Mutsig 2.0 software with TCGA mutation data to identify genes with significant mutations in their MAF files. The threshold used was $P < 0.05$.

Construction of a prognostic gene signature

We chose the genes that were significantly related to OS and genes that were exhibited amplification, deletion or mutation. We further used the Random Survival Forest algorithm to rank the importance of prognostic genes. Like Jin et al. [39], we used the R package random Survival Forest to screen the prognostic genes. We set the number of Monte Carlo iterations to 100 and the number of steps forward to 5, and identified the genes whose relative importance as characteristic genes was greater than 0.4. In addition, we carried out a multivariate Cox regression analysis, and the following risk scoring model was constructed using the Equation 2:

$$RickScore = \sum_{k=1}^n Exp_k * e^{HR_k} \quad (2)$$

where N is the number of prognostic genes, Exp_k is the expression value of the prognostic genes, and e^{HR_k} is the estimated regression coefficient of genes in the multivariate Cox regression analysis.

Functional enrichment analyses

GO and KEGG pathway enrichment analysis was performed using the R package cluster profiler for genes [40], to identify over-represented GO terms in three categories (biological processes, molecular function and cellular component) as well as over-represented KEGG pathway terms. For this analyses, a false discovery rate < 0.05 was considered to indicate statistical significance.

GSEA [41] was performed using the <http://software.broadinstitute.org/gsea/downloads.jsp> website with the MSigDB [42] C2 Canonical pathways gene set collection, which contains 1320 gene sets. Gene sets with a false discovery rate < 0.05 after performing 1000 permutations were considered to be significantly enriched.

Statistical analysis

When the median risk score in each data set was used as a cutoff to compare survival risk between high-risk and low-risk groups, a Kaplan-Meier (KM) curve was

drawn. Multivariate Cox regression analysis was used to test whether gene markers were independent prognostic factors. Significance was defined as $P < 0.05$. All analyses were performed using R 3.4.3.

Abbreviations

HNSCC: head and neck squamous cell carcinoma; TCGA: the cancer genome atlas; GEO: group on earth observations; SNPs: single nucleotide polymorphisms; CNV: copy number variation; SCC-A: squamous cell carcinoma antigen; HPV: human papilloma virus; MFA: the mutation annotation information; OS: overall survival; GO: gene ontology; KEGG: kyoto encyclopedia of genes and genomes; KM: Kaplan-Meier; HR: higher risk; PPI: protein-protein interaction.

CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

FUNDING

The work was supported by The Natural Science Foundation of China (no.81660294, 81602389, 81472696, 81202128, 81272974) and The Natural Science Youth Funds of Jiangxi Province (no.2016BAB215249).

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin.* 2016; 66:7–30. <https://doi.org/10.3322/caac.21332> PMID:26742998
2. Leemans CR, Braakhuis BJ, Brakenhoff RH. The molecular biology of head and neck cancer. *Nat Rev Cancer.* 2011; 11:9–22. <https://doi.org/10.1038/nrc2982> PMID:21160525
3. Chauhan SS, Kaur J, Kumar M, Matta A, Srivastava G, Alyass A, Assi J, Leong I, MacMillan C, Witterick I, Colgan TJ, Shukla NK, Thakar A, et al. Prediction of recurrence-free survival using a protein expression-based risk classifier for head and neck cancer. *Oncogenesis.* 2015; 4:e147. <https://doi.org/10.1038/oncsis.2015.7> PMID:25893634
4. Zhao X, Sun S, Zeng X, Cui L. Expression profiles analysis identifies a novel three-mRNA signature to predict overall survival in oral squamous cell carcinoma. *Am J Cancer Res.* 2018; 8:450–61. PMID:29637000
5. De Cecco L, Bossi P, Locati L, Canevari S, Licitra L. Comprehensive gene expression meta-analysis of head and neck squamous cell carcinoma microarray data

- defines a robust survival predictor. *Ann Oncol.* 2014; 25:1628–35.
<https://doi.org/10.1093/annonc/mdu173>
PMID:[24827125](https://pubmed.ncbi.nlm.nih.gov/24827125/)
6. Tian S, Meng G, Zhang W. A six-mRNA prognostic model to predict survival in head and neck squamous cell carcinoma. *Cancer Manag Res.* 2018; 11:131–42.
<https://doi.org/10.2147/CMAR.S185875>
PMID:[30588115](https://pubmed.ncbi.nlm.nih.gov/30588115/)
 7. Yeon Yeon S, Jung SH, Jo YS, Choi EJ, Kim MS, Chung YJ, Lee SH. Immune checkpoint blockade resistance-related B2M hotspot mutations in microsatellite-unstable colorectal carcinoma. *Pathol Res Pract.* 2019; 215:209–14.
<https://doi.org/10.1016/j.prp.2018.11.014>
PMID:[30503610](https://pubmed.ncbi.nlm.nih.gov/30503610/)
 8. Padhi SS, Roy S, Kar M, Saha A, Roy S, Adhya A, Baisakh M, Banerjee B. Role of CDKN2A/p16 expression in the prognostication of oral squamous cell carcinoma. *Oral Oncol.* 2017; 73:27–35.
<https://doi.org/10.1016/j.oraloncology.2017.07.030>
PMID:[28939073](https://pubmed.ncbi.nlm.nih.gov/28939073/)
 9. Hou SQ, Ouyang M, Brandmaier A, Hao H, Shen WH. PTEN in the maintenance of genome integrity: from DNA replication to chromosome segregation. *BioEssays.* 2017; 39:1700082.
<https://doi.org/10.1002/bies.201700082>
PMID:[28891157](https://pubmed.ncbi.nlm.nih.gov/28891157/)
 10. Soussi T, Wiman KG. TP53: an oncogene in disguise. *Cell Death Differ.* 2015; 22:1239–49.
<https://doi.org/10.1038/cdd.2015.53> PMID:[26024390](https://pubmed.ncbi.nlm.nih.gov/26024390/)
 11. Llovet JM, Ricci S, Mazzaferro V, Hilgard P, Gane E, Blanc JF, de Oliveira AC, Santoro A, Raoul JL, Forner A, Schwartz M, Porta C, Zeuzem S, et al, and SHARP Investigators Study Group. Sorafenib in advanced hepatocellular carcinoma. *N Engl J Med.* 2008; 359:378–90.
<https://doi.org/10.1056/NEJMoa0708857>
PMID:[18650514](https://pubmed.ncbi.nlm.nih.gov/18650514/)
 12. Cheng AL, Kang YK, Chen Z, Tsao CJ, Qin S, Kim JS, Luo R, Feng J, Ye S, Yang TS, Xu J, Sun Y, Liang H, et al. Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma: a phase III randomised, double-blind, placebo-controlled trial. *Lancet Oncol.* 2009; 10:25–34.
[https://doi.org/10.1016/S1470-2045\(08\)70285-7](https://doi.org/10.1016/S1470-2045(08)70285-7)
PMID:[19095497](https://pubmed.ncbi.nlm.nih.gov/19095497/)
 13. Siow ZR, De Boer RH, Lindeman GJ, Mann GB. Spotlight on the utility of the Oncotype DX® breast cancer assay. *Int J Womens Health.* 2018; 10:89–100.
<https://doi.org/10.2147/IJWH.S124520>
PMID:[29503586](https://pubmed.ncbi.nlm.nih.gov/29503586/)
 14. Bhutiani N, Egger ME, Ajkay N, Scoggins CR, Martin RC 2nd, McMasters KM. Multigene signature panels and breast cancer therapy: patterns of use and impact on clinical decision making. *J Am Coll Surg.* 2018; 226:406–412.e1.
<https://doi.org/10.1016/j.jamcollsurg.2017.12.043>
PMID:[29366844](https://pubmed.ncbi.nlm.nih.gov/29366844/)
 15. Wang SY, Dang W, Richman I, Mougalian SS, Evans SB, Gross CP. cost-effectiveness analyses of the 21-gene assay in breast cancer: systematic review and critical appraisal. *J Clin Oncol.* 2018; 36:1619–27.
<https://doi.org/10.1200/JCO.2017.76.5941>
PMID:[29659329](https://pubmed.ncbi.nlm.nih.gov/29659329/)
 16. Kopetz S, Taberero J, Rosenberg R, Jiang ZQ, Moreno V, Bachleitner-Hofmann T, Lanza G, Stork-Sloots L, Maru D, Simon I, Capellà G, Salazar R. Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *Oncologist.* 2015; 20:127–33.
<https://doi.org/10.1634/theoncologist.2014-0325>
PMID:[25561511](https://pubmed.ncbi.nlm.nih.gov/25561511/)
 17. Tan IB, Tan P. Genetics: an 18-gene signature (ColoPrint®) for colon cancer prognosis. *Nat Rev Clin Oncol.* 2011; 8:131–33.
<https://doi.org/10.1038/nrclinonc.2010.229>
PMID:[21304502](https://pubmed.ncbi.nlm.nih.gov/21304502/)
 18. Maak M, Simon I, Nitsche U, Roepman P, Snel M, Glas AM, Schuster T, Keller G, Zeestraten E, Goossens I, Janssen KP, Friess H, Rosenberg R. Independent validation of a prognostic genomic signature (ColoPrint) for patients with stage II colon cancer. *Ann Surg.* 2013; 257:1053–58.
<https://doi.org/10.1097/SLA.0b013e31827c1180>
PMID:[23295318](https://pubmed.ncbi.nlm.nih.gov/23295318/)
 19. Zhao X, Li Y, Wu H. A novel scoring system for acute myeloid leukemia risk assessment based on the expression levels of six genes. *Int J Mol Med.* 2018; 42:1495–507.
<https://doi.org/10.3892/ijmm.2018.3739>
PMID:[29956722](https://pubmed.ncbi.nlm.nih.gov/29956722/)
 20. Granell M, Urbano-Ispizua A, Pons A, Aróstegui JI, Gel B, Navarro A, Jansa S, Artells R, Gaya A, Talarn C, Fernández-Avilés F, Martínez C, Rovira M, et al. Common variants in NLRP2 and NLRP3 genes are strong prognostic factors for the outcome of HLA-identical sibling allogeneic stem cell transplantation. *Blood.* 2008; 112:4337–42.
<https://doi.org/10.1182/blood-2007-12-129247>
PMID:[18772453](https://pubmed.ncbi.nlm.nih.gov/18772453/)
 21. Pan JX, Qu F, Wang FF, Xu J, Mu LS, Ye LY, Li JJ. Aberrant SERPINE1 DNA methylation is involved in carboplatin induced epithelial-mesenchymal transition in epithelial ovarian cancer. *Arch Gynecol Obstet.*

- 2017; 296:1145–52.
<https://doi.org/10.1007/s00404-017-4547-x>
PMID:28975405
22. Azimi I, Petersen RM, Thompson EW, Roberts-Thomson SJ, Monteith GR. Hypoxia-induced reactive oxygen species mediate N-cadherin and SERPINE1 expression, EGFR signalling and motility in MDA-MB-468 breast cancer cells. *Sci Rep*. 2017; 7:15140.
<https://doi.org/10.1038/s41598-017-15474-7>
PMID:29123322
23. Yu XM, Jaskula-Sztul R, Georgen MR, Aburjania Z, Somnay YR, Levenson G, Sippel RS, Lloyd RV, Johnson BP, Chen H. Notch1 signaling regulates the aggressiveness of differentiated thyroid cancer and inhibits SERPINE1 expression. *Clin Cancer Res*. 2016; 22:3582–92.
<https://doi.org/10.1158/1078-0432.CCR-15-1749>
PMID:26847059
24. Trabert B, Eldridge RC, Pfeiffer RM, Shiels MS, Kemp TJ, Guillemette C, Hartge P, Sherman ME, Brinton LA, Black A, Chaturvedi AK, Hildesheim A, Berndt SI, et al. Prediagnostic circulating inflammation markers and endometrial cancer risk in the prostate, lung, colorectal and ovarian cancer (PLCO) screening trial. *Int J Cancer*. 2017; 140:600–10.
<https://doi.org/10.1002/ijc.30478> PMID:27770434
25. Mazzoccoli G, Paziienza V, Panza A, Valvano MR, Benegiamo G, Vinciguerra M, Andriulli A, Piepoli A. ARNTL2 and SERPINE1: potential biomarkers for tumor aggressiveness in colorectal cancer. *J Cancer Res Clin Oncol*. 2012; 138:501–11.
<https://doi.org/10.1007/s00432-011-1126-6>
PMID:22198637
26. Ramos-García P, González-Moles MA, Ayén Á, González-Ruiz L, Ruiz-Ávila I, Gil-Montoya JA. Prognostic and clinicopathological significance of CTTN/cortactin alterations in head and neck squamous cell carcinoma: systematic review and meta-analysis. *Head Neck*. 2019; 41:1963–78.
<https://doi.org/10.1002/hed.25632> PMID:30597652
27. Zhang S, Qi Q. MTSS1 suppresses cell migration and invasion by targeting CTTN in glioblastoma. *J Neurooncol*. 2015; 121:425–31.
<https://doi.org/10.1007/s11060-014-1656-2>
PMID:25385572
28. Hu X, Moon JW, Li S, Xu W, Wang X, Liu Y, Lee JY. Amplification and overexpression of *CTTN* and *CCND1* at chromosome 11q13 in Esophagus squamous cell carcinoma (ESCC) of North Eastern Chinese Population. *Int J Med Sci*. 2016; 13:868–74.
<https://doi.org/10.7150/ijms.16845>
PMID:27877079
29. Long HC, Gao X, Lei CJ, Zhu B, Li L, Zeng C, Huang JB, Feng JR. miR-542-3p inhibits the growth and invasion of colorectal cancer cells through targeted regulation of cortactin. *Int J Mol Med*. 2016; 37:1112–18.
<https://doi.org/10.3892/ijmm.2016.2505>
PMID:26952924
30. Han J, Jackson D, Holm J, Turner K, Ashcraft P, Wang X, Cook B, Arning E, Genta RM, Venuprasad K, Souza RF, Sweetman L, Theiss AL. Elevated d-2-hydroxyglutarate during colitis drives progression to colorectal cancer. *Proc Natl Acad Sci USA*. 2018; 115:1057–62.
<https://doi.org/10.1073/pnas.1712625115>
PMID:29339485
31. Qin F, Song Z, Chang M, Song Y, Frierson H, Li H. Recurrent cis-SAGE chimeric RNA, D2HGDH-GAL3ST2, in prostate cancer. *Cancer Lett*. 2016; 380:39–46.
<https://doi.org/10.1016/j.canlet.2016.06.013>
PMID:27322736
32. Brehmer S, Pusch S, Schmieder K, von Deimling A, Hartmann C. Mutational analysis of D2HGDH and L2HGDH in brain tumours without IDH1 or IDH2 mutations. *Neuropathol Appl Neurobiol*. 2011; 37:330–32.
<https://doi.org/10.1111/j.1365-2990.2010.01114.x>
PMID:20727073
33. Krell D, Assoku M, Galloway M, Mulholland P, Tomlinson I, Bardella C. Screen for IDH1, IDH2, IDH3, D2HGDH and L2HGDH mutations in glioblastoma. *PLoS One*. 2011; 6:e19868.
<https://doi.org/10.1371/journal.pone.0019868>
PMID:21625441
34. Kou Y, Koag MC, Lee S. N7 methylation alters hydrogen-bonding patterns of guanine in duplex DNA. *J Am Chem Soc*. 2015; 137:14067–70.
<https://doi.org/10.1021/jacs.5b10172> PMID:26517568
35. Kou Y, Koag MC, Lee S. Structural and kinetic studies of the effect of guanine N7 alkylation and metal cofactors on DNA replication. *Biochemistry*. 2018; 57:5105–16.
<https://doi.org/10.1021/acs.biochem.8b00331>
PMID:29957995
36. Wichmann G, Rosolowski M, Krohn K, Kreuz M, Boehm A, Reiche A, Scharrer U, Halama D, Bertolini J, Bauer U, Holzinger D, Pawlita M, Hess J, et al, and Leipzig Head and Neck Group (LHNG). The role of HPV RNA transcription, immune response-related gene expression and disruptive TP53 mutations in diagnostic and prognostic profiling of head and neck cancer. *Int J Cancer*. 2015; 137:2846–57.
<https://doi.org/10.1002/ijc.29649> PMID:26095926
37. Guo JC, Wu Y, Chen Y, Pan F, Wu ZY, Zhang JS, Wu JY, Xu XE, Zhao JM, Li EM, Zhao Y, Xu LY. Protein-coding genes combined with long noncoding RNA as a novel

- transcriptome molecular staging model to predict the survival of patients with esophageal squamous cell carcinoma. *Cancer Commun (Lond)*. 2018; 38:4.
<https://doi.org/10.1186/s40880-018-0277-0>
PMID:[29784063](https://pubmed.ncbi.nlm.nih.gov/29784063/)
38. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhi R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011; 12:R41.
<https://doi.org/10.1186/gb-2011-12-4-r41>
PMID:[21527027](https://pubmed.ncbi.nlm.nih.gov/21527027/)
39. Meng J, Li P, Zhang Q, Yang Z, Fu S. A four-long non-coding RNA signature in predicting breast cancer survival. *J Exp Clin Cancer Res*. 2014; 33:84.
<https://doi.org/10.1186/s13046-014-0084-7>
PMID:[25288503](https://pubmed.ncbi.nlm.nih.gov/25288503/)
40. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012; 16:284–87.
<https://doi.org/10.1089/omi.2011.0118>
PMID:[22455463](https://pubmed.ncbi.nlm.nih.gov/22455463/)
41. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*. 2007; 23:3251–53.
<https://doi.org/10.1093/bioinformatics/btm369>
PMID:[17644558](https://pubmed.ncbi.nlm.nih.gov/17644558/)
42. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011; 27:1739–40.
<https://doi.org/10.1093/bioinformatics/btr260>
PMID:[21546393](https://pubmed.ncbi.nlm.nih.gov/21546393/)

SUPPLEMENTARY MATERIAL

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Tables 1–6

Supplementary Table 1. Amplified 247 genes.

Supplementary Table 2. 302 genes with significant mutation frequencies.

Supplementary Table 3. Information of training set.

Supplementary Table 4. Clinical information of training set.

Supplementary Table 5. Information of test set.

Supplementary Table 6. Clinical information of test set.

Supplementary Table 7. The top twenty enriched pathways.

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
KEGG_ADHERENS_JUNCTION	73	0.604583	1.982844	0	0.076978	0.042
KEGG_FOCAL_ADHESION	199	0.548716	1.832598	0.014199	0.257997	0.24
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	55	0.478323	1.751252	0.00789	0.378554	0.417
KEGG_RENAL_CELL_CARCINOMA	70	0.464708	1.714178	0.00996	0.388839	0.51
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	212	0.423736	1.680424	0.010267	0.402515	0.592
KEGG_ECM_RECEPTOR_INTERACTION	83	0.595027	1.672061	0.061181	0.35479	0.61
KEGG_AXON_GUIDANCE	128	0.43319	1.669205	0.006424	0.310781	0.615
KEGG_TGF_BETA_SIGNALING_PATHWAY	85	0.465713	1.622249	0.027668	0.380817	0.719
KEGG_WNT_SIGNALING_PATHWAY	149	0.398066	1.59946	0.010616	0.392898	0.76
KEGG_GAP_JUNCTION	89	0.424439	1.583867	0.023207	0.388653	0.785
KEGG_THYROID_CANCER	29	0.468404	1.56782	0.044834	0.391882	0.81
KEGG_ERBB_SIGNALING_PATHWAY	87	0.420793	1.547303	0.025263	0.403438	0.829
KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC	74	0.473225	1.542216	0.066528	0.383884	0.835
KEGG_TIGHT_JUNCTION	130	0.381005	1.530401	0.041152	0.382432	0.847
KEGG_PATHWAYS_IN_CANCER	324	0.360142	1.497865	0.035124	0.433172	0.884
KEGG_VIBRIO_CHOLERAЕ_INFECTION	52	0.40149	1.483408	0.036961	0.442476	0.894
KEGG_PANCREATIC_CANCER	70	0.408252	1.476431	0.065606	0.43181	0.897
KEGG_ALPHA_LINOLENIC_ACID_METABOLISM	19	-0.72946	-2.11189	0	0.012797	0.009
KEGG_LINOLEIC_ACID_METABOLISM	29	-0.64152	-1.84077	0.001976	0.215058	0.23
KEGG_ARACHIDONIC_ACID_METABOLISM	57	-0.52228	-1.82148	0.005952	0.168108	0.266
KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION	46	-0.61053	-1.62009	0.067864	0.626243	0.702